BEEM: Data-driven Building Energy bEnchMarking for Singapore

Pandarasamy Arjunan[a], Kameshwar Poolla[b], Clayton Miller[c]

[a]*Berkeley Education Alliance for Research in Singapore, Singapore*
[b]*University of California, Berkeley, United States*
[c]*Building and Urban Data Science (BUDS) Lab, National University of Singapore (NUS), Singapore*

Abstract

Building energy use benchmarking is the process of measuring the energy performance of buildings, relative to their peer group, for creating awareness and identifying energy-saving opportunities. In this paper, we present the design and implementation of *BEEM*, a data-driven energy use benchmarking system for buildings in Singapore. The peer groups for comparison are established using a public energy disclosure data set. We use an ensemble tree algorithm for accurately modeling building energy use and for identifying the most influential factors. Our models reduce the prediction error from 24.39% to 6.04%, on average, when compared to the baseline linear regression models, which was used in the previous energy efficiency labeling program in Singapore, and outperforms ten other recent models. Using the prototype implementation of BEEM, we benchmarked three building types, office (290), hotel (203), and retail (125) and compared their rating. The code repository and the accompanying data set are released as an open-source repository for the community use.

Keywords: Building energy benchmarking, building energy labeling, Regression analysis, Gradient boosting trees, Feature interaction, Interpretable machine learning

Email addresses: samy@bears-berkeley.sg (Pandarasamy Arjunan), poolla@berkeley.edu (Kameshwar Poolla), clayton@nus.edu.sg (Clayton Miller)

1. Introduction

Commercial buildings account for approximately one-third of global energy consumption and greenhouse gas emissions. Several government agencies and policymakers have started to implement energy benchmarking programs as one of the approaches for improving building energy efficiency. Benchmarking is the process of measuring the energy performance of a building with an established peer group. It helps in creating awareness, identifying energy-saving opportunities, and prioritizing energy management action plans. Realizing the potential of energy benchmarking, cities worldwide started to benchmark their building stock and reported 3%–8% reductions in energy consumption [1].

Singapore is a tropical city-state country where air-conditioning is required through out the year due to the tropical climate. As part of the *Green Building Masterplan* [2], Singapore has implemented several programs and measures toward reducing the energy footprint of its building stock. The Building and Construction Authority (BCA) has introduced the *Green Mark* scheme [3], a point-based rating system similar to LEED [4], that assesses the environmental sustainability levels of buildings and assign grades. The BCA also benchmarks the country's building stock using a simple Energy Performance Indicator (EPI) called Energy Use Intensity (EUI).

1.1. Previous work in data-driven benchmarking methods

Every building is different in terms of its physical and operational characteristics, such as size, age, geometry, occupancy, schedule, and appliance usage. Building's energy use is influenced by these multitudes of building attributes and their interactions in a complex way. In addition to this, meteorological conditions, such as air temperature and relative humidity, also affect energy usage significantly. Hence, it is essential to normalize the building energy usage for all influential factors to make fair comparisons between buildings. Many approaches have been proposed in the literature for normalizing building energy usage. EUI is a commonly used measure which is expressed as energy usage per unit area, e.g. kWh/m^2. It is easy to compute and interpret EUI and it has been used in many benchmarking studies. However, EUI is an unfair metric because it overlooks other influential factors occupancy, operational hours, and other building attributes. There are other

metrics that are specific to different building types such as energy usage per worker for offices and energy usage per bed for hotels. However, all these metrics have similar limitations as EUI.

In response to the limitations of EUI, data-driven predictive models have been adopted. Unlike EUI, these predictive models can account for multiple influential factors thus enables a fair benchmarking system. Multiple Linear Regression (MLR) models have been widely used in several benchmarking systems. This include the *Energy Star* system for the USA and Canada, Singapore [5, 6], China [7, 8], South Korea [9, 10], and Taiwan [11]. MLR models finds a linear fit between building attributes and energy use. The primary advantage of MLR models is the ease of interpretation of model coefficients due to their linear and additive properties. However, such linear models are inadequate in modeling the complex relationships between energy use and building attributes, which is often non-linear. Due to this, MLR models are often found to be a poor performer, as reported in many studies [12]. In order to develop a fair benchmarking system, the underlying energy use prediction model needs to be highly accurate by reflecting all combination of relationships and their interactions between the building attributes and energy use.

The recent studies have adopted non-linear models for energy benchmarking. These nonlinear models are proven to achieve better performance in terms of prediction accuracy when compared with the linear models [12, 13]. However, due to their complex nature, it is difficult to interpret the predictions of these nonlinear models out of the box unlike the linear models, e.g., which factors contributed to prediction results for each building. Furthermore, unsupervised clustering methods have also been used to group the buildings based on their similarity in energy use and building attributes [14]. Other contemporary benchmarking approaches have also used the econometric-based Stochastic Frontier Analysis [15] and Data Envelopment Analysis [16]. However, these approaches are sensitive to outliers, and they also lack model interpretability. A comparison data-driven energy benchmarking approaches from around the world are shown in Table 1.

Ref.	Location	Building types	No. of buildings	No. of attributes	Algorithm	Rating
[17]	EnergyStar, USA	16 types		6 to 8	MLR	Point (1-100)
[18]	USA	Office	242	5	MLR and RF	
[7]	China	Office	88	10	MLR	Point (1-100)
[9]	South Korea	Office	1,072	11	DT and ANOVA	5 Grades (A-E)
[11]	Taiwan	Hotel	45	6	MLR	NA
[8]	China	Campus buildings	13	11	MLR	5 Grade (A-E)
[19]	Hong Kong	Office	30	5	MLR	
[20]	Taiwan	School and universities	74	4	MLR	
[21]	Brazil	Bank branches	1,890		MLR	
[22]	Taiwan	Office	47		DEA	Point (1-100)
[23]	Ireland	Primary school			Energy Plus	7 Grades (A-G)
[24]	Greece	School	320		Fuzzy clustering	5 Grades (A-E)
[25]	Hawaii	Office and classrooms	60	10	ANN	
[26]	UK	School	7,700	23	ANN	
[27]	Greece	Hotel	90		k-means	5 Grades
[28]	Italy	Healthcare center	100	11	LMEM and CART	
[6]	Singapore	Hotel	29	3	MLR	Point (1-100)
[5]	Singapore	Office				

Table 1: Summary of data-driven energy benchmarking methods globally

1.2. Need for a holistic energy benchmarking system for Singapore

Despite initiatives in other contexts, Singapore has not had a systematic, active benchmarking system for assessing the energy performance of buildings for over ten years. There are limited energy benchmarking studies specific to the Singapore context. More than a decade ago, a benchmarking approach for office buildings was presented in [29, 5]. After normalizing energy usage using regression analysis, buildings whose energy efficiency falls within the nation's top 25% are considered as the most efficient buildings. A subset of these buildings that meet specific physical and occupancy criteria was awarded *Energy Smart Office* label. In another study [6], Singapore hotels were benchmarked using the same approach. Both the studies have used linear regression models on a data set of fewer than 100 buildings for each building type. More recent research is focused on identifying key factors influencing the energy usage in 56 air-conditioned office buildings using clustering method [30]. While there are sophisticated benchmarking systems in other countries, such as the *ENERGY STAR Portfolio Manager* [17] for the USA and Canada, they may not be directly applicable to Singapore's tropical climate and vast differences in the energy and appliance usage patterns. For example, decentralized split air-conditioning systems are common, and they are required around the year in Singapore buildings.

