https://dx.doi.org/10.1016/j.enbuild.2022.111909

This is the Pre-Published Version.

GEIN: An interpretable benchmarking framework towards all building types based on machine learning

Xiaoyu Jin[1], Fu Xiao[1, 2, *], Chong Zhang[1], Ao Li[1]

[1.] Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University

[2.] Research Institute for Smart Energy, The Hong Kong Polytechnic University

[*] Corresponding author: linda.xiao@polyu.edu.hk

Abstract

Building energy performance benchmarking is adopted by many countries in the world as an effective tool to reduce energy consumption at city or country level. Machine learning holds a lot of promise for quickly and correctly predicting energy consumption from massive data, thereby it's suitable for large-scale performance assessment. However, there is a severe problem of data imbalance in building types in many datasets. Due to the lack of samples for some types of buildings, unfavorable results, such as low accuracy of prediction, are produced sometimes. Meanwhile, the poor interpretability of machine learning models makes it difficult to promote the benchmarking frameworks based on machine learning. Therefore, this study proposed a novel machine learning based building

© 2022. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/

performance benchmarking framework with improved generalization and interpretability. A reliable yet convenient data augmentation approach was established to overcome the data imbalance problem while avoiding overfitting problem. Superior results were obtained in case studies using three city-level open-source building datasets from two different countries. A complete rating framework was also proposed, with proper explanations of results at sample level. The performance of this rating framework was verified by comparing with other data-driven benchmarking frameworks. Moreover, the importance of variables was quantified and ranked, which can be a significant reference for data collectors and publishers. The results demonstrated that data augmentation can effectively solve the problem of data imbalance, which enables the universality of machine learning based benchmarking on all types of buildings. And the proposed GEIN benchmarking framework can also effectively address the issues of interpretability.

Keywords: Interpretable Building Energy Benchmarking, EUI Prediction, Machine Learning, Data Augmentation, GEIN

1. Introduction

Globally, urban areas account for more than 70% of final energy consumption and greenhouse gas emissions [1]. The building industry in city consumes a substantial amount of energy, thus there are great opportunities for energy conservation [2]. Building energy benchmarking is an effective way to provide supervision towards energy usage in buildings. As a result of benchmarking policies, building owners are more likely to take energy-saving measures, city managers and policymakers can also gain more insight from benchmarking results at the urban level. The effect of building energy performance benchmarking on energy saving is noteworthy. For example, according to a 2012 investigation by the

US Environmental Protection Agency (EPA), Energy Star benchmarks resulted in a 7% reduction in energy use for over 35,000 buildings over four years [3]. Moreover, in Australia, the National Australian Built Environment Rating System (NABER) enabled large buildings to achieve an average of 33% energy use reduction in ten years [4].

Rapidly and precisely assessing building energy efficiency at the metropolitan level has become a crucial demand for city planners and policymakers. Some existing benchmarking methods rely on comparing expected energy usage to measured building operation data. For example, the EnergyStar benchmarking framework has been adopted by nearly 25% of buildings in U.S., making it an industry-leading benchmarking tool. EnergyStar adopts the weighted ordinary least squares regression to estimate the expected building EUI based on operating hours, number of workers, floor areas, etc. [5]. However, as several studies have shown, the regression models used by EnergyStar were frequently ineffective in capturing variations in city-level building energy data [3, 6] due to over simplification. Some benchmarking frameworks, such as the EPLabel for evaluating European buildings [7], estimate energy usage using simulation methods. This kind of benchmarking framework shows advantage in reflecting details in individual buildings and contributes to the fairness of evaluation, and can also identify the improvement potential for retrofit measures. But the simulation based frameworks requires detailed building information as model inputs and takes a long time to develop [8], which result that it is very time and labor consuming to apply those framework to a large number of buildings at city level.

Meanwhile, as proven by a growing number of studies, machine learning models are capable of quickly providing precise estimations of city level building energy consumption. The benchmarking approaches based on machine learning models has made notable progress, such as DUE-B [9], which uses Classification Regression Tree to perform the classification of building groups by performance,

then identifies the abnormal building samples. The GREEN grading framework [6] used Decision Tree and XGBoost for estimating expected EUI, then compare it with the measured EUI values to evaluate building performance. EnergyStar++ [3], a modified version of EnergyStar, applies Random Forests and XGB to calculate the EUI, and then analyze variable importance and interpret the model for policymakers.

However, despite the fact that some benchmarking frameworks (e.g., EnergyStar [5], EPC England and Wales [10]) claim over 20 building property clusters, numerous studies have developed machine learning models with extremely limited (no more than 6) property types of buildings [3]. Most research just excluded minority types of buildings. Only a small number of studies have comprehensive coverage of all property types [6] [11-22], while many of the results for buildings from minority categories are not reliable [11]. This issue occurs because of the high dependence of machine learning models on data volume. Some building types only have a tiny number of samples, which are insufficient for training machine learning models. For some cases, the number of minority types of buildings can be several thousand times less than the sample number of majority types, resulting in a substantial imbalance problem known as the long tail problem in machine learning [23], as an example shown in Table 1 of Section 3. It is unreasonable to simply eliminate the data on the tail, which excludes minority types of buildings from city-level building energy benchmarking. The long-tail problem restricts the applications of machine learning based energy benchmarking towards comprehensive types of buildings at city level. Therefore, this issue requires urgent attention.

Moreover, the interpretation of machine learning models is another issue. Because machine learning is a black box model without using physical mechanisms, there is still significant skepticism in industry about employing it for practical applications. This is the reason why more and more pieces of literature are calling for the interpretability of machine learning models [3] [24], which can help the

decision makers to understand and trust the mechanisms underlying model. Most of the researchers who tried to explain the black box models mainly did this on the global basis of the model, without further explanation at the sample (i.e., individual building in this study) level. In other words, a lot of research has been done concerning how the inputs affect the overall results, but cannot provide further information on how the model make decision on each sample. For example, Yue P. et al. [25] evaluated variable importance for all testing samples using Random Forests. Jonathan et al. [11] quantified the variable importance by defining significance value for Lasso Regression, as well as leveraging the feature importance for random forests [26]. The SHapley Additive exPlanation (SHAP) method was also adopted to estimate the significance level of input variables [3] [6]. Only one study explained the result of each sample, with the method SHAP force plot [3]. It showed intuitively how the input features influence the energy prediction results. But it only compared the predicted value with the group mean value, without explaining the relationship between the predicted value and real consumption of buildings. Therefore, it is necessary to create a model that can explain the variance between expected and actual energy consumption for each building.

To address the aforementioned research gaps, this study proposed a generalizable and interpretable benchmarking framework" GEIN, whose name takes a combination of the first two letters of the word "generalizable" and "interpretable". The GEIN framework is based on machine learning, to enable a wider application of machine learning to performance benchmarking of all building types with improved interpretability. In this study, city level open-source datasets were leveraged to develop Energy Use Intensity (EUI) prediction model using machine learning methods. A simple and effective approach was developed to solve the long tail problem on building types, thereby improve the generalization capability of machine learning-based benchmarking methods in evaluating various types of buildings. To make GEIN easier to be adopted, this research aims to provide a more concise

and unified method. Therefore, all building types are trained by one integrated model, rather than being trained separately. This makes the calculation more convenient and can include as many building types as possible. The EUI prediction results are regarded as the expected energy consumption of the building. Performance scores of buildings were assigned based on the comparison of predicted and real EUI values. Superiority of this benchmarking framework was further verified by comparing GEIN with EnergyStar. Each building's score is interpreted by using the Local Interpretable Model-agnostic Explanations (LIME) method. Furthermore, input features were categorized into several categories with different importance ranks to identify the most relevant variables for benchmarking.

This paper is structured as follows. Section 2 presents the overall methodology of the research, as well as the principle of each sub-section. A brief description of the open-source data used in this study is included in Section 3. Results are summarized in Section 4, including EUI prediction result, scoring of buildings, explanation on the results by LIME and aggregated feature importance for the models. A discussion of limitations and future works are included in Section 5. And the conclusions are summarized in Section 6.

2. Methodology

The core idea of GEIN is to accurately estimate the energy consumption of a building in terms of Energy Use Intensity (EUI) which is the total annual energy use divided by the total area, then compare the measured EUI values with the estimated values for scoring the energy performance of buildings. The methodology framework of GEIN is shown in Figure 1.

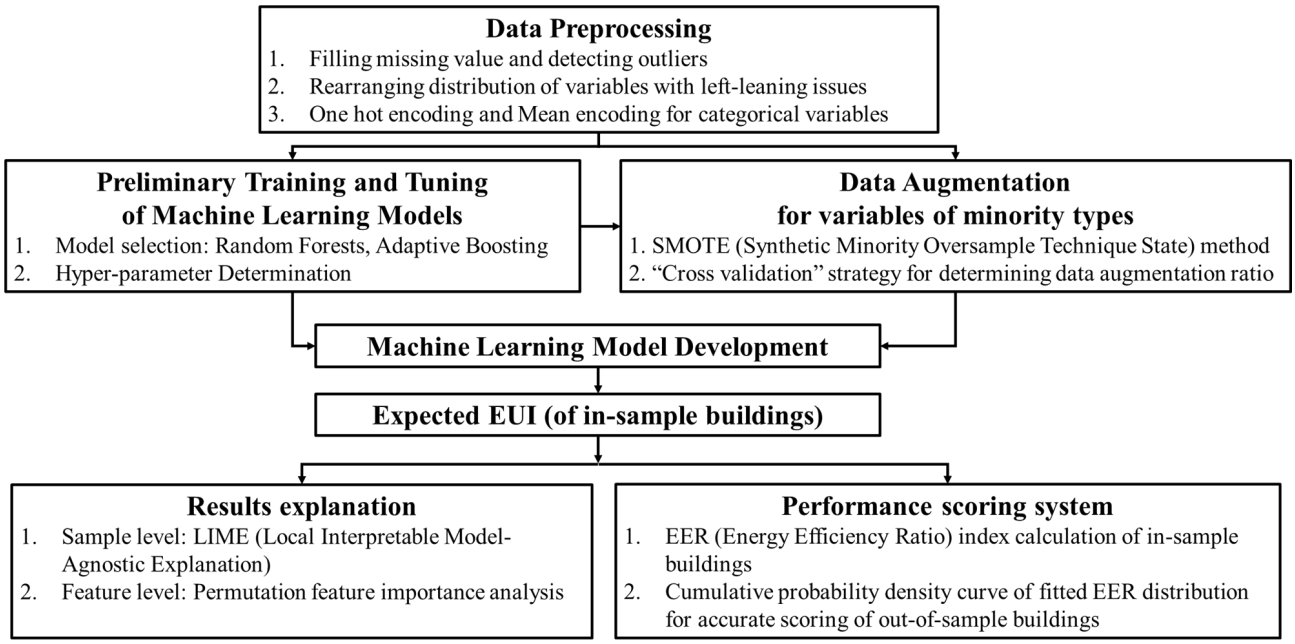


Figure 1. Methodology Diagram of GEIN

Firstly, raw data needs preprocessing to fit machine learning models, as shown in Figure 1. After data preprocessing, two machine learning methods (i.e., Random Forests (RF), Adaptive Boosting (Adaboost)) are used for preliminary training and tuning of the EUI prediction model. To tackle the long-tail issue on the variable of building property type, the Synthetic Minority Oversample Technique (SMOTE) is used for data augmentation on minority types. A cross validation-based strategy is developed to prevent overfitting problem and determine the optimal number of sample generation. A machine learning model can be developed using outcomes from previous steps, including cleaned datasets, appropriate machine learning models and generated samples. Then the expected EUI of existing buildings can be determined. Buildings are scored based on their expected EUI generated by prediction model and real EUI provided in measurement dataset. The results are interpreted from both sample level and model level by Local Interpretable Model-Agnostic Explanation (LIME) and feature importance analysis respectively.

