Applied Energy 235 (2019) 1551–1560

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning

Cheng Fan[a,b], Fu Xiao[b], Chengchu Yan[c,*], Chengliang Liu[d], Zhengdao Li[a], Jiayuan Wang[a]

[a] Department of Construction Management and Real Estate, College of Civil Engineering, Shenzhen University, Shenzhen, China
[b] Department of Building Services Engineering, The Hong Kong Polytechnic University, Hong Kong, China
[c] College of Urban Construction, Nanjing Tech University, Nanjing, China
[d] Aviation University of Airforce, Changchun, China

HIGHLIGHTS

- A methodology is developed to explain and evaluate prediction model performance.
- A novel metric, i.e., the trust, is developed to evaluate prediction validity.
- Useful insights on model inference mechanism can be extracted for interpretation.
- It helps to break the tradeoff between model complexity and interpretability.

ARTICLE INFO

Keywords:
Building energy management
Interpretable machine learning
Data-driven models
Building operational performance
Big data analytics

ABSTRACT

The development of advanced data-driven approaches for building energy management is becoming increasingly essential in the era of big data. Machine learning techniques have gained great popularity in predictive modeling due to their excellence in capturing nonlinear and complicated relationships. However, it is a big challenge for building professionals to fully understand the inference mechanism learnt and put trust into the prediction made, as the models developed are typically of high complexity and low interpretability. To enhance the practical value of advanced machine learning techniques in the building field, this study proposes a comprehensive methodology to explain and evaluate data-driven building energy performance models. The methodology is developed based on the framework of interpretable machine learning. It can help building professionals to understand the inference mechanism learnt, e.g., why a certain prediction is made and what are the supporting and conflicting evidences towards the prediction. A novel metric, i.e., trust, is proposed as an alternative approach other than conventional accuracy metrics to evaluate model performance. The methodology has been validated based on actual building operational data. The results obtained are valuable for the development of intelligent and user-friendly building management systems.

1. Introduction

Building operations account for approximately 80–90% of the total energy consumption throughout the whole building life-cycle [1]. The energy saving potential in building operations is typically large due to the wide existence of improper control strategies and operating faults [2]. Conventional approaches to building energy saving rely heavily on domain expertise and engineering experience, which may not be efficient and flexible for generalization. The recent development of Building Automation Systems (BASs) has enabled the real-time monitoring and controls over building operations. As a result, massive amounts of building operational data are being collected and stored in BASs. It is desired to utilize big data-driven methods to discover useful knowledge from building operational data, based on which semi- or fully-automated energy conservation measures are developed.

The unique characteristics of building operational data have imposed great challenges for efficient and effective knowledge extraction. Firstly, building operational data are large-scale and high-dimensional. Secondly, the underlying relationships among building variables are typically nonlinear and have temporal dependencies. Therefore, advanced data analytics are needed to ensure the validity and reliability of data analysis results. In the past few years, researchers and building

* Corresponding author.
E-mail address: chengchu.yan@njtech.edu.cn (C. Yan).

https://doi.org/10.1016/j.apenergy.2018.11.081
Received 31 May 2018; Received in revised form 31 October 2018; Accepted 22 November 2018
Available online 27 November 2018
0306-2619/

professionals have made substantial efforts in bridging the knowledge gap between advanced data analytics and building energy management [3,4]. Big data analytics can be generally classified into two categories, i.e., supervised and unsupervised learning [5,6]. Supervised learning focuses on developing prediction models, either for regression or classification tasks. By contrast, unsupervised learning explores the intrinsic data structures, associations and correlations. Previous studies mainly investigated the potential of supervised learning in analyzing building operational data [7,8]. Machine learning techniques have been used as the main tools to develop prediction models of building cooling or heating load [9,10], building energy consumption [11,12], indoor environment [13,14], and system performance indices [15,16]. To accurately capture the complicated relationships between input and output variables, the supervised learning techniques adopted are typically of high complexity, such as artificial neural networks [17,18], support vector machines [12,19] and decision-tree based ensembles [20,21].

Compared with statistical methods (e.g., multiple linear regression), machine learning techniques generally lead to more accurate predictions. Nevertheless, there is an intrinsic trade-off between model interpretability and model complexity, e.g., machine learning models are "black-boxes" to the users and it is very difficult to understand the inference mechanism learnt. Previous studies in the building field mainly focused on developing accurate models, while overlooking the model interpretability. It should be noted that model interpretability can significantly influence the model applicability in practice. Firstly, prediction accuracy alone is not enough to fully justify the validity of prediction models. For example, a classification task is to be performed to classify the building energy consumption into two levels, i.e., *High* and *Low*. If the relative frequencies of data at *High* and *Low* levels are 95% and 5% respectively, a seemingly satisfactory classification accuracy of 95% can be achieved by simply predicting all cases as *High*. However, such model does not present any practical value. Secondly, building professionals are typically suspicious towards the prediction results unless they can fully understand the model's inference mechanism. More importantly, what building professionals need in practice is not only a single prediction, but also explanations on the decision-making process, e.g., why a certain prediction is made and what are the supporting and conflicting evidences towards it. Therefore, to fully realize the value of advanced machine learning techniques in the building field, it is essential to break the trade-off between model complexity and model interpretability.

Interpretable machine learning is an emerging subject in the field of big data analytics [22,23]. It aims to provide methods and tools to enhance the model interpretability without sacrificing the model complexity. Considering the practical difficulties faced by building professionals in utilized advanced supervised learning techniques, interpretable machine learning is very promising for the development of smart and user-friendly building energy management systems. To the best of the authors' knowledge, there is no such methodology available to address the interpretability of complicated prediction models in the building industry. In this study, a novel methodology is proposed to explain and evaluate data-driven building energy performance models. It serves as a solid