In this paper, we present BEEM, a data-driven building energy benchmarking system for Singapore. BEEM consists of four parts: (1) peer groups are established based on building type (e.g., office and retail), and data are extensively cleaned for further refining them, (2) highly accurate prediction models are developed, using the CatBoost [31] algorithm, for modeling peer group's energy usage based on various building attributes, (3) the relative energy performance level of each building is calculated and mapped to a five-point-scale letter grade using an univariate clustering algorithm for ease of understanding, and (4) finally, the visual explanation for individual model predictions is presented by leveraging recent Explainable Artificial Intelligence approaches.

In addition to the structure outlined, we also compare the performance of the prediction model used in our system to a linear regression model that has been used in existing benchmarking systems, including *Energy Star* and a former Singaporean labeling system [5, 6] and various state of the art models. Our models achieve significant improvement over baseline and reduce error from 24.39% to 6.04%, on average.

The major contributions of this work are:

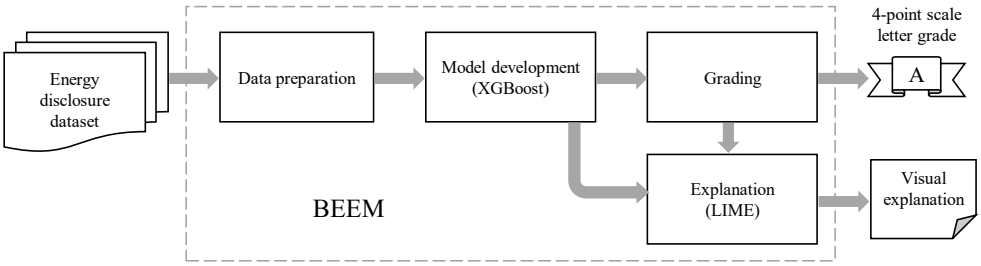

Figure 1: An overview of BEEM benchmarking system for Singapore buildings

- The *Data-driven Building Energy bEnchMarking (BEEM)* system is developed as energy benchmarking approach for the Singapore buildings that utilizes advanced modelling and context-specific features for the Singapore context.
- BEEM is systematically assessed through the interactions of building attributes. This effort shows that models with interaction effects achieve better accuracy than a baseline model with only main effects.
- A reproducible code repository and the accompanying dataset is released as an open source for the community use and deployment.

This paper is organized as follows. In Section 2, the proposed BEEM benchmarking system is outlined in detail in the context of application to a group of buildings in the Singapore context. Section 3 outlined the results of an implementation on a sample data set of buildings and discusses the strengths and drawbacks of the approach. Section 3 outlines a comparison of the proposed method as compared to other building benchmarking systems in the built environment and discusses the advantages of machine learning explainability. Finally, in Section 4 there is an overview of limitations, future work, and conclusions for the BEEM system.

2. Methodology

The methodology developed for the BEEM system builds upon previous work focused on the context of North America [32]. This work focused on the development of a novel modeling and explainable machine learning workflow to build upon the EnergyStar rating system. In comparison with [32], the proposed BEEM system uses a more accurate and robust gradient boosting model called CatBoost (See Section 2.2.1). This model is combined with a

Table 2: List of building attributes available from the BCA 2017 data set

S.No	Variable name	Description
1	AirconFA	Total air-conditioned floor area (m^2)
2	NonAirconFA	Total non air-conditioned floor area (m^2)
3	Age	Age of the building
4	IsPublic	Is public sector building? (Yes/No)
5	Occupancy	Average monthly occupancy rate (%)
6	AirconType	Type of air-conditioning system such as:
		Water-cooled chilled water plant
		Air-cooled chilled water plant
		District cooling plant
		Split units or unitary systems
7	AirconAge	Age of the air-conditioning system
8	AirconEff	Air-conditioning system efficiency (kW/RT)
9	LED	LED light usage (%)
10	Rooms	Number of rooms (only for hotels)

novel model agnostic XAI technique called LIME [33] that provides grade explanation to each building. Both these features were not studied together, to the best of our knowledge, in the context of energy benchmarking, especially to the Singaporean buildings. Furthermore, we also present the design and implementation of an end-to-end benchmarking system including an user interface application. This section outlines the BEEM methodology in four steps: (1) data preparation, (2) model development, (3) grading, and (4) explanation of grade. An overview of this process is shown in Figure 1.

2.1. Building peer group data preparation

The first phase of the method involves defining the peer group and performing data cleaning. A peer group is a group of buildings with similar operational characteristics, for example, office or hotel. We use energy disclosure data set released by the Building and Construction Authority (BCA) for the year 2017 for defining peer group samples [34]. BCA publishes the Building Energy Benchmarking Report (BEBR) annually since 2014, to monitor the building energy performance of Singapore's building stock. Under the Building Control Act, building owners have been required to submit building related information and energy consumption data to BCA on an annual basis since 2013. The information thus collected was analysed to establish the national building energy benchmarks for Singapore's built environment.

Table 3: List of analytical filters applied during the data cleaning process.

Building type	Attribute	Filters (acceptable range)
Office	GFA	Between 1000 and 80,000 m^2
	EUI	Less than 500 $kWh/m^2/year$
	ChillerType	Not using *Water Cooled Packaged Unit*
	Age	Less than 100 years
Hotel	GFA	Between 100 and 60,000 m^2
	EUI	Less than 750 $kWh/m^2/year$
	ChillerType	Not using *Water Cooled Packaged Unit*
	Age	Less than 100 years
	Rooms	Less than 600 rooms
Retail	GFA	Between 1000 and 80,000 m^2
	EUI	Less than 1,000 $kWh/m^2/year$
	ChillerType	Not using *Water Cooled Packaged Unit*
	Age	Less than 100 years

This data set contains detailed building attributes and energy use information of 1145 buildings. We split this data set into different groups based on the building type. Though there are six building types in this data set, we selected only office, hotel, and retail buildings. Other building types, such as hospital, had very few samples or too many missing values in building attributes. After selecting building samples, we carefully cleaned the building attributes in each peer group. Specifically, we removed outliers and inconsistent samples based on data distribution and statistical measures, similar to the analytical filters in the *Energy Star* system [17]. The list of filters that we applied during the data cleaning is summarized in Table 3.