2.1 Data Preprocessing

The quality of data determines the upper limit of machine learning model performance[27]. In this study, the raw data was processed mainly from these aspects: null value and outlier handling, distribution rearranging and variable encoding.

Because EUI is the target variable, samples with null EUI values were eliminated in the first place. The columns with too many null values (>40%) were then eliminated as well. And there are two ways to fill null values. In some datasets, some variables have both official statistics and self-reported values. This research maintained the priority of official reports, and the statistics null values can be filled with self-reported records. If no self-reported values were available, the null values were filled using K-Nearest Neighbor (KNN) imputation [28], which is a useful method for estimating a value based on other nearby data. Furthermore, histograms of variables revealed that some datasets had evident outliers that needed to be removed. Samples with negative "building age" values, for example, were removed, as are some old buildings (e.g., over 200 years old).

The next step is to rearrange the distribution of variables. Some variables have severe skewing issues, and their distribution curves generally appear distorted to the left. The tail portion of left skewed data may be mistaken as outliers by the statistical model, which has a negative impact on the model's performance [29], especially for regression-based models. Therefore, it is necessary to transform the skewed data into a Gaussian distribution or a normal distribution.

Then, categorical variables were encoded into numbers because some machine learning models only support numerical inputs. Two encoding methods were adopted in this study. First is one hot encoding. It converts each categorical value into a new categorical column and assigns a binary value of 1 or 0

to those columns, and was performed to convert variables with restricted categories (e.g., the "AC inspection condition" variable meet the criteria for including values of "yes" and "no"). However, for variables containing great number of categories (e.g., "Community district" variable contains over 100 different districts), the one hot coding method makes input matrix extremely sparse. The performance of machine learning models may be severely harmed by this problem [30] [31, 32]. Therefore, mean encoding (target encoding) was adopted for these types of variables, which is to label a category by the mean value of outputs from all the samples in this category. For example, the variable "Property Type" is represented by the averaged EUI values of buildings in the same "Property Type" group in the training dataset. This method does not affect the volume of the data and contributes to faster learning [30].

2.2 Preliminary Training and Tuning of Machine Learning Models

To obtain expected EUI of buildings, two machine learning algorithms (RF and Adaboost) are selected to develop the building EUI prediction models because of their good performance in forecasting EUI value in prior studies [3] [33]. Since both of these algorithms are the extension of Decision Tree, the principles of these three algorithms are briefly described here.

Decision Tree is an important machine learning method using tree-like mode to make prediction. It works by continuously splitting the data into different categories according to certain parameters. For a Decision Tree model, there are some crucial elements: 1) the root node, representing the entire set of samples, 2) the leaf nodes, standing for the class labels, 3) the branches, indicating the properties of the link that leads to these class labels in the tree structure [34]. Decision trees that can use continuous values as target variables are called regression trees.

Random Forests always generate a large number of decision trees during training [35]. In this study, the mean or average forecast of the individual trees is returned for regression problems. Assuming the training data set contains N samples, P independent variables, and one dependent variable, apply the Boostrap sampling method to extract N samples from the original training set with replacement to create a single decision tree. Finally, k datasets are generated through numerous rounds of sampling, and are then integrated into a random forest with k trees.

There are some important hyperparameters for Random Forests. The number of trees (expressed as "n_estimators" in sklearn) and the max depth for each decision tree (max_depth) are important parameters to control over-fitting. The number of features to consider when looking for the best split (max_features) generally has a positive correlation with model performance, but if this parameter is too large, it will reduce the diversity of a single tree in Random Forests, and affect calculation speed. The smallest number of samples needed to split an internal node (min_samples_split) limits the conditions for the continued division of the subtree. If the number of samples of a node is less than this value, it will not continue to try to select the optimal feature for division, but if the sample size is very large, increasing this value have a positive impact on the performance of the model [34].

The Adaboost algorithm [36] implements the weighted operation of multiple basic decision trees $f(x)$. The basic principle follows Equation (1).

$$F(x) = \sum_{m=1}^{M} \alpha_m f_m(x) = F_{m-1}(x) + \alpha_m f_m(x) \quad (1)$$

where $F(x)$ is the final boosting tree composed of M basic Decision Trees, $F_{m-1}(x)$ represents the lifting tree after $m-1$ rounds of iteration, α_m is the corresponding weight of the m^{th} basic decision tree, $f_m(x)$ is the m^{th} basic decision tree. Different weights are set for the sample points based on the

classification result of the previous basic decision tree. If the prediction is wrong, the weight of the sample in the next decision tree will be increased, vice versa.

For Adaboost, the hyperparameters are similar to RF [34] when the base estimator is also Decision Tree. In addition, learning rate (learning_rate) represents the weight applied to each classifier in each upgrade iteration, which is the α_m in the previous paragraph. A higher learning rate will increase the contribution of each classifier.

In order to determine the best combination of hyperparameters for the models, random search with cross validation test is applied. With these developed models, further research of data augmentation in Section 2.3, as well as EUI prediction can be conducted.

2.3 Data Augmentation

The data augmentation method Synthetic Minority Oversample Technique (SMOTE) was adopted to solve the problem of data imbalance (long-tail) on the variable of "building type" ("property use type") in building energy datasets. This section presents the rationales for employing SMOTE method, as well as its basic principle.

For imbalanced datasets, some classes can be far more dominant with regard to other remaining classes. Generally, this problem can be solved by removing samples from majority classes (e.g., random under-sampling), adding samples to minority classes (e.g., random over-sampling), or a mix of the two. Random under-sampling process is carried out by randomly eliminating samples until a balance of sample number is reached between majority and minority classes [29]. The problem of under-sampling is that it may result in the omission of potentially relevant data [37]. However, the

numbers of majority type samples in the building datasets can be drastically larger than its counterpart of minority types. To achieve data balance, only using under-sampling will be waste a considerable amount of data. Similarly, random over-sampling refers to the replication of samples from the minority group at random. The procedure of replication is repeated until a balance is achieved with respect to the majority class [29]. But by making exact copies of the existing samples, the severe problem of overfitting is highly likely to occur [37].

Besides sampling on the original datasets, data augmentation is a more optimized approach to leverage the existing data. Data augmentation refers to methods for increasing the amount of data by adding slightly changed copies of current data or creating new synthetic data from existing data[38]. Widely adopted approaches include oversampling based methods such as SMOTE, neural-network based method Generative Adversarial Network (GAN), etc.

Generative Adversarial Network (GAN) is an unsupervised learning method that involves pitting two neural networks against one another [39]. The generative confrontation network is composed of a generative network and a discriminant network. But the extremely small sample size of some building types disables this study from training GAN.

The SMOTE algorithm is another widely adopted data augmentation method, which works by selecting K nearest neighbors, connecting them and creating a composite sample in space. The algorithm uses feature vectors and their nearest neighbors to calculate the distance between these vectors, multiplies the difference by a random number between (0, 1) and add it back to the feature [40]. Unlike GAN, this algorithm doesn't require a number of input samples to generate more data. Even with very few samples, it can generate new data for the minority class. And the sample generation

number can be assigned manually. Moreover, the algorithm is relatively simple thereby easy to compute. Therefore, this study used SMOTE to settle the long tail problem.

Despite the advantages brought by data augmentation, it doesn't necessarily mean that data augmentation amount should be as more as better. A plain assumption is that with the increase of generated sample number, the performance of models will become better, but when there are too many created samples, features of the original data will be overwhelmed [37]. In other words, the new samples produced by data augmentation depict a feature distribution characteristic for the minority types. But it may deviate from the actual situation and mislead the model if there are too many forged samples. Therefore, a tradeoff between model performance improvement and the model's resemblance to the real data is needed, a proper data augmentation number (i.e., sample generation amount) should be selected.

A "cross validation" strategy is specially designed for this research to determine this optimal data augmentation number, as shown in Figure 2. The minority group samples are separated into five folds, four of which are employed by SMOTE to create data. The generated data is joined with the majority group's training set to create the model's training set. One of the folds is kept and joined with the majority group's test set to make up the testing set. In this way, the characteristics of the original minority group can be preserved. And the performance index of maintained data is calculated separately as an important performance index, which represents the proportion of the dependent variable's variation that is explained by independent variables.

The R-squared value (R^2) is used as an important index to support for the cross-validation analysis [42]. It is a statistical measure that quantifies the proportion of variation explained by an independent variable or variables in a regression model for a dependent variable. R^2 value reveals how much the

variation of one variable explains the variance of the second variable, whereas correlation explains the strength of the relationship between an independent and dependent variable. Suppose a data set includes n observation values of $y_1, \dots y_n$, and the corresponding model prediction values are $f_1, \dots f_n$ respectively, R^2 of the data set is calculated by Equation (2) below.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

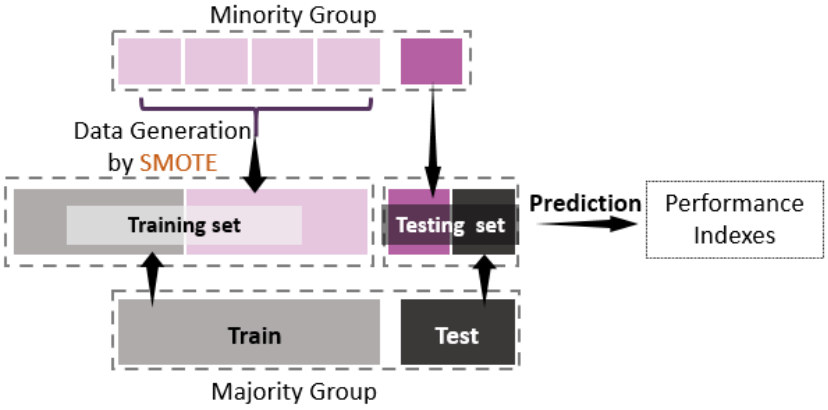


Figure 2. Performance Index Calculation for one Round

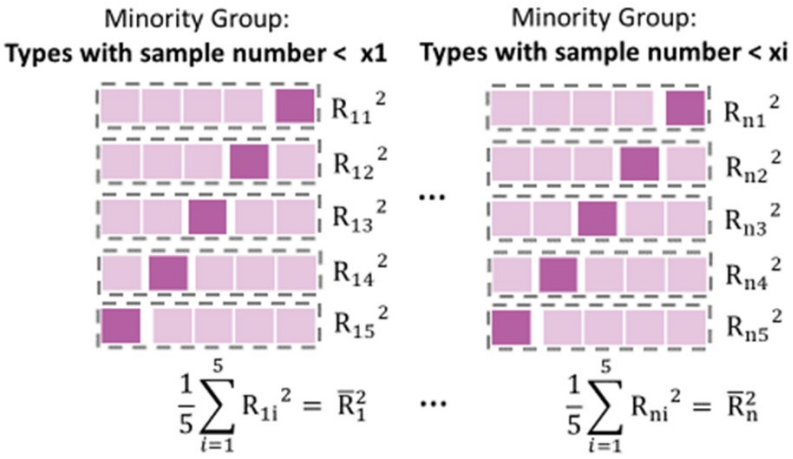


Figure 3. Cross Validation for Different Data Generation Number

Similarly, changing the maintained fold gets the R^2 value for other folds (e.g.,$R_{12}^2 \sim R_{15}^2$), as shown in Figure 3. After a 5-fold calculation, the mean value of R^2 can be determined as the performance index ($\overline{R}_1^2$) of a certain data generation number x1. Then gradually increase the data generation number, the corresponding R^2 can be calculated ($\overline{R}_n^2$).

This study uses R^2 as the essential index to do cross validation analysis, while R^2 may also have bias on data concentration. Therefore, the Root Mean Squared Error (RMSE), shown in Equation (3), is also calculated for the model at each round. To avoid the influence by data augmentation, the calculation of RMSE eliminates the generated samples.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_{i-} f_i)^2}{n}} \tag{3}$$

It should also be noticed that during each round of the data generation, the generated target number xi is set identical for buildings of all minority types, in order to balance the types of generated data sets and better solve the data imbalance problem.