We also created derived building attributes based on available data, such as air-conditioned and non-air-conditioned floor area and age of the building based on Temporary Occupation Permit/Certificate Of Statutory Completion. After cleaning, there are 290, 203, and 125 samples for office, hotel, and retail buildings, respectively.

The list of building attributes used in this study and their description are provided in Table 2. Further, the descriptive statistics of office buildings after cleaning is shown in Table 4. The histogram of EUI values of all three building types are shown in Figure 2. It is to be noted from Table 2 that the most of the existing building use benchmarking systems also used only the most significant building attributes (approximately 10) based on data

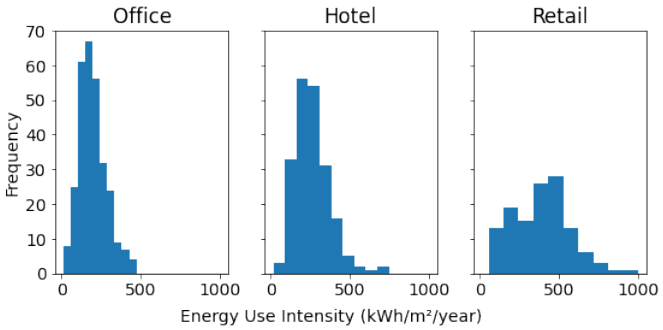

Figure 2: Histograms of Energy Use Intensity of office, hotel, and retail buildings after data cleaning.

availability. This is one of the limitations of data-driven energy benchmarking as collecting complete list of factors is practically difficult and doing so also will limit the usability of benchmarking.

Table 4: Descriptive statistics of the office buildings after data cleaning.

Attribute	Mean	STD	Min	25%	50%	75%	Max
AirconFA	13152.1	13565.9	0.0	2493.7	8838.0	18356.4	61153.0
NonAirconFA	3596.7	5073.4	0.0	455.0	1743.2	4503.6	32587.4
Age	24.1	16.7	1.0	15.0	21.0	31.0	118.0
Occupancy	89.7	15.1	5.0	85.0	95.0	100.0	100.0
ChillerAge	8.2	7.3	0.0	3.0	6.0	12.0	40.0
ChillerEff	0.8	0.2	0.5	0.6	0.7	0.9	1.6
LED	18.3	28.3	0.0	0.0	2.0	25.0	100.0

2.2. Benchmarking model development

After establishing peer group buildings for each building type, the next step is model development. In this step, the relationship between energy use and available building attributes of each peer group are fit into a model. This model will be used to get the estimated energy usage of a new building that needs to be benchmark. The key priority in this step is to find the model that is the most accurate in predicting the energy use of the building based on its attributes in order to ensure a fair comparison of buildings. We use all the available building attributes, as listed in Table 2, as the predictors of total energy usage (kWh). It is important to note that we use total energy usage as the dependent variable, instead of using EUI, because we approach also

provides explanation on which attributes influence the grades on individual buildings (See Sections 3.3 and 3.4).

2.2.1. Gradient boosting and CatBoost model

The primary model that is implemented in the BEEM system is CatBoost [31]. CatBoost is an efficient implementation of gradient boosting based on decision trees. We choose CatBoost model because it offers several advantages over other modelling techniques. Firstly, CatBoost models are more accurate and perform better than many contemporary non-linear methods such as XGBoost, LigtGBM, and deep learning techniques [31]. It was also included in many of the solutions of the ASHRAE Great Energy Predictor III competition held in 2019 [35]. Secondly, it uses symmetric or oblivious decision trees as the base predictor. This can reduce over-fitting thus more generalizable model. Moreover, the underlying decision trees are suitable for capturing higher-order feature interactions inherently, e.g., what is the combined effect of gross floor area and occupancy on energy use? Thirdly, CatBoost can handle both numerical and categorical data inherently. Since many building attributes in our peer-group dataset are categorical in nature, it is easy to include those attributes into the model, instead of using an explicit one-hot encoding technique. Catboost can also handle missing data inherently. These unique features also helps in model interpretation as to understand the significance of each building attribute on energy use. Finally, CatBoost is suitable for small datasets and works faster unlike the deep learning based models that require huge amount of training data and computing resources.

Let's consider the training dataset $\mathcal{D} = \{(\mathbf{x}_k, y_k)\}_{k=1..n}$, where $\mathbf{x}_k = (x_k^1, \ldots, x_k^m)$ is a vector of m features and $y_k \in \mathbb{R}$ is a target. The learning task involves training a function $F : \mathbb{R}^m \to \mathbb{R}$ which minimizes the expected loss function:

$$\mathcal{L}(F) := \mathbb{E}L(y, F(\mathbf{x})) \tag{1}$$

The gradient boosting procedure involves fitting a sequence of approximation functions $F^t : \mathbb{R}^m \to \mathbb{R}$, $t = 0, 1, \ldots$ sequentially. Each F^t is obtained from the previous approximation F^{t-1} and included into the a additive manner to the final model:

$$F^t = F^{t-1} + \alpha h^t \tag{2}$$

Where, α is a step size. h^t is the base predictor which is chosen from H, a family of functions, to minimize the expected loss $\mathcal{L}$:

$$h^t = \arg\min_{h \in H} \mathcal{L}\left(F^{t-1} + h\right) = \arg\min_{h \in H} \mathbb{E}L\left(y, F^{t-1}(\mathbf{x}) + h(\mathbf{x})\right) \tag{3}$$

This minimization problem is solved by taking a negative gradient step using least-square approximation.

$$h^t = \arg\min_{h \in H} \mathbb{E}\left(-g^t(\mathbf{x}, y) - h(\mathbf{x})\right)^2 \tag{4}$$

A CatBoost algorithm is an efficient implementation of gradient boosting based on oblivious decision trees. A oblivious decision or symmetric trees are balanced thus less prone to over fitting and helps in faster prediction. A decision tree is built recursively by splitting the features $\mathbb{R}^m$ into disjoint sets until some splitting criteria is met. An example tree is shown in Figure 3. A decision tree is formerly defined as:

$$h(\mathbf{x}) = \sum_{j=1}^{J} b_j \mathbb{1}_{\{\mathbf{x} \in R_j\}} \tag{5}$$

In addition, the CatBoost algorithm addresses two key limitations found in existing gradient boosting techniques such as XGBoost. Firstly, it proposes an *ordered target statistics* technique to handle categorical features efficiently.

A straight forward way

$$\hat{x}_k^i = \frac{\sum_{j=1}^{n} \mathbb{1}_{\left\{x_j^i = x_k^i\right\}} \cdot y_j + ap}{\sum_{j=1}^{n} \mathbb{1}_{\left\{x_j^i = x_k^i\right\}} + a} \tag{6}$$

Where, i is feature index, n is the number of instances, $\hat{x}_k^i$ is the expected target category, and a is the weight of p that is the prior which is usually set by average value of the dataset. Secondly, it recognizes and addresses the *prediction shift* problem which is caused due to target leakage. CatBoost addresses this issue by using a novel *ordering boosting* in which several random permutation of samples are used simultaneously to reduce the variance and each subsequent tree is built using unbiased samples from the previous trees. More details about the working of this algorithm can be found in [31].

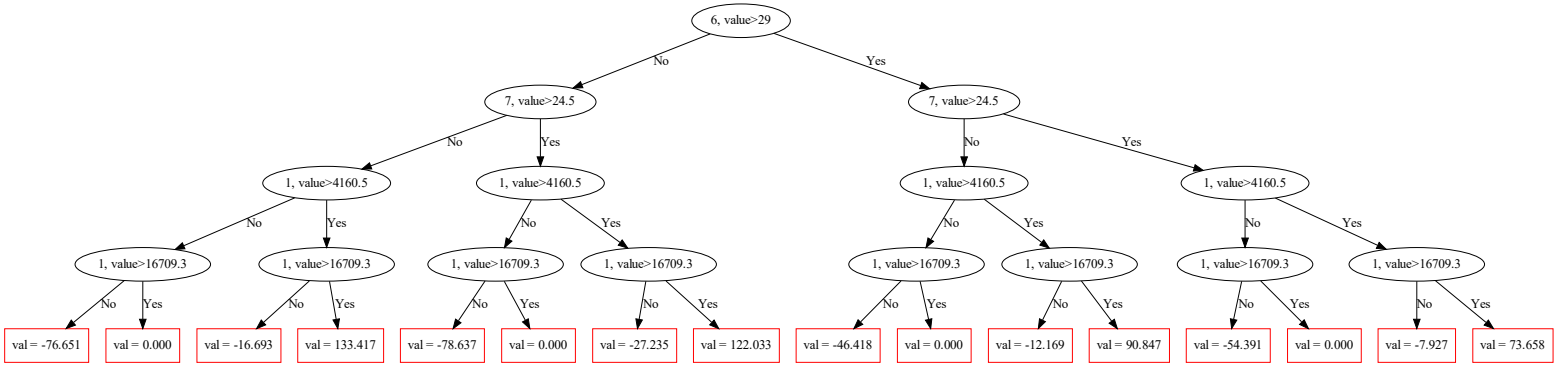

Figure 3: An example symmetric decision tree of a CatBoost model. This tree has a depth of four. Each node contains the condition based on a specific building attribute (shown as index) to branch to the next level.

Table 5: List of CatBoost parameters tuned using grid search.

Parameter	Description	Range / Step
iterations	The maximum number of trees	50 - 1000 / 50
depth	Depth of the tree	2 - 10 / 1
learning_rate	Learning rate	0.01 - 0.1 / 0.01
l2_leaf_reg	Coefficient at the L2-regularization term	1 - 10 / 1

2.2.2. Tuning hyper-parameters and model selection

CatBoost offers several parameters that can be tuned to select an optimal model. The grid search method is employed for selecting optimal parameter values using a 10-fold cross-validation approach. The four most important hyper-parameters chosen for turning are: (a) The maximum number of trees (*iterations*), (b) Depth of individual decision trees (*depth*), (c) Learning rate (*learning_rate*) used to reduce th gradient steps which also affects the training time, and (d) The coefficient of cost function's L2 regularization term (*l2_leaf_reg*). These hyper-parameter names and their corresponding search range is given in Table 5. All other parameters of the model were to set to their defaults.

2.3. Contemporary models

In addition to CatBoost, we also study and compare the performance of contemporary data-driven prediction models that are used in recent energy benchmarking studies [36, 32]. A brief explanation of those models are given below and their performances are compared in Section 3.2.

2.3.1. Multiple linear regression and variants

Multiple linear regression models have been widely used in many energy benchmarking studies. It fits linear approximation function between energy use (dependent variable) and building attributes (independent variables or predictors). An MLR is defined as:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j} + \epsilon_i, \quad i = 1, 2, ..., n \tag{7}$$

Here, Y_i is the dependent variable, β_0 is model's offset term, $X_{i,*}$ are a vector of p independent variables, β_j are weights or coefficients of the predictor variables, n is the total number of examples, and ϵ_i is the residual

or error term. The weights β_* are estimated using Ordinary Least Squares (OLS) [37]. The OLS method estimate the coefficients or weights for each predictor by minimizing a cost function. There are several variants of MLR models, such as ridge and lasso regression, to overcome the bias-variance problem in MLR models. In general, a regularization term will be added to the cost function.

In *ridge regression*, a penalty term, also called as L2-regularization, is included to the model that penalizes sum of squared coefficients of the predictors. The ridge regression model is defined as:

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{8}$$

Where, $\lambda > 0$ is the complexity parameter to control the penalty. When $\lambda = 0$, the model is similar to the OLS, whereas a larger λ lead to high penalty to the coefficients. This helps reduce the model's complexity and multicollinearity.

Whereas in *lasso regression*, or Least Absolute Shrinkage and Selection Operator, a penalty term, also called as L1-regularization is included to the model that penalizes sum of squared coefficients of the predictors. The lasso regression model is defined as:

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \tag{9}$$

Elastic Net is also variant of MLR in which the penalties of Ridge and Lasso regression are combined to get benefits of both. It is defined as:

$$\lambda \sum_{j=1}^{p} \left(\alpha\beta_j^2 + (1-\alpha)\left|\beta_j\right| \right) \tag{10}$$

Where, α is the controlling parameter to decide between ridge ($\alpha = 0$) and lasso ($\alpha = 1$) regression.