2.4 Building Energy Performance Scoring

The building performance are determined by a comparison of expected EUI and real EUI. The index of building performance is defined as energy efficiency ratio (EER), as shown in Equation (4).

$$EER = \frac{EUI_{predicted}}{EUI_{real}} \tag{4}$$

The principle of index EER is easy to understand. The $EUI_{predicted}$ stands for an expected energy consumption of the buildings. Compared with this value, a lower EUI_{real} indicates lower power consumption during operation, and the performance is better. Therefore, a high EER value indicates excellent performance, and vice versa.

The performance score is calculated based on the position of its EER value in the whole dataset. The distribution of discrete EER values of in-sample buildings, which are the buildings in the datasets used to develop the EUI model, was fitted using a continuous function for smooth scoring. Even though Gamma distribution has been widely adopted by the existing benchmarking frameworks and research [2], there is no obvious evidence showing that Gamma distribution is the most suitable curve for fitting EER distribution of a massive amount of building samples in different nations. In this study, different distribution curves including *lognorm*, *gamma*, *logistic* and *laplace* were compared, by using the toolkit of optimal curve fitting in *seaborn* [43]. The r^2 value was calculated for each distribution, in order to quantify the quality of fitting. A corresponding value on the distribution curve can be found for each EER value.

The cumulative probability density curve can be determined based on corresponding distribution curve. The in-sample and out-of-sample buildings can all be benchmarked by the cumulative probability density curve. No matter what the EER value is, the corresponding probability value can be found on this curve. The performance score of a building equals to 100 times the cumulative probability density of its EER, ranking between 0~100, the higher the better.

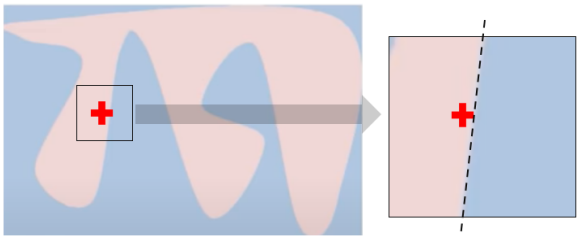
2.5 Local Interpretable Model-Agnostic Explanation (LIME)

An effective energy benchmarking framework should not only provide the score, but also give a proper explanation of it. This study used Local Interpretable Model-Agnostic Explanation (LIME) to explain the predicted EUI result of each sample.

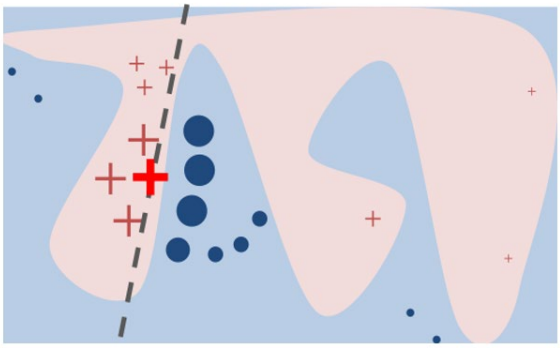
As mentioned in previous section, a number of studies used SHAP method to interpret the model by estimating the significance levels of different input variables. However, the toolkit hasn't been

developed comprehensively, which restricts its application on some algorithms such as Adaboost. Comparatively, LIME, developed by Marco T. et al in 2016 [44], has excellently wide applicability because of its calculation principle. The essence idea of LIME can be summarized as follows. If the viewing angle is narrowed to a small enough neighborhood for each sample point, a simple model can be used to explain this sample. Just as shown in Figure 4(a), by only focusing on a sample's neighbors, this model can be described as a line at the sample level, even if the whole picture is far more complex than that.

The mechanism of LIME includes the following steps [45]. Take the bold red cross in Figure 4(b) as an example. Firstly, select a data sample of interest (e.g., the bold red cross denoted as x) which requires an explanation of its black box prediction. Then sample points around, and get the black box predictions for these points. In this study, the sample of interest stands for a building requiring for score explanation, black box model stands for the EUI prediction model RF or Adaboost. Thirdly, weigh samples according to their proximity to x, which are denoted by the blue dots and light red crosses and crosses in Figure 4(b). The weight is shown on the graph as the size of these icons. Then, a simple and interpretable model (e.g., Linear Regression, Decision Tree) can be trained using these weighted samples, shown as a grey line. Finally, this interpretable model can be used to explain the results by quantifying each variable's contribution at sample level.



(a) Narrowing down the model to a small neighborhood



(b) Prediction based on weighted samples nearby

Figure 4. Schematic Diagram of LIME [44]

The local surrogate models with interpretability constraints can be stated mathematically in Equation (5) as follows [46].

$$explanation(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{5}$$

The explanation model for x is the model g (e.g., linear regression model), which evaluates how close the explanation is to the prediction of the original model f (e.g., an Adaboost model) while keeping the model complexity $\Omega(g)$ low (e.g., prefer fewer features). G stands for the group of plausible explanations, such as all linear regression models. The proximity measure specifies the size of the neighborhood that we evaluate for the explanation around instance x. In practice, LIME only optimizes the loss part. The complexity must be determined by the users, for example, they can choose the maximum number of characteristics that the linear regression model can employ.

2.6 Principle of Feature Importance Analysis

To provide priority order for input features and identify the most relevant variables for benchmarking and provide a reference of data disclosure, a feature importance analysis is required.

In this study, the importance of features was measured by permutation importance, calculated as follows. First, a scoring-based baseline measure is assessed using a dataset defined by the X. The measure is then tested after a feature column is permuted from the validation set. The difference between the baseline metric and the metric from permuting the feature column is defined as the permutation significance [34]. Outline of the permutation importance algorithm is as follows.

The inputs include a fitted predictive model m and a tabular dataset D. Then compute the reference score s of model m on dataset D (e.g., the R^2 for a regressor). For each feature (i.e., a column of D), repeat the following operations k times: randomly shuffle column j of dataset D to generate a corrupted version of the data named $\widetilde{D_{k,j}}$, then compute the score $s_{k,j}$ of model m on corrupted data $\widetilde{D_{k,j}}$. The importance value i_j for feature f_j can be defined as:

$$i_j = s - \frac{1}{K}\sum_{k=1}^{K} s_{k,j} \tag{5}$$

For each case of this study, the feature importance values were calculated separately because each case an utilized individual model with different input matrixes. In order to explore the commonality of variable requirements for all case studies and get a universal ranking of variable importance, the feature importance values need to be aggregated. To do so, there should be an approach to add up the importance values properly. But there are some issues that should be considered before the aggregation of importance values.

Firstly, for one model, the summation of feature importance values always equals to 1, which means the value of importance will be influenced by the total input number. For example, the most important variable may obtain an importance value of 0.3 for a model with six input variables, while for a model with more than 15 inputs, even the variable with the highest importance ranking obtained only around

0.1. To balance the bias brought by input number, the importance values of j-th variable in i-th case (I_{ij}) were multiplied by the model input number (N_i) before further calculation, as shown in Equation (6).

$$I_{mij} = N_i \times I_{ij} \tag{6}$$

Another issue is the input variance. Same kinds of variables may be used repeatedly as inputs in all cases (e.g., property use type), while some variables are only used in one case. Therefore, the average by occurrence (I_j) was determined as the aggregated importance index, as shown in Equation (7). Finally, the variable importance values were ranked according to I_j. For example, the variable "building age" occurred in two cases, its importance value should be: summation of I_m on "building age" in 2 cases, divided by 2.

$$I_j = \frac{summation\ of\ I_{mij}\ for\ the\ same\ variable\ of\ all\ the\ cases}{occurance\ time\ of\ this\ variable} \tag{7}$$

3. Data Description and Preprocessing

This study relies on open-source databases, which are all publicly available at reference URLs. By using these datasets, three cases were examined in this research. Case 1 employed building benchmarking data from the New York City Energy and Water Disclosure for Local Law 84 (LAW84) [47], as well as a geographical database PLUTO [48] to offer additional building attributes. Key variables include property type, land use purpose, location information (from council to community), building physical information (gross floor area, frontage, etc.), building age, occupancy rate, assessed land value and total value of building, proportion of commercial and residential area.

Case 2 used data from Chicago Energy Benchmarking [49], which had a lower sample size (2716 buildings). Variables include property type, community area, gross floor area, building age, building number, and natural gas usage. The purpose of this case is to check if EUI prediction can be made with a limited quantity of data since the sample and variable numbers are small.

Case 3 leveraged an extraction of benchmarking data from Energy Performance of Buildings in England and Wales [10], containing 45,230 building samples. A total of 14 variables were selected, related with building type, city and district, AC power, AC installation and inspection, floor area, renewable energy usage, HVAC type, occupancy, fuel type, district heating, and grid-supplied electricity. This case is to verify the applicability of the methodology framework towards building under benchmarking and recording frameworks in different countries.

There are more than 20 different categories of buildings for each above mentioned three cases. As an example, the building category distribution of case 1 is shown in Table 1. It can be seen that the building energy consumption statistics of New York City suffers from the long-tail problem on building categories. The sample numbers of some minority types are too small, compared to other building categories. The building category distributions of the other two datasets are summarized in Appendix A, both of them have similar problems. Therefore, it can be demonstrated that the long tail problem of building types is common in building energy databases.

Table 1. Building Category Distribution of
New York City Energy and Water Disclosure for Local Law 84 for Case 1

Building Category	**Sample Number**
Residential Buildings	18462
Office	2524
School	2179
Storage	754
Mall	651
Utility	492

Hospital	275
Worship Facility	170
Recreation	164
Service	138
Meeting	83
Indoor Sports	79
Laboratory	24
Restaurant	24
Prison/Incarceration	21
Personal Services	6
Ice/Curling Rink	5
Zoo	4
Data Center	4
Stadium (Open)	2

Variables of these datasets were preprocessed following the steps described in Section 2.1. The preprocessing measures include null value and outlier processing, distribution rearranging and encoding of categorical variables. Information of processed input variables is listed in Appendix B.

After completing the preceding steps, the data can be utilized to train a model. Preliminary tuning by grid searching was conducted to get the optimal models and their hyperparameters for the three cases, as introduced in Section 2.3. The results of grid searching are summarized in Table 2.

Table 2. Optimal Model and Parameters for Case 1 – Case 3

Case 1	Optimal Algorithm	Adaboost
	n_estimators	1000
	learning_rate	1.00E-04
	base estimator	DecisionTreeRegressor (max_depth=90, min_samples_split=5, min_samples_leaf=2)
Case 2	Optimal Algorithm	Random Forests
	n_estimators	1000
	min_samples_split	2
	min_samples_leaf	1
	max_features	sqrt
	max_depth	20
Case 3	Optimal Algorithm	Adaboost
	n_estimators	2000
	loss	linear
	learning_rate	1.00E-02
	base estimator	DecisionTreeRegressor (max_depth=70, min_samples_split=2, min_samples_leaf=1)

4. Results and Discussion

In this section, the proposed GEIN building energy benchmarking framework was implemented to perform EUI prediction and scoring of three different building database cases. The effectiveness of using data augmentation to improve the accuracy of building EUI prediction was investigated. Moreover, the scoring results of the EnergyStar and proposed GEIN benchmarking framework were compared to identify the superiority of the GEIN. Then interpretability and feature importance of the GEIN benchmarking were also analyzed.

4.1 Data Augmentation and EUI Prediction Results

As illustrated in Section 2.3, the proposed SMOTE method was used to generate more samples for minority building types of three cases. As an example, Figure 5 shows the effect of data augmentation on the sample number of retail buildings in case 3. An intuitive insight can be gained from the EUI distribution that sample number increased significantly after data augmentation, meanwhile, the EUI distribution remained similar. In order to quantify the similarity of these two distributions, a Kolmogorov-Smirnov test (K-S test) was performed. The hypothesis in this study is that the distributions before and after data augmentation are the same. The package of K-S test in SciPy was applied [50], which has the output of *statistics* and *pvalue*. Among them, *statistics* stands for absolute max distance, known as *D statistics* in K-S test. The closer *statistic* is to 0, the more similar these two distributions are. And *pvalue* indicates the confidence in the hypothesis. If *pvalue* is greater than the significance level, usually 0.05, the hypothesis can be determined to be true, as shown in Figure 5. The K-S test was also applied to other building categories, and the results are attached to Appendix C. According to the high *pvalue*, it can be seen that after generating data for all minority types of buildings, the EUI distributions are the same as before.

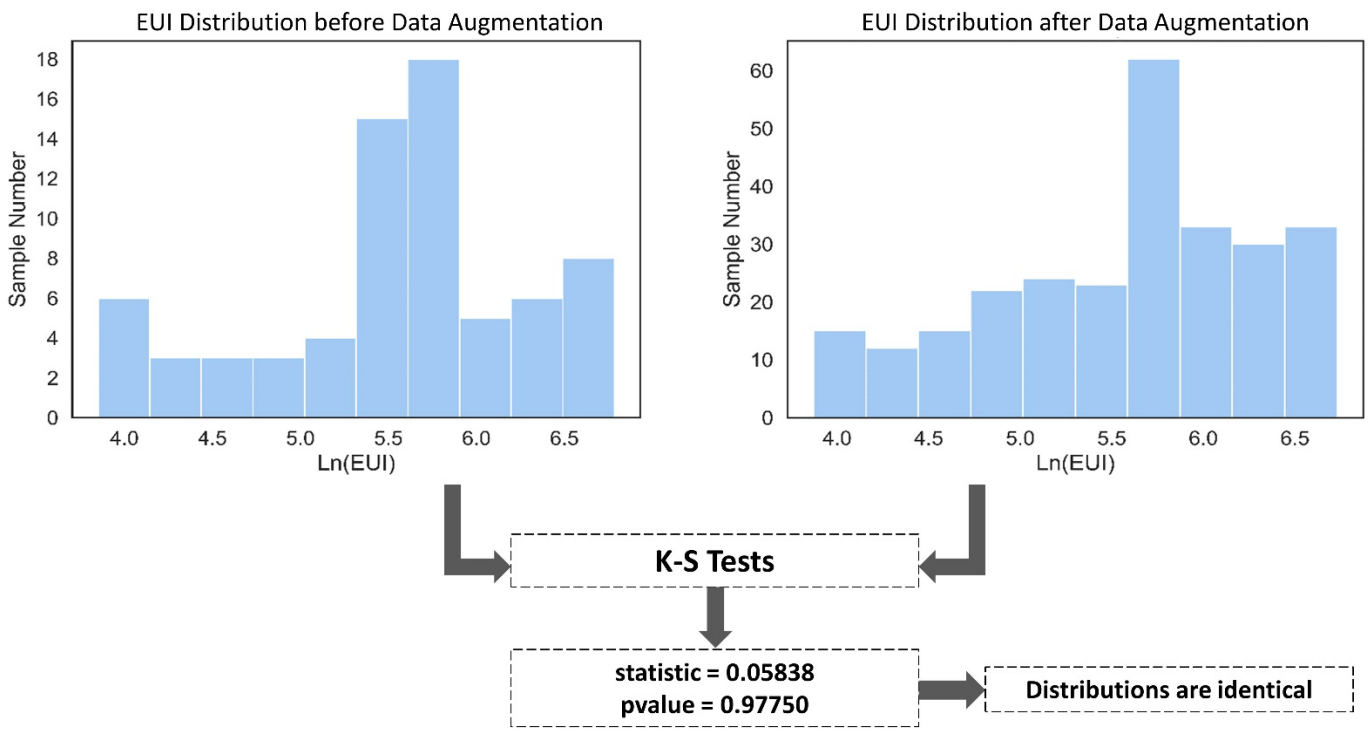


Figure 5. EUI Distributions of Retail Buildings (before & after data augmentation) and K-S Test

Then, the influence of data augmentation number on the performance of EUI prediction model was evaluated, in order to determine the optimal data augmentation numbers for all the cases. The R-squared (R^2) values of all testing data (R^2_{model}), data of majority building types (R^2_{major}) and the data of maintained minority building types (R^2_{minor}) were used as model performance indexes.

Figure 6 presents the effect of the data augmentation for three cases, it can be observed that performance on minority types (orange line) develops just as expected: data augmentation improves the performance at first, and when the augmentation number exceeds a critical value, further increasing is detrimental. This critical value occurred for all three cases, further proving the universality of this phenomenon. The critical point also represents the optimal data augmentation number for EUI prediction model, as shown in Table 3.

By observing R^2_{model} (purple curves) and R^2_{major} (grey curves), it can be concluded that the augmentation of minority samples doesn't affect the model's performance on other types of buildings, but showed a positive contribution. This should be because with more balanced samples on various building categories, the model can better distinguish different types of buildings, so that it makes more reliable prediction.

Table 3. Optimal Data Augmentation Numbers

	Sample Number before Data Augmentation	Optimal Data Augmentation Number for Each Minority Building Type	Sample Number after Data Augmentation
Case 1	28807	600	33570
Case 2	2716	55	4157
Case 3	45230	400	48981

The observation of RMSE further proves the rationality of this optimal point. For these three cases, RMSE values show a significant decrease when data augmentation begins. After a few rounds, increasing the augmented sample number doesn't bring significant benefit, such as case 1 and case 3, or even increase the error, such as case2. Even if the turning point is not exactly the same as the optimal point of minority type, it can also show that after the optimal point, it is not beneficial to continue to generate samples.

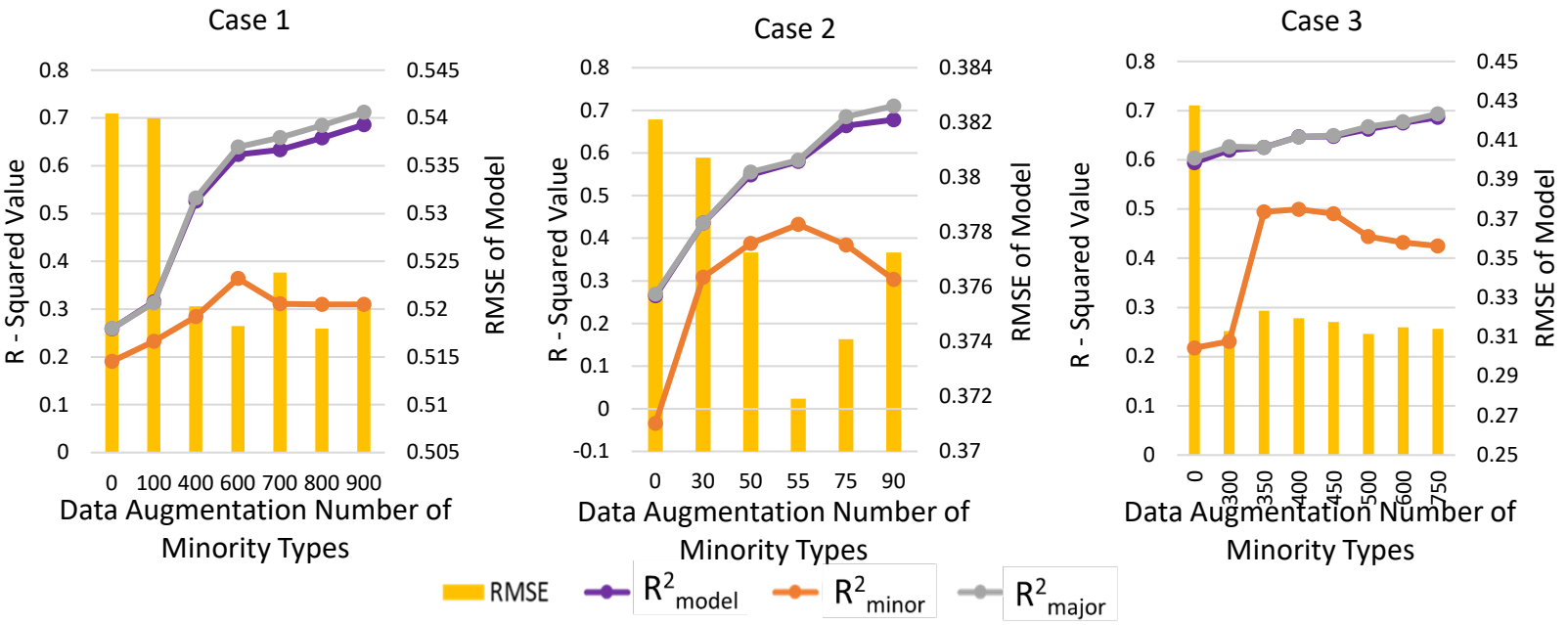


Figure 6. Influence of Data Augmentation Number on the Performance of EUI Prediction Model

The most proper amount of generated samples (listed in Table 3) and optimal hyperparameters (listed in Section 2.2) can be used to provide an accurate prediction of EUI. Moreover, the mean squared error (MSE) and Mean Absolute Percentage Error (MAPE) are used to determine the performance improvement of the EUI prediction model. The MSE and MAPE can be calculated by Equations (8) and (9), respectively.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2 \tag{8}$$

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - f_i}{y_i}\right| \tag{9}$$

where y_i is observation values; f_i is prediction values; n is the total sample number. Table 4 presents the effectiveness of data augmentation on the performance improvement of EUI prediction model. It can be seen that using data augmentation can eliminate the long tail problem of original datasets and improve the accuracy and reliability of the EUI prediction model.

Table 4. Effectiveness of Data Augmentation on Performance of EUI Prediction Model

	MSE			R^2_{model}			**MAPE**		
	Case 1	**Case 2**	**Case 3**	**Case 1**	**Case 2**	**Case 3**	**Case 1**	**Case 2**	**Case 3**
Before Data Augmentation	0.282	0.132	0.101	0.263	0.265	0.594	13.21%	6.23%	3.89%
After Data Augmentation	0.209	0.093	0.092	0.600	0.585	0.655	7.00%	4.60%	3.77%

There are also similar studies making EUI prediction for multiple types of buildings. For the fairness of comparison, the performance indexes used in the research mentioned below are the result of the testing set. For example, Travis and Michael [15] developed a regression-based model to predict the EUI of over 10 property use types, including food service, health care, office, lodging buildings, etc. The R^2 value of this model on testing dataset is just 0.4. Caleb et Al. [11] tested the performances of different machine learning models for predicting energy consumption on 20 types of buildings, while the results on different types of buildings vary greatly. Even if some majority types obtained R^2 values of 0.37-0.88, and some minority types get R^2 values between negative and 0.27. Moreover, some studies utilized the identical datasets of this study, for example, Sokratis and Constantine [6] tried to establish a novel benchmarking framework based on the dataset of LAW 84, and the R^2 value of the model is about 0.3. Jonathan et al. [51] also utilized LAW 84 and PLUTO to estimate the annual energy consumption of the buildings in the city, they compared different machine learning methods and further identified the optimal one, which obtains a minimum MSE of 0.293. But for this study, as shown in Table 4, the MSE value is only 0.209 for case 1, which used the same datasets as the above studies. It can be concluded that the proposed data-augmentation based EUI prediction model can significantly improve the prediction accuracy compared with the existing studies.

4.2 Scoring of GEIN

As illustrated in Section 2.4, the index of EER was used to score the buildings of three cases by comparing the measured and expected EUI (i.e., EUI predicted by machine learning model). Then, the EER values of three cases were fitted to four commonly used distributions (Lognorm, Gamma, Logistic, Laplace distribution) for determining the optimal curve of each case. Figure 7 illustrates the *EER* distribution and its corresponding curve fitting process for case 3. It can be found that the Laplace distribution of *EER* values achieves the highest R^2 value and is the optimal curve fitting for case 3. Laplace distribution is also the optimal curve fitting for the *EER* values of case 1 and case 2, as shown in Table 5.