2.3.2. Ensemble models

In recent year, ensemble learning techniques gained wide acceptance due to their improved performance compared to the classical machine learning models. The main idea of ensemble learning is to build a more accurate and

generalized prediction model by combining the predictions of a collection of base models. These base models are generally simple weak learners built on a subset of the training samples. There are two ways to build and combine the predictions of base models: *Bagging* and *Boosting*. In bagging, also called as bootstrap aggregation, a homogeneous weak models are trained independently (often in parallel to each other). The predictions of all these models are averaged to make a final result. Bagging :

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{i=1}^{B} \hat{f}^{*i}(x) \tag{11}$$

Where, $\hat{f}^{*i}$ are the predictions from B bootstrap samples. Bagging helps reduce the variance of the model though in. Whereas in boosting, homogeneous weak models are trained sequentially in an additive manner and their predictions are combined by adding them. In this paper, three widely used ensemble models are studied and their performances are compared with the proposed CatBoost model. They are: (1) Random forest, (2) AdaBoost, and (3) XGBoost. All these techniques use decision tree or a variant as its base learner.

Random forest [38] is based on bagging with some modifications. A random forest model with B trees is built by drawing $Z^i, i = 1, 2, ..., B$ bootstrap samples each of size N from the training set. Then a random forest tree T_i is grown for each bootstrap sample independently. For the regression problems, predictions from all random forest trees T_i are averaged to make the final prediction for a new sample x.

$$\hat{f}_{\text{rf}}^{B}(x) = \frac{1}{B}\sum_{i=1}^{B} T_i(x) \tag{12}$$

AdaBoost [39], as the name implies, is based on boosting technique. In AdaBoost, base learners are trained on weighted versions of the training samples. Initially, the weight of each training sample (x_i, y_i) is set to $w_i = 1/N$. In the subsequent iterations, these weights are modified based on previous model's error, i.e., increase the weights of the samples that had high prediction error and decrease the weights of those samples with low error. This weight adjustment procedure forces the model to focus on those samples with high error in each subsequent step. The final model will have weighted collection of these base models that are trained in each step.

XGBoost [40] is another widely used ensemble model based on boosting. It uses the gradient decent based boosting to optimize the loss function and uses both L1 and L2 regularization terms to prevent model overfitting. An XGBoost model has a collection of Classification and Regression Trees (CART) [41]. A CART model is similar to normal decision tree except that leave nodes contain the real score. Similar to other boosting based models, the final prediction is made by summing the predictions of each CART model.

There are other nonlinear machine learning models exist, such as Support Vector Machine and Neural Networks, for energy use prediction. However, those models are excluded in this study because they are often found to be under performing than the ensemble models [42].

2.4. Grading

After selecting optimal models, the final step in benchmarking involves grading buildings based on their energy performance levels relative to the respective peer groups. The relative energy performance of a building, called Energy Efficiency Ratio (EER), is calculated as:

$$EER = \frac{Actual\ energy\ usage}{Expected\ energy\ usage} \tag{13}$$

Here, the expected energy usage is the model-predicted energy usage for the building that is similar to the peer group samples. An EER value less than 1 indicates that the building consumes less energy than the peer group, whereas an amount more than 1 suggests the building consumes more energy than the peer group. The EER values are calculated for each building. Next, we split the sorted EER list into five disjoint groups using an univariate clustering algorithm. The cluster boundaries are used to create a *grade lookup table* with grades from A to E. We have chosen five grades in this work inline with recent energy benchmarking systems [9, 8, 24].

2.5. Grade explanation with LIME

After assigning grades to each building based on their energy efficiency ratio, the proposed BEEM system provides explanations of those grades, e.g., which factors make the building energy efficient or inefficient. In this work, our CatBoost model is augmented with Explainable Artificial Intelligence (XAI) techniques. Particularly, a model agnostic method called Local Interpretable Model-agnostic Explanations (LIME) [33] is used in the BEEM

system. Unlike the MLR models, in which the coefficients denote the influence of each factor on an average on energy use of all samples, LIME can provide explantation for individual predictions. With LIME in place, facility managers can receive the understand which factors are dominant and influencing high or low energy use.

The *explanation* provided by LIME for data instance x is defined as:

$$\text{explanation}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{14}$$

where $f()$ is the original model, $g()$ is the local explanation for instance x that minimizes the loss function L that measures the fidelity between $f()$ and $g()$, while keeping the model complexity $\omega(g)$ low. π_{x_*} denotes a neighborhood of $\underline{x}_*$ in which approximation is sought. $g()$ belongs to a class of interpretable models, $\mathcal{G}$, such as linear models or decision trees [33].

LIME is an example of *sparse explainer* that is suitable for interpreting machine learning models with a large number of predictors. The important idea behind LIME is to train a local surrogate, interpretable model that approximates the predictions of the underlying black-box model. Since LIME is a readily available off-the-shelf tool, we decided to use it in our experiments.

3. Results

3.1. Model implementation and benchmarking grade calculation

We implemented the proposed benchmarking system and a web interface in *R* and *Shiny*. We used the *scikit-learn* [43] for developing the baseline prediction models and *catboost.ai* [44] for developing CatBoost models, and for tuning hyper-parameters and cross-validation. We used the python-based LIME library[1] to explain which factors influence the energy usage in each building. We also release the code repository and the accompanying dataset in open source[2] for community usage. Using our implementation, we benchmarked all buildings in our public dataset. The predicted energy usage of each building was calculated from the model trained on all peer group samples except the current one. Table 6 shows the final grade lookup table for office, hotel and retail building types that are used in this study. In Figure 4, we compare EER value range for different grades. Out of 290, 23 offices

[1] `https://github.com/marcotcr/lime`

[2] `https://github.com/samy101/BEEM/`

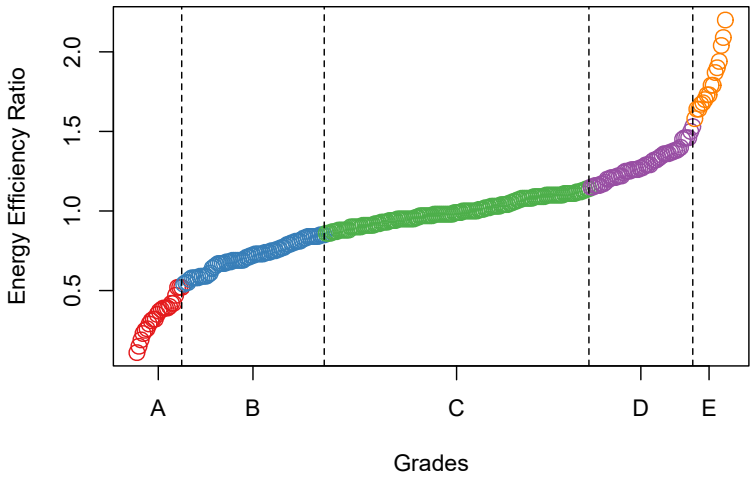

Figure 4: The EER range for different grades in the grade lookup table for office buildings. The X-axis denotes grade labels while the y-axis is Energy Efficiency Ratio. The dashed vertical bar denotes the boundary between different grades.