Table 5. R^2 Value of Distributions for Three Cases

Case	**R^2 Value of Distributions**			
	Gamma	Lognorm	Logistic	Laplace
Case 1	0.38	0.41	0.68	0.98
Case 2	0.52	0.55	0.77	0.99
Case 3	0.45	0.48	0.73	0.98

The cumulative probability curve of *EER* distribution was applied for the scoring of in-sample buildings. Figure 8 illustrates the cumulative probability curves of the measured *EER* values and Laplace distribution fitted *EER* values for 100 randomly selected samples of case 3. Therefore, the ultimate energy performance score of each sample is determined by its *EER* value corresponding likelihood of occurrence on the Laplace fitted cumulative probability curve. To be more specific, as shown in Figure 8, the blue dot represents a building sample with an EER of 1.0, its corresponding Laplace cumulative probability is 0.5, so its ultimate score should be 100 times Laplace cumulative probability, which is 50. Similarly, since the Laplace cumulative probability curve is continuous, the in-sample building samples can always find their position on the curve, and get their corresponding

scores. A scoring table of case 3 is presented in Table 6, which lists the corresponding EER for score 10~90 at an interval of 10. The scoring tables for the other two cases are listed in Appendix D.

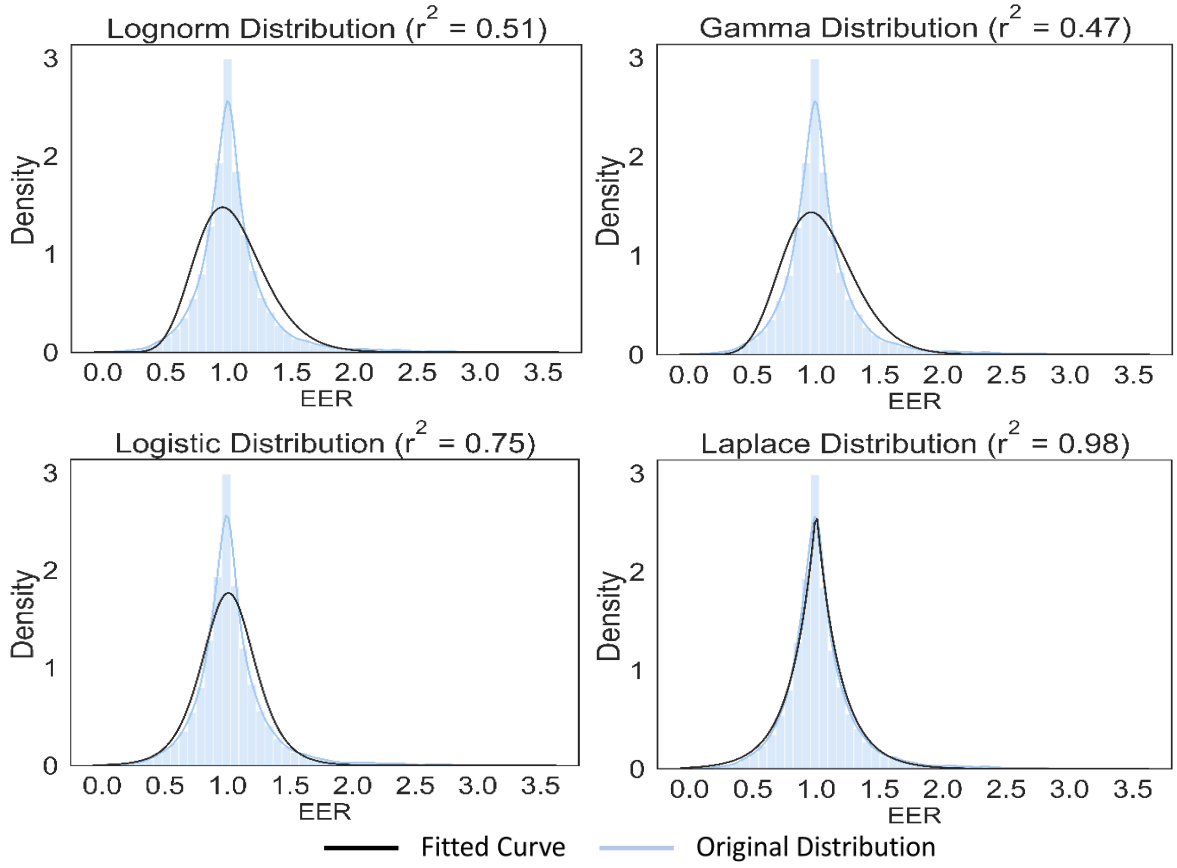


Figure 7. Curve Fitting of EER for Case 3: (a) Lognorm Distribution Curve; (b) Gamma Distribution Curve; (c) Logistic Distribution Curve; (d) Laplace Distribution Curve

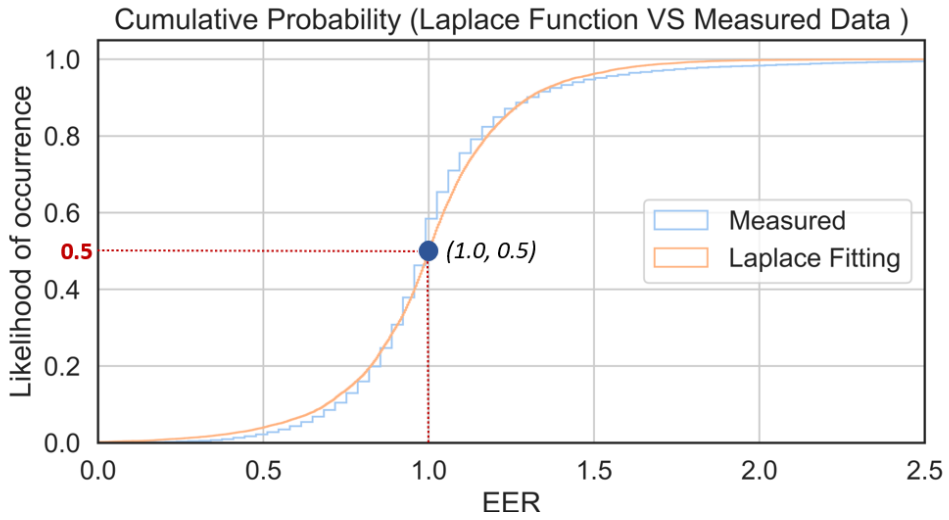


Figure 8. Cumulative Probability Curve of Laplace Fitted and Measured EER Values for Case 3

A proper comparison with other existing benchmarking frameworks is required to verify the superiority of this grading framework. EnergyStar score was used for this comparison because the proposed GEIN and EnergyStar have similar rating principles, which is also based on the comparison of predicted and measured data. So far, a number of building samples of case 1 have received their corresponding EnergyStar scores [47]. Therefore, the scoring results of EnergyStar were compared with those of GEIN for the building samples of case 1.

Table 6. Score Table of Case 3

EER	**Score**
0.696	10
0.827	20
0.903	30
0.958	40
1.000	50
1.042	60
1.097	70
1.174	80
1.304	90

To quantify the superiority of the proposed benchmarking method, previous studies tried to compare R^2 values on the validation dataset using the proposed model and EnergyStar's regression framework

[6], or calculate the scores' correlation with energy consumption [9]. Essentially, the building benchmarking framework based on EUI predictions is to determine the energy consumption level of a building in peer groups. The buildings of same peer groups have similar features such as property type and area. Therefore, a good benchmarking framework should not only have strong relevance with energy consumption, but should also consider other factors. Since the previous section has proven the model's superiority in R^2 value, this study uses Pearson Correlation Coefficient (PCC) to quantify the relevance levels.

PCC is a measure of the strength and direction of association that exists between two variables measured on at least an interval scale [52]. Given a pair of random variables (X, Y), the PPC between them is defined as the covariance of the two variables divided by the product of their standard deviations, as shown in Equation (10).

$$\rho X, Y = \frac{cov\,(X,Y)}{\sigma_X \sigma_Y} \tag{10}$$

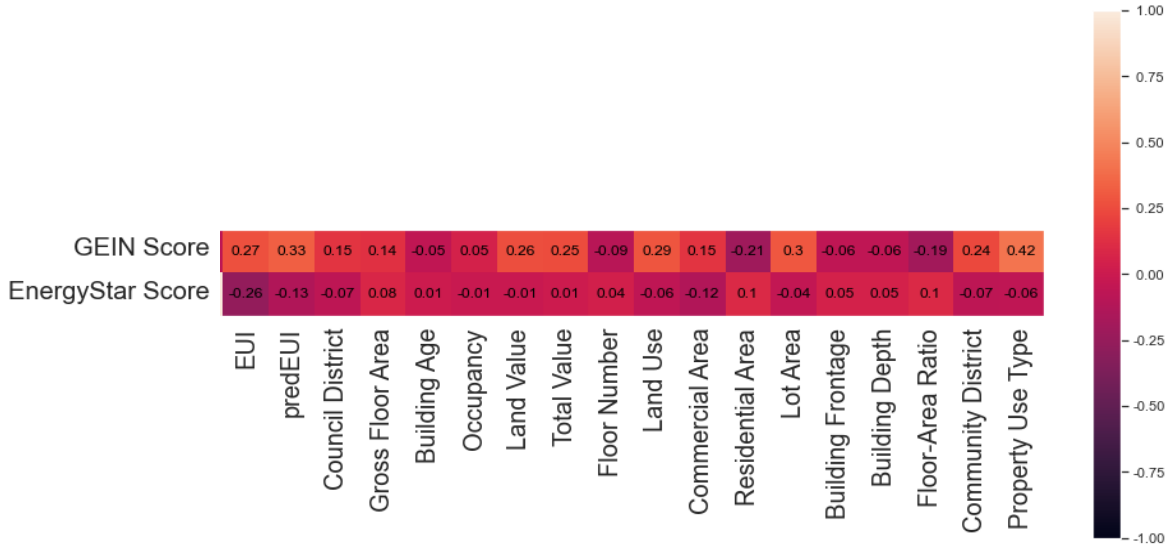


Figure 9. Heat Map of Pearson's Correlation Matrix for Comparing EnergyStar and GEIN Scores (Lighter colors indicate stronger correlation)

For GEIN and EnergyStar scores, their PCCs with energy consumption (EUI) and other factors of buildings are calculated as matrix. A heat map of PCC matrix is presented in Figure 9 for visualizing the difference between EnergyStar and GEIN. Since the assigned scores of GEIN and EnergyStar both have positive correlation coefficients with most of the other variables, a set of higher correlation coefficients can indicate a closer relationship with all other variables. To be more specific, for the benchmarking scores of these two rating frameworks, the column with lighter colors on the heat map has a stronger correlation with energy consumption, property use type, and other inputs. And it's intuitive that the column of assigned score has a lighter color, indicating a more reliable and equitable evaluation. Therefore, it can be found that the proposed GEIN benchmarking framework can better reflect the building energy use condition as well as building features.

4.3 Explanation of Results at Sample Level

There are some samples getting very low or high scores from this benchmarking framework. For example, case 1 has 556 and 363 buildings scored less than 10 and higher than 90 respectively. Therefore, it's worthwhile clarifying the reasons of these low or high scores for better understanding how the proposed GEIN benchmarking framework work on those sampled buildings, as a proper explanation of building performance rating can help evaluators understand the results better, meanwhile increasing the credibility of the rating framework.

In the framework of GEIN, the expected EUI calculation (EUI prediction) determines the value of EER. Therefore, the scores can be understood when the explanation is made on EUI prediction results. Therefore, in this section, in order to get the reasons behind the scores, the EUI prediction results are explained by the Local Interpretable Model-Agnostic Explanation (LIME) method introduced in Section 2.5.

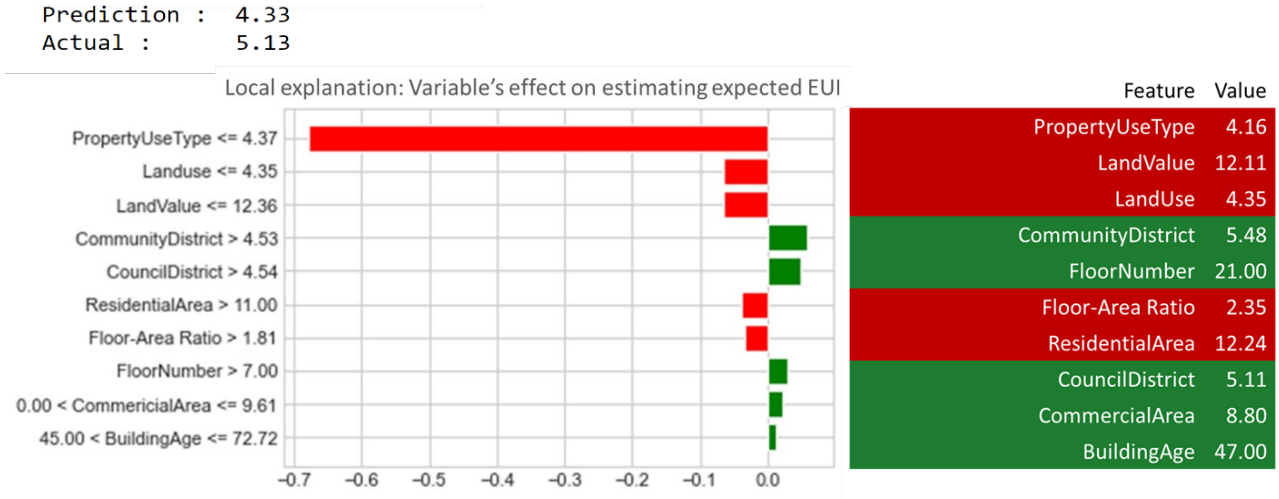


Figure 10. Local Explanation of a Low-scored Building Sample

Figure 10 shows an example of LIME-based interpretation on a low-scored (GEIN score <10) building sample of case 1. The input values are presented in the table on the right, and the red cells represent negative effects, green stands for positive. According to the benchmarking principle, a much lower expected EUI than the actual value can lead to a low score. On the left, the bar graph quantifies the impact of each input on EUI prediction, it explains why the model thinks the building should consume less energy, and the top ten most impactful variables of this sample are listed. The negative value on the horizontal axis means a reducing effect of on expected EUI estimation, presented in red; and the positive value means that the input contributes to an increase on expected EUI, shown in green. On the vertical axis, the inequality following each input variable name represents a boundary of positive and negative effects. For example, "PropertyUseType <= 4.37" means 4.37 is the threshold of the variable "PropertyUseType". For this building sample, the input of "PropertyUseType" no more than 4.37 reduces the predicted energy consumption, vice versa.

Since the inputs of GEIN include little variables related to building operation, it's hard to identify the causation in terms of building operation for unusual performance scores. But it's feasible to use LIME to gain more insight of the building's energy consumption level in various peer groups. The graph in

Figure 10 shows a high positive impact from the input variable of "PropertyUseType". Since this categorical variable was encoded by mean encoding method, the value of "PropertyUseType" implies the average energy consumption level of the same property use type. Therefore, in practice, this building sample consumes more energy than its peer group buildings with the same property use type.

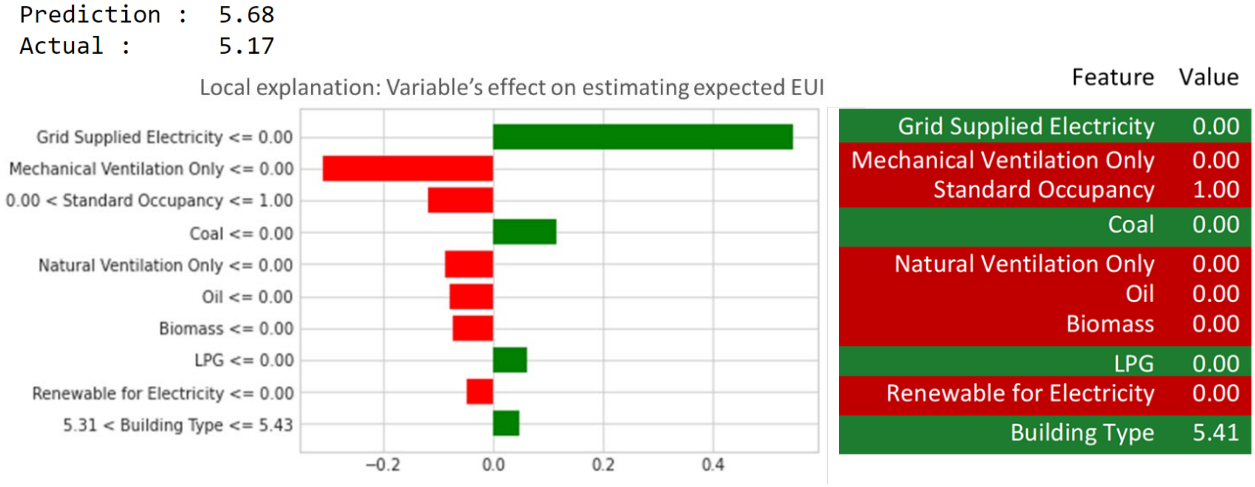


Figure 11. Local Explanation of a High-scored Building Sample

Similarly, an explanation can be made on the samples with high scores. Figure 11 shows an example of LIME explainer on a high-scored (score > 90) building sample in case 3. This building consumes much less energy compared with the buildings which also don't use grid-supplied electricity ("Grid Supplied Electricity = 0"). But HVAC type of this building actually lowers the expected EUI value. The increasing effect from coal, LPG and building type also shows a lower consumption of this building compared with buildings using similar fuels, or belonging to the same type.

4.4 Feature Importance Analysis

It is necessary to clarify what kinds of variables are more important for building energy benchmarking, and which building features relevant departments should pay more attention to when collecting and

publishing data for building benchmarking purpose. The feature importance analysis may provide an alternative solution to these questions.

For EUI prediction, the variable importance on each case was quantified by permutation importance as mentioned in Section 2.6. After integrating feature importance for the three cases, the final values of feature importance are given in Figure 12. A total of four levels of importance are defined, based on the integrated importance value: not essential (0-0.5), significant (0.5-1), important (1-2.5) and very important (>2.5).

As shown in Figure 12, the variable "Property Type" has such a significant advantage that it is the only feature at the "Very Important" level. The "Assessed value of building/land" has a fairly high priority level. Gross floor area, estimated AC power, building depth and frontage, lot area, floor-area ratio, and community district are all included at the "Important" level. Some of the "Important" variables (building depth, frontage, and floor-area ratio) come from the geographical dataset PLUTO, revealing the merits of applying an external database on EUI prediction and benchmarking.

Furthermore, implications can be driven from variables that represent comparable information. For example, most of the variables about area are at the "Important" level; the importance ranking is: Gross Floor Area > Floor Area Ratio > Lot Area > Residential Area. The importance ranking for location-related factors is Community District > Council District> City. This indicates that location features with higher accuracy are more important. For the features related to energy consumption, the numerical variable "Estimated AC power" is considerably more essential than categorical variables such as "HVAC type", "AC installation", and "AC inspection".

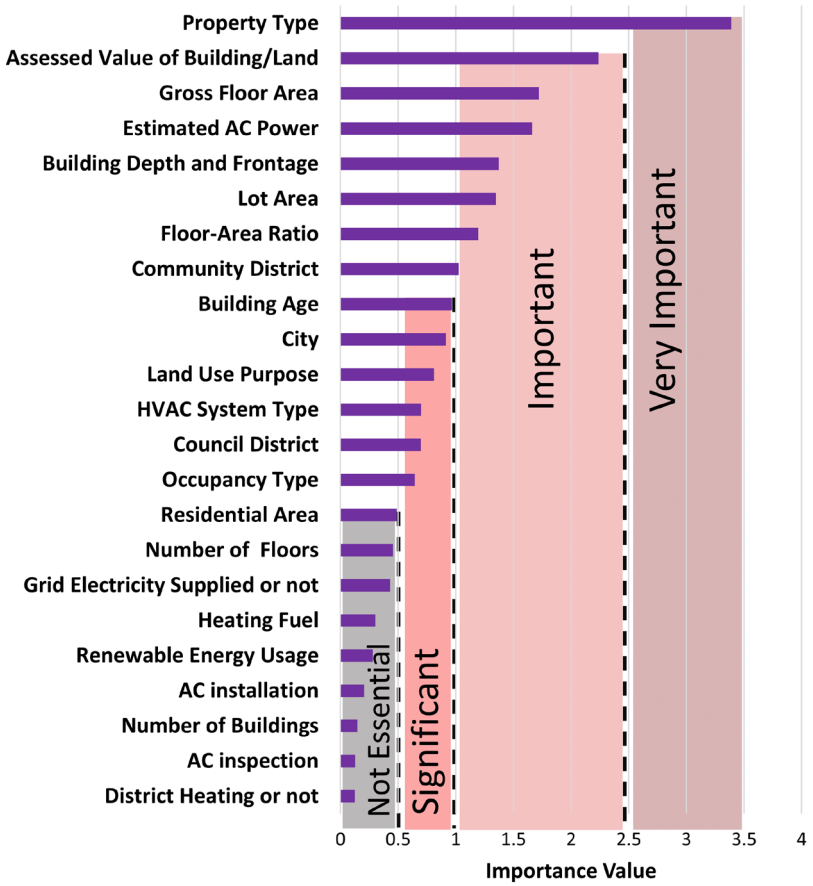
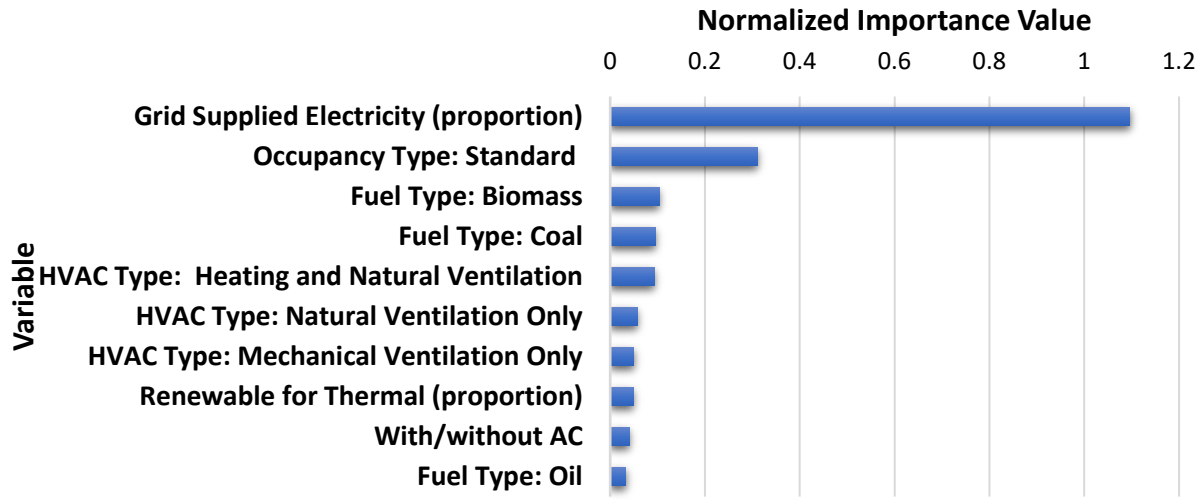
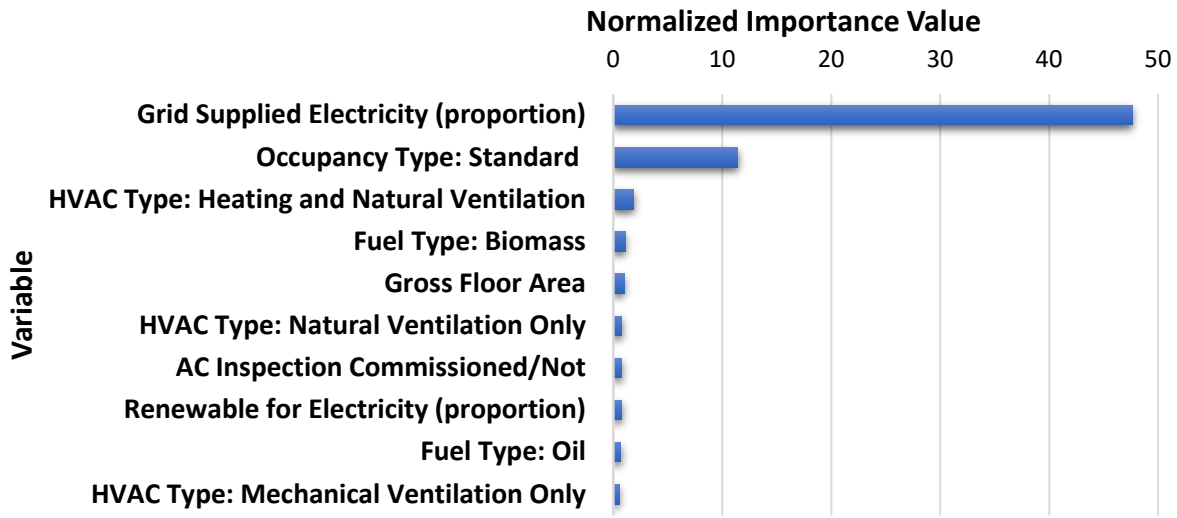