Table 6: The grade lookup table for office, hotel and retail buildings.

Grade	Office	Hotel	Retail
A	< 0.52	< 0.82	< 0.64
B	$0.53 - 0.85$	$0.83 - 1.06$	$0.64 - 0.93$
C	$0.85 - 1.14$	$1.07 - 1.32$	$0.94 - 1.12$
D	$1.15 - 1.58$	$1.33 - 1.76$	$1.12 - 1.63$
E	> 1.58	> 1.76	> 1.63

were assigned with grade A. These offices are considered to be the most energy efficient because their EER values are much lower (< 0.52) than other buildings in the peer group.

3.2. Comparison of model performance

It is essential to use an accurate model to develop a robust and fair benchmarking system. The robustness and fairness of a benchmarking system lie in the underlying predictive model's accuracy. We implemented all MLR and ensemble models, as described in the Section 2.2 for all three building types using the same set of building attributes listed in Table 2. We use Leave-One-Out Cross-Validation (LOOCV) procedure to validate the model

Table 7: Comparison of symmetric Mean Absolute Percentage Error (%) between various state of the art models and CatBoost.

Model	Hotel	Office	Retail	Average
CatBoost	**7.21**	**10.50**	**4.93**	**7.55**
RandomForest	18.34	18.84	21.65	19.61
Bagging	18.21	18.98	21.59	19.59
GradientBoosting	19.39	19.00	24.35	20.91
BayesianRidge	19.43	20.46	24.33	21.41
LassoLars	25.58	21.06	29.68	25.44
Ridge	25.77	21.06	29.63	25.49
Lasso	25.78	21.08	29.70	25.52
Linear	25.72	21.14	29.72	25.53
ElasticNet	25.96	21.19	30.42	25.86
AdaBoost	32.17	27.85	26.79	28.94

performance. LOOCV is a more robust approach because one model is developed and validated for each sample, using all other samples as training set. We use a scale-independent measure called symmetric Mean Absolute Percentage Error (SMAPE) to compare model performances. It is defined as:

$$sMAPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|P_i - A_i|}{(|A_i| + |P_i|)/2} \tag{15}$$

Where, A_i and P_i are the actual and predicted values. We compare the $sMAPE$ of proposed CatBoost and other models for three building types in Table 7. MLR models have also been used in earlier studies on benchmarking office and hotel buildings in Singapore [5, 6]. We observed that CatBoost models for all three buildings performed better than other models by achieving lowest $sMAPE$ of 7.55%.

3.3. Important building attributes

Next, we analyze how much impact each building attribute has on energy usage in our CatBoost. The feature importance plot for all three building types are shown and compared in Figure 5. We can observe that air-conditioned floor area ($AirconFA$) is the most dominant influencing attribute on energy use followed by non air-conditioned floor area ($NonAirconFA$). The combined floor area account for close to 80% influence of energy use.

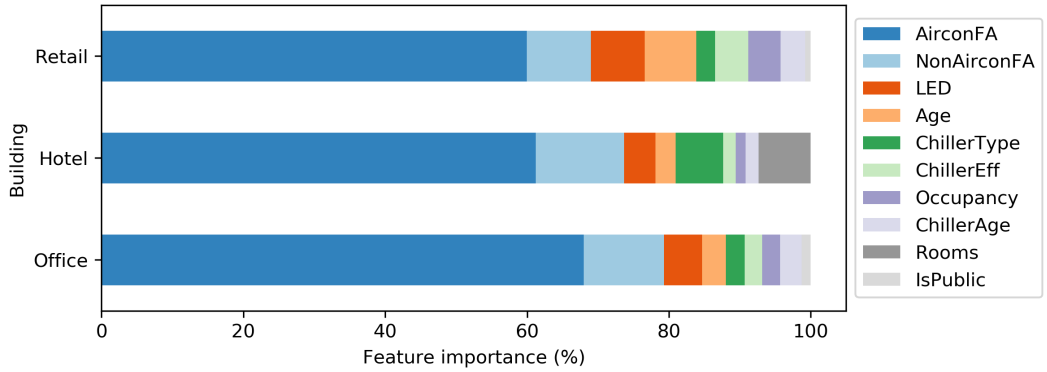

Figure 5: Feature importance using CatBoost models.

Moreover, it is interesting to note that the influence of floor area is high in office building followed by hotel and then retail buildings. It is very common that office building will have more air-conditioned space for workers than Hotel and Retail. The next dominant building attribute is percentage of LED lights usage followed by the age of the buildings. The type of chiller system used is next important attribute that has a large contribution in the hotel buildings. This is due to the fact that many hotels in Singapore different air-conditioning system (split air-conditioning) compared to the office and retail buildings. Moreover, the number of hotel rooms has a large impact in hotel buildings. Finally, it is also interesting to note that the list most dominant building attributes ($AirconFA$, $NonAirconFA$, LED, Age, and $ChillerType$) are very common for all three building types.

3.4. Grade explanation

In addition to grade assignment, the proposed BEEM system also provides insights on which building attributes influence the energy use in individual buildings. Note that the feature importance plots shown in Figure 5 are limited to providing only the overall importance of each building attribute. Figure 6 shows the LIME explanation for an office building that consumed lower energy (actual = $4572.8kWh$) than the peer group (predicted = $4743.2kWh$). The EER of this building is 0.96, and it gets a B grade as per the grade lookup table. In Figure 6, the x-axis denotes LIME values, and the y-axis refers to building attributes in decreasing order of importance. The contribution of each building attribute on energy use is shown as horizontal bars. The red color bars indicate negative contri-