Figure 12. Ranking of Integrated Feature Importance Values

On the other hand, since the benchmarking process may refer to different variables for different building types, an importance analysis by building types will reveal the influencing factors within one type of building and bring more insight for data collection. By using LIME, the sample level feature importance matrix can be obtained, therefore, the combination of each sample within a building type can be regarded as feature importance matrix of this type of building. Figure 13 shows importance values normalized by sample number. The most significant 10 variables are shown for hotel and hospital buildings.



(a) Feature Importance Values for Hotel Buildings in Case 3



(b) Feature Importance Values for Hospital Buildings in Case 3

Figure 13. Examples of Feature Importance Ranking by Building Types

The feature importance ranking can answer the question of what kind variables should be focused on during energy benchmarking at the urban scale. Data collectors should guarantee the data quality for the important variables to get a more accurate and fair rating.

5. Limitations and Future work

The above section presents the main results of this study and demonstrates that the proposed GEIN benchmarking framework can effectively address the issues of interpretability of scoring results and applicability of all building types. Further discussions are also needed for clarifying the opportunities and challenges of GEIN.

It should be noted that there exists an intrinsic limitation for this kind of comparison-based benchmarking mechanism, including EnergyStar and GEIN. For GEIN, the performance score is determined based on comparing predicted EUI and real EUI. Here the predicted EUI is regarded as the expected energy consumption of the building, and this is based on the assumption that the EUI prediction model is very reliable and accurate so that it can estimate a building's expected energy consumption corresponding to its certain features. But the reliability of this kind of benchmarking is difficult to quantify. In other words, if the difference between real EUI and predicted EUI is large, sometimes it's hard to identify whether it's the problem of the building performance or the model, even if this concern can be settled to some extend by sample level explanation in this study.

In terms of feature importance analysis, there might be some factors that can affect the fairness of importance value aggregation. Different building samples or datasets will result in different feature importance raking, and the ultimate importance values are represented by the average occurrence in this study. However, some variables might be very important for a certain case, while not essential for other cases. But if this variable is not included in the case where it should be less important, the average by occurrence will overestimate its importance. Also, it seems that some of the variables show limited influence on building benchmarking. Neglecting these variables may achieve trade-off between accuracy and calculation efficiency.

Moreover, this study was conducted based on publicly available data. The importance of some unpublished features of buildings may have no chance to be investigated. This means that some potential problems during operation cannot be displayed by this benchmarking framework in the current stage. On the other hand, the scoring framework proposed in this study only guarantees statistical fairness and reasonableness, without considering the evaluator's intentions and policies in practice. For example, a targeted building energy efficiency policy is about to be implemented, and the evaluator will be recommended to adjust certain levels of thresholds in order to more accurately define the target group.

Therefore, the future work will apply this benchmarking framework with other evaluation frameworks, utilize the advantages of this framework to quickly and accurately analyze city-level building energy ratings, and make up for potential model reliability problems. Further investigations will be carried out to estimate the significance of different input variables, which may involve cooperation with the city's building data management department. Moreover, this building energy data and performance rating framework is expected to be related to more specific local issues or policies.

6. Conclusions

Machine learning is an effective tool to benchmark a large number of buildings' energy performance efficiently. However, machine learning methods may lead to undesirable results on some minority types of buildings due to the lack of data. On the other hand, the lack of interpretability has always been a barrier for broader applications of machine learning models. In this study, a machine learning based benchmarking framework GEIN was developed to provide a generalizable and interpretable building energy rating framework for all building types. The framework was specially designed to address the issue on data imbalance of building categories through data augmentation by

using a concise method SMOTE. The LIME method was adopted to address the interpretability issue of machine learning model, making the results of GEIN benchmarking framework more understandable.

The GEIN benchmarking framework was implemented to perform EUI prediction and scoring of three building database cases, containing 28807, 2716, and 45230 real building samples respectively. The effectiveness of using data augmentation to improve the accuracy of building EUI prediction was investigated and compared with the existing studies. Moreover, a comparison of scoring results of the EnergyStar and GEIN were carried to identify the superiority of the GEIN. The interpretability and feature importance analysis of the benchmarking framework were also analyzed. The main findings of this study can be concluded as follows: (1) Data augmentation process of GEIN framework can solve the data imbalance issue on building categories and significantly improve the EUI prediction accuracy of the whole building sample database, especially for the minority building categories; (2) The scoring results of the GEIN benchmarking framework show stronger explainable relationship with practical energy use scenarios and building features than that of the EnergyStar scores; (3) The GEIN framework realizes the interpretability of machine learning based scoring results at the individual building sample level, which will help the users to understand the proposed benchmarking framework and increase its reliability; (4) Feature importance analysis provides a guideline of building energy benchmarking to identify the most crucial and negligible variables or building features, which are significant for improving the accuracy and calculation efficiency of benchmarking.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors gratefully acknowledge the support of this research by the National Key Research and Development Program of China (2021YFE0107400), Hong Kong Scholars Program (XJ2019044) and the Research Grant Council of the Hong Kong SAR (152133/19E).

Appendixes

Appendix A. Building Category Distribution of Case 2 and Case 3

Energy Performance of Buildings in England and Wales for Case 3

Building Category	Sample Number
School	21728
Office	5877
University	5173
Clinic	2261
Dry Sports and Leisure Facility	2053
Cultural Activities	1374
Hospital	1295
Long Term Residential	1238
Emergency Service	975
General Accommodation	914
Covered Car Park	493
Entertainment Halls	330
Swimming Pool	315
Workshop	298
Fitness and Health Centre	227
Restaurant	179
High Street Agency	121
Bar/Pub/Club	121
Retail	74
Storage Facility	56
Public Waiting	39
Cold Storage	37
Hotel	20
Public Buildings with Light Usage	15
Food Store	7

Chicago Energy Benchmarking for Case 2

Building Category	Sample Number	Building Category	Sample Number
Multifamily Housing	1342	Museum	6
K-12 School	398	Enclosed Mall	6
Office	322	Library	6
Hotel	84	Other - Mall	6
College/University	76	Performing Arts	5
Senior Care Community	62	Automobile Dealership	5
Retail Store	58	Financial Office	5
Supermarket/Grocery Store	35	Other - Entertainment/Public Assembly	5
Mixed Use Property	33	Other - Specialty Hospital	4
Hospital	27	Meeting	4
Residence Hall/Dormitory	24	Movie Theater	3
Strip Mall	23	Other - Education	3
Other - Recreation	17	Bank Branch	3
Medical Office	15	Adult Education	3
Other	13	Other - Lodging/Residential	2
Residential Care Facility	11	Outpatient Rehabilitation/Physical Therapy	2
Laboratory	10	Pre-school/Daycare	2
Worship Facility	10	Courthouse	2
Sale Center	9	Prison/Incarceration	2
Indoor Sports	9	Service	2

Appendix B. Input Variables

Case 1 Inputs

Categorical Variable	
Name	Encoding Method
Council District	Mean Encoding
Property Use Type	Mean Encoding
Land Use	Mean Encoding
Numerical Variable	
Name	Range
ln(GrossFloorArea (ft^2))	10~16
Building Age	0~120 (years)
Occupancy Rate	0~100%
ln(Assessed Land Value (dollars))	0~22
ln(Assessed Total Value(dollars))	0~23
Number of floors	0~90
ln(Commercial Area (ft^2))	0~17
ln(Residential Area(ft^2))	0~16
ln(Lot Area (ft^2))	7~20
ln(Building front (feet))	0~9
ln(Building depth (feet))	0~10

Case 2 Inputs

Categorical Variable	
Name	Encoding Method
Natural Gas Usage/not	One Hot
Community Area	Mean Encoding
Primary Property Type	Mean Encoding
Numerical Variable	
Name	Range
ln(Gross Floor Area (ft^2))	10~17
Number of Buildings	1~236
Building Age	0~157 (years)

Case 3 Inputs

Categorical Variable	
Name	Encoding Method
Building Type	Mean Encoding
City	Mean Encoding
Area	Mean Encoding
With/without AC	one hot
HVAC Type	one hot
AC Inspection	one hot
Occupancy Type	one hot
Fuel Type	one hot
Numerical Variable	
Name	Range
AC Power	0~12044 (kW)
ln(Total Floor Area (m^2))	5.5~13
Renewable for Electricity	0.9~100%
Grid Supplied Electricity (proportion)	0~100%
District Heating (proportion)	0~100%
Renewable for Thermal (proportion)	0~100%
Renewable for Electricity (proportion)	0~100%

Appendix C. Results of Kolmogorov-Smirnov Test

Kolmogorov-Smirnov Test Results of Case 1

Building Category	statistic	pvalue	Identical Distribution?
Utility	0.0134	1.0000	yes
Hospital	0.0464	0.7928	yes
Worship Facility	0.0683	0.5726	yes
Recreation	0.0490	0.9224	yes
Service	0.0594	0.8212	yes
Meeting	0.0917	0.6049	yes
Indoor Sports	0.0574	0.9640	yes
Laboratory	0.1417	0.6960	yes
Restaurant	0.1250	0.8246	yes
Prison	0.1560	0.6497	yes
Personal Service	0.3417	0.4020	yes
Ice/Curling Rink	0.3550	0.4586	yes
Zoo	0.3367	0.5273	yes
Data Center	0.3600	0.4407	yes
Stadium (Open)	0.2167	0.9342	yes

Kolmogorov-Smirnov Test Results of Case 3

Building Category	statistic	pvalue	Identical Distribution?
Swimming Pool	0.0289	0.9977	yes
Workshop	0.0225	1.0000	yes
Fitness and Health Centre	0.0343	0.9946	yes
Restaurant	0.0553	0.8475	yes
Bar/Pub/Club	0.0614	0.8785	yes
High Street Agency	0.0845	0.5409	yes
Retail	0.0584	0.9775	yes
Storage Facility	0.0927	0.7866	yes
Public Waiting	0.0836	0.9503	yes
Cold Storage	0.0747	0.9854	yes
Hotel	0.1300	0.8687	yes
Public Buildings With Light Usage	0.2108	0.4782	yes
Food Store	0.3283	0.4495	yes

Kolmogorov-Smirnov Test Results of Case 2

Building Category	statistic	pvalue	Identical Distribution?
Residential Care Facility	0.2303	0.7000	yes
Strip Mall	0.0725	1.0000	yes
Other - Recreation	0.1137	0.9936	yes
Medical Office	0.1000	1.0000	yes
Other	0.1949	0.8074	yes
Residence Hall/Dormitory	0.0417	1.0000	yes
Laboratory	0.2667	0.6264	yes
Worship Facility	0.1333	0.9988	yes
Sale Center	0.1222	0.9995	yes
Indoor Sports	0.3000	0.4829	yes
Museum	0.1333	1.0000	yes
Enclosed Mall	0.3333	0.6652	yes
Library	0.3333	0.6652	yes
Other - Mall	0.1333	1.0000	yes
Performing Arts	0.3333	0.6652	yes
Automobile Dealership	0.2000	0.9904	yes
Financial Office	0.2667	0.8782	yes
Other - Entertainment/Public Assembly	0.3333	0.6652	yes
Other - Specialty Hospital	0.2667	0.8782	yes
Meeting	0.3333	0.6652	yes
Movie Theater	0.2667	0.8782	yes
Other - Education	0.2667	0.8782	yes
Bank Branch	0.2333	0.9482	yes
Adult Education	0.2000	0.9904	yes
Other - Lodging/Residential	0.2000	0.9904	yes
Outpatient Rehabilitation/Physical Therapy	0.2000	0.9904	yes
Pre-school/Daycare	0.2000	0.9904	yes
Courthouse	0.2000	0.9904	yes
Prison/Incarceration	0.2000	0.9904	yes
Service	0.2000	0.9904	yes

Appendix D. Score Tables of Case 1 and Case 2

Score Table of Case 1

EER	Score
0.630	10
0.790	20
0.883	30
0.949	40
1.000	50
1.051	60
1.117	70
1.211	80
1.899	90

Score Table of Case 2

EER	Score
0.694	10
0.826	20
0.902	30
0.957	40
1.000	50
1.042	60
1.096	70
1.173	80
1.304	90

References

[1] F. Johari, G. Peronato, P. Sadeghian, X. Zhao,J. Widén. Urban Building Energy Modeling: State of the Art and Future Prospects. Renewable and Sustainable Energy Reviews **128** (2020) 109902.

[2] Z. Wei, W. Xu, D. Wang, L. Li, L. Niu, W. Wang, B. Wang,Y. Song. A Study of City-Level Building Energy Efficiency Benchmarking System for China. Energy and Buildings **179** (2018) 1-14.

[3] P. Arjunan, K. Poolla,C. Miller. Energystar++: Towards More Accurate and Explanatory Building Energy Benchmarking. Applied Energy **276** (2020) 115413.

[4] F. C, Apec Workshop on Energy Intensity Reduction in the Apec Regions, in Asia-Pacific Economic Cooperation (Apec) Workshop, K.S. Wong, Editor. 2021.

[5] Energy Star: Comparison of Energy Consumption Level Based on Actual Building Operation, https://www.energystar.gov/buildings/benchmark?testEnv=false.

[6] S. Papadopoulos,C.E. Kontokosta. Grading Buildings on Energy Performance Using City Benchmarking Data. Applied Energy **233** (2019) 244-253.

[7] B. Bordass. Energy Performance in Use: And Government Policy. Retrofit for Purpose: RIBA Publishing; 2019, p. 13-21.

[8] Energy Performance of Buildings Directive (Epbd), http://www.estif.org/policies/epbd0/.

[9] Z. Yang, J. Roth,R.K. Jain. Due-B: Data-Driven Urban Energy Benchmarking of Buildings Using Recursive Partitioning and Stochastic Frontier Analysis. Energy and Buildings **163** (2018) 58-69.

[10] Energy Performance of Buildings Data: England and Wales, https://epc.opendatacommunities.org/.

[11] C. Robinson, B. Dilkina, J. Hubbs, W. Zhang, S. Guhathakurta, M.A. Brown,R.M. Pendyala. Machine Learning Approaches for Estimating Commercial Building Energy Consumption. Applied energy **208** (2017) 889-904.

[12] J. Ma,J.C.P. Cheng. Identifying the Influential Features on the Regional Energy Use Intensity of Residential Buildings Based on Random Forests. Applied energy **183** (2016) 193-201.

[13] D. Hsu. Identifying Key Variables and Interactions in Statistical Models of Building Energy Consumption Using Regularization. Energy **83** (2015) 144-155.

[14] P. Torres, M. Blackhurst,N. Bouhou. Cross Comparison of Empirical and Simulated Models for Calculating Residential Electricity Consumption. Energy and Buildings **102** (2015) 163-171.

[15] T. Walter,M.D. Sohn. A Regression-Based Approach to Estimating Retrofit Savings Using the Building Performance Database. Applied energy **179** (2016) 996-1005.

[16] W. Zhang, S. Guhathakurta, R. Pendyala, V. Garikapati,C. Ross. A Generalizable Method for Estimating Household Energy by Neighborhoods in Us Urban Regions. Energy Procedia **143** (2017) 859-864.

[17] W. Zhang, C. Robinson, S. Guhathakurta, V.M. Garikapati, B. Dilkina, M.A. Brown,R.M. Pendyala. Estimating Residential Energy Consumption in Metropolitan Areas: A Microsimulation Approach. Energy **155** (2018) 162-173.

[18] F.R. Cecconi, N. Moretti,L.C. Tagliabue. Application of Artificial Neutral Network and Geographic Information System to Evaluate Retrofit Potential in Public School Buildings. Renewable and Sustainable Energy Reviews **110** (2019) 266-277.
[19] J.A. Fonseca, I. Nevat,G.W. Peters. Hybrid and Multi-Scale Modelling of the Energy Demand of the Building Stock of the United States. ETH Zurich Research Collection (2019).
[20] U. Ali, M.H. Shamsi, M. Bohacek, K. Purcell, C. Hoare, E. Mangina,J. O'Donnell. A Data-Driven Approach for Multi-Scale Gis-Based Building Energy Modeling for Analysis, Planning and Support Decision Making. Applied Energy **279** (2020) 115834.
[21] J. Morris, J. Harrison, D. Allinson,K. Lomas. Towards Benchmarking English Residential Gas Consumption. in XXXIII International Conference on Energy Efficiency and Renewable Energy. 2012. Paris: Loughborough University.
[22] D.E. Marasco,C.E. Kontokosta. Applications of Machine Learning Methods to Identifying and Predicting Building Retrofit Opportunities. Energy and Buildings **128** (2016) 431-441.
[23] C. Anderson. The Long Tail: Why the Future of Business Is Selling Less of More. Book The Long Tail: Why the Future of Business Is Selling Less of More: Hachette Books; 2006.
[24] A. Li, F. Xiao, C. Zhang,C. Fan. Attention-Based Interpretable Neural Network for Building Cooling Load Prediction. Applied Energy **299** (2021) 117238.
[25] Y. Pan,L. Zhang. Data-Driven Estimation of Building Energy Consumption with Multi-Source Heterogeneous Data. Applied Energy **268** (2020) 114965.
[26] J. Roth, B. Lim, R.K. Jain,D. Grueneich. Examining the Feasibility of Using Open Data to Benchmark Building Energy Usage in Cities: A Data Science and Policy Perspective. Energy Policy **139** (2020) 111327.
[27] V. Gudivada, A. Apon,J. Ding. Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations. International Journal on Advances in Software **10**(1) (2017) 1-20.
[28] M. Kuhn,K. Johnson. Applied Predictive Modeling. Book Applied Predictive Modeling: Springer; 2013.
[29] A. Fernández, S. García, M. Galar, R.C. Prati, B. Krawczyk,F. Herrera. Learning from Imbalanced Data Sets. Book Learning from Imbalanced Data Sets: Springer; 2018.
[30] D. Micci-Barreca. A Preprocessing Scheme for High-Cardinality Categorical Attributes in Classification and Prediction Problems. ACM SIGKDD Explorations Newsletter **3**(1) (2001) 27-32.
[31] P. Rodríguez, M.A. Bautista, J. Gonzalez,S. Escalera. Beyond One-Hot Encoding: Lower Dimensional Target Embedding. Image and Vision Computing **75** (2018) 21-31.
[32] P. Cerda, G. Varoquaux,B. Kégl. Similarity Encoding for Learning with Dirty Categorical Variables. Machine Learning **107**(8) (2018) 1477-1494.
[33] M. Lokhandwala,R. Nateghi. Leveraging Advanced Predictive Analytics to Assess Commercial Cooling Load in the Us. Sustainable Production and Consumption **14** (2018) 66-81.
[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss,V. Dubourg. Scikit-Learn: Machine Learning in Python. the Journal of machine Learning research **12** (2011) 2825-2830.
[35] T.K. Ho. Random Decision Forests. in 3rd International Conference on Document Analysis and Recognition. 1995. IEEE.

[36] Y. Freund, R. Schapire,N. Abe. A Short Introduction to Boosting. Journal-Japanese Society For Artificial Intelligence **14**(771-780) (1999) 1612.
[37] G.M. Weiss, K. McCarthy,B. Zabar. Cost-Sensitive Learning Vs. Sampling: Which Is Best for Handling Unbalanced Classes with Unequal Error Costs? Dmin **7**(35-41) (2007) 24.
[38] C. Shorten,T.M. Khoshgoftaar. A Survey on Image Data Augmentation for Deep Learning. Journal of Big Data **6**(1) (2019) 1-48.
[39] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville,Y. Bengio. Generative Adversarial Nets. Advances in neural information processing systems **27** (2014).
[40] N.V. Chawla, K.W. Bowyer, L.O. Hall,W.P. Kegelmeyer. Smote: Synthetic Minority over-Sampling Technique. Journal of artificial intelligence research **16** (2002) 321-357.
[41] M. Stone. Cross‐Validatory Choice and Assessment of Statistical Predictions. Journal of the royal statistical society: Series B (Methodological) **36**(2) (1974) 111-133.
[42] S.A. Glantz, B.K. Slinker,T.B. Neilands. Primer of Applied Regression and Analysis of Variance. Mcgraw-Hill. Inc., New York (1990).
[43] Seaborn Distplot, https://seaborn.pydata.org/generated/seaborn.distplot.html.
[44] M.T. Ribeiro, S. Singh,C. Guestrin. " Why Should I Trust You?" Explaining the Predictions of Any Classifier. in 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.
[45] C. Molnar. Interpretable Machine Learning.8.2 Local Surrogate (LIME): Lulu. com; 2020.
[46] C. Molnar. Interpretable Machine Learning. A Guide for Making Black Bos Models Explainable,5.72020.
[47] Energy and Water Data Disclosure for Local Law 84 2020 (Data for Calendar Year 2019), https://data.cityofnewyork.us/Environment/Energy-and-Water-Data-Disclosure-for-Local-Law-84-/qb3v-bbre.
[48] Pluto and Mappluto, https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page.
[49] Chicago Energy Benchmarking, https://data.cityofchicago.org/Environment-Sustainable-Development/Chicago-Energy-Benchmarking/xq83-jr8c.
[50] C. SciPy. Scipy 1.0: Fundamental Algorithms for Scientific Computing in Python'. Nature (2020).
[51] J. Roth, A. Martin, C. Miller,R.K. Jain. Syncity: Using Open Data to Create a Synthetic City of Hourly Building Energy Estimates by Integrating Data-Driven and Physics-Based Methods. Applied Energy **280** (2020) 115981.
[52] J. Benesty, J. Chen, Y. Huang,I. Cohen. Pearson Correlation Coefficient. Noise Reduction in Speech Processing: Springer; 2009, p. 1-4.