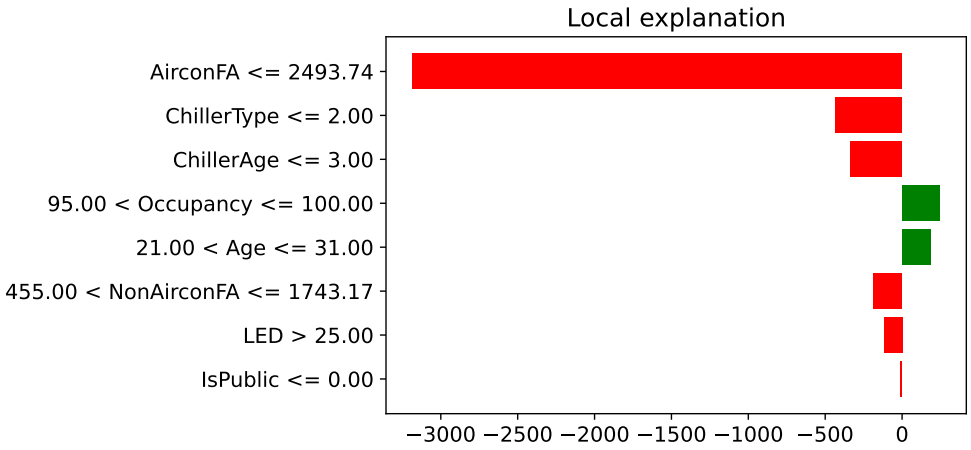

Figure 6: An example of LIME based explanation provided for a office building that consumes less energy. The EER of this building is 0.96.

bution, whereas green bars indicate positive contribution. From Figure 6, we can observe that this office has relatively lower air-conditioned floor area ($AirconFA < 2493.74kWh$), new ($ChillerAge <= 3$) water-cooled chilled water plant ($ChillerType$), moderate non air-conditioned floor area ($455kWh < NonAirconFA <= 1743.17kWh$), and more LED lighting systems ($LED > 25\%$). These factors contribute to lower the energy usage in this building. We can also observe that other factors such as high occupancy ($95\% < Occupancy <= 100\%$') and medium building age ($21 < Age <= 31$) contribute to high energy usage. Still, their combined contribution (all green bars) is much less than the factors that lower the energy usage (all red bars). It is to be noted the LIME-based grade explanation is provided for individual buildings (local explanation), unlike the global feature importance plot as shown in Figure 5. This grade explanation chart is potentially helpful for building managers, providing insights into understanding the factors responsible for high or low energy usage.

3.5. Comparison with Energy Smart and Green Mark systems

The results of the BEEM development show that it is a viable option for a Singapore-specific building energy benchmarking methodology that specifically captures the relevant variables to rate buildings in this context. In this section, we compare the application of BEEM to previous and existing building energy benchmarking methods from the Singapore context and discuss the differences and impact on real-world implementation.

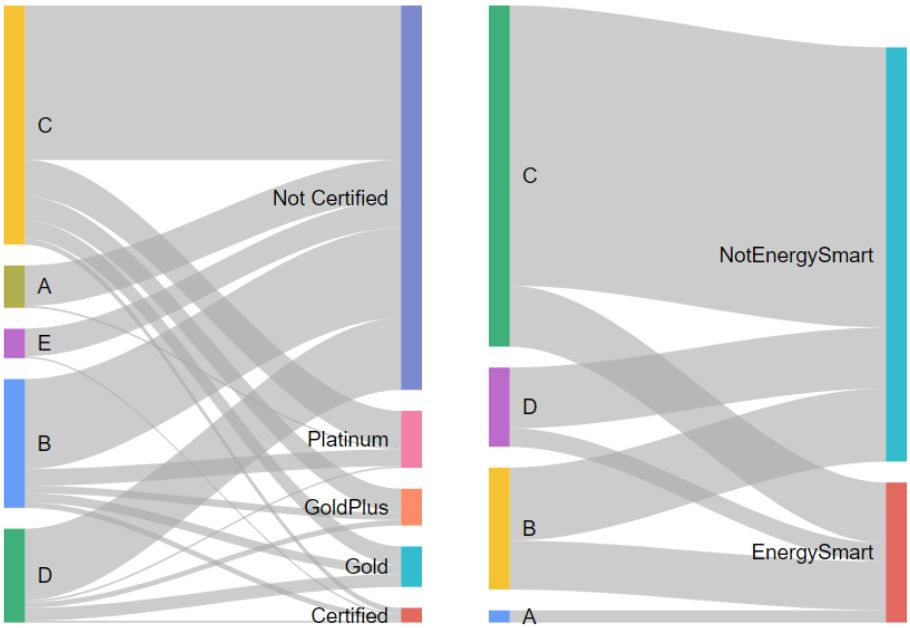

Figure 7: Comparison of grade distribution between BEEM, Energy Smart and Green Mark for the office buildings.

The two baseline building energy benchmarking systems that we will compare BEEM to are Singapore's previous Energy Smart [5] and the current Green Mark [3] energy rating systems. Figure 7 illustrates a sankey diagram of the comparisons of three systems. In the Energy Smart system, which is similar to the *Energy Star*, top 25% of the energy efficient buildings are rated as *Energy Smart Offices*. Whereas, Green Mark is a point based rating system that assigns three labels, namely Platinum, Gold Plus and Gold, based on the earned points corresponding to meeting predefined standard energy efficiency measures. This comparison shows that BEEM is classifying buildings in a more granular way as compared to Greenmark and EnergySmart. The rating systems have a significant amount of overlap in terms of which levels *good* or *poor* performing buildings are captured, but there are many exceptions.

4. Conclusion

We presented the design and implementation of BEEM, an energy use benchmarking system for Singapore buildings. Our approach differs from others by using nonlinear algorithms for accurately modeling building en-

ergy usage and the visual explanation of factors affecting energy usage in an individual building. There are some limitations in our approach and scope for improvement. The BCA's energy disclosure data set, used for establishing the peer group, may not be the nationally representative building samples. There is no documentation available on how BCA selected these buildings. Though we have carefully filtered the samples for each peer group, further refinement is required, using additional building characteristics, to make our approach more generalizable. The evaluation of our system is limited to comparing the performance of our models to the baseline. Because it is difficult to measure and compare the actual energy saving potential of our system with others unless the system is deployed. This is the limitation of any new benchmarking approach. Further, conducting a field study is a future work to evaluate the effectiveness and usability of our visual explanation of grades.

5. Acknowledgement

This work was supported by the Republic of Singapore's National Research Foundation (NRF) through a grant to the Berkeley Education Alliance for Research in Singapore (BEARS) for the SingaporeBerkeley Building Efficiency and Sustainability in the Tropics (SinBerBEST) Program.

References

[1] N. Mims, S. R. Schiller, E. Stuart, L. Schwartz, C. Kramer, R. Faesy, Evaluation of U.S. Building Energy Benchmarking and Transparency Programs: Attributes, Impacts, and Best Practices, Technical Report 1393621, 2017.

[2] Building and Construction Authority of Singapore, 3rd Green Building Masterplan, `http://www.bca.gov.sg/GreenMark/others/3rd_Green_Building_Masterplan.pdf`, 2014. Accessed: 2019-10-01.

[3] Building and Construction Authority of Singapore, Green Mark Scheme, `https://www.bca.gov.sg/greenmark/green_mark_buildings.html`, 2019. Accessed: 2019-10-01.

[4] Leadership in Energy and Environmental Design, LEED green building certification, `http://leed.usgbc.org/leed.html`, n.d. Accessed: 2019-10-01.

[5] S. E. Lee, P. Rajagopalan, Building energy efficiency labeling programme in singapore, Energy Policy 36 (2008) 3982–3992.

[6] W. Xuchao, R. Priyadarsini, L. S. Eang, Benchmarking energy use and greenhouse gas emissions in singapore's hotel industry, Energy policy 38 (2010) 4520–4527.

[7] 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.

[8] Y. Ding, Z. Zhang, Q. Zhang, W. Lv, Z. Yang, N. Zhu, Benchmark analysis of electricity consumption for complex campus buildings in china, Applied Thermal Engineering 131 (2018) 428–436.

[9] H. S. Park, M. Lee, H. Kang, T. Hong, J. Jeong, Development of a new energy benchmark for improving the operational rating system of office buildings using various data-mining techniques, Applied energy 173 (2016) 225–237.

[10] D. W. Kim, Y. M. Kim, S. E. Lee, Development of an energy benchmarking database based on cost-effective energy performance indicators: Case study on public buildings in south korea, Energy and Buildings 191 (2019) 104–116.

[11] J. C. Wang, A study on the energy performance of hotel buildings in taiwan, Energy and Buildings 49 (2012) 268–275.

[12] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, X. Zhao, A review of data-driven approaches for prediction and classification of building energy consumption, Renewable and Sustainable Energy Reviews 82 (2018) 1027–1047.

[13] S. Papadopoulos, C. E. Kontokosta, Grading buildings on energy performance using city benchmarking data, Applied Energy 233 (2019) 244–253.

[14] X. Gao, A. Malkawi, A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm, Energy and Buildings 84 (2014) 607–616.

[15] 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.

[16] S.-H. Yoon, C.-S. Park, Objective building energy performance benchmarking using data envelopment analysis and monte carlo sampling, Sustainability 9 (2017) 780.

[17] Benchmark with EPA ENERGY STAR® Portfolio Manager, `https://www.energystar.gov/buildings/facility-owners-and-managers/existing-buildings/use-portfolio-manager`, n.d. Accessed: 2019-10-01.

[18] A. Kaskhedikar, P. T Agami Reddy PhD, Use of random forest algorithm to evaluate model-based eui benchmarks from cbecs database, Ashrae Transactions 121 (2015) 17.

[19] R. Jing, M. Wang, R. Zhang, N. Li, Y. Zhao, A study on energy performance of 30 commercial office buildings in hong kong, Energy and Buildings 144 (2017) 117–128.

[20] J. C. Wang, A study on the energy performance of school buildings in taiwan, Energy and Buildings 133 (2016) 810–822.

[21] E. H. Borgstein, R. Lamberts, Developing energy consumption benchmarks for buildings: Bank branches in brazil, Energy and Buildings 82 (2014) 82–91.

[22] W.-S. Lee, K.-P. Lee, Benchmarking the performance of building energy management using data envelopment analysis, Applied Thermal Engineering 29 (2009) 3269–3273.

[23] P. Hernandez, K. Burke, J. O. Lewis, Development of energy performance benchmarks and building energy ratings for non-domestic buildings: An example for irish primary schools, Energy and buildings 40 (2008) 249–254.

[24] M. Santamouris, G. Mihalakakou, P. Patargias, N. Gaitani, K. Sfakianaki, M. Papaglastra, C. Pavlou, P. Doukas, E. Primikiri, V. Geros, et al., Using intelligent clustering techniques to classify the

energy performance of school buildings, Energy and buildings 39 (2007) 45–51.

[25] M. Yalcintas, An energy benchmarking model based on artificial neural network method with a case example for tropical climates, International Journal of energy research 30 (2006) 1158–1174.

[26] S.-M. Hong, G. Paterson, D. Mumovic, P. Steadman, Improved benchmarking comparability for energy consumption in schools, Building Research & Information 42 (2014) 47–61.

[27] I. Farrou, M. Kolokotroni, M. Santamouris, A method for energy classification of hotels: A case-study of greece, Energy and Buildings 55 (2012) 553–562.

[28] A. Capozzoli, M. S. Piscitelli, F. Neri, D. Grassi, G. Serale, A novel methodology for energy performance benchmarking of buildings by means of linear mixed effect model: The case of space and dhw heating of out-patient healthcare centres, Applied Energy 171 (2016) 592–607.

[29] H. Sun, S. Lee, R. Priyadarsini, X. Wu, Y. Chia, H. Majid, Building energy performance benchmarking and simulation under tropical climatic conditions (2006) 1–9.

[30] C. Deb, S. E. Lee, Determining key variables influencing energy consumption in office buildings through cluster analysis of pre-and post-retrofit building data, Energy and Buildings 159 (2018) 228–245.

[31] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, Catboost: unbiased boosting with categorical features, Advances in Neural Information Processing Systems 31 (2018) 6638–6648.

[32] P. Arjunan, K. Poolla, C. Miller, EnergyStar++: Towards more accurate and explanatory building energy benchmarking, Appl. Energy 276 (2020) 115413.

[33] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144.

[34] Building and Construction Authority (BCA), BCA Building Energy Benchmarking Report 2017, Technical Report, 2017.

[35] C. Miller, P. Arjunan, A. Kathirgamanathan, C. Fu, J. Roth, J. Y. Park, C. Balbach, K. Gowri, Z. Nagy, A. D. Fontanini, J. Haberl, The ASHRAE great energy predictor III competition: Overview and results, Science and Technology for the Built Environment (2020) 1–21.

[36] 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.

[37] J. Friedman, T. Hastie, R. Tibshirani, The elements of statistical learning, volume 1, Springer series in statistics New York, 2001.

[38] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[39] Y. Freund, R. E. Schapire, A decision-theoretic generalization of online learning and an application to boosting, Journal of computer and system sciences 55 (1997) 119–139.

[40] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM, pp. 785–794.

[41] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth, 1984.

[42] Z. Wang, R. S. Srinivasan, A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models, Renewable and Sustainable Energy Reviews 75 (2017) 796–808.

[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, Journal of machine learning research 12 (2011) 2825–2830.

[44] CatBoost - open-source gradient boosting library, `https://catboost.ai/`, n.d. Accessed: 2019-10-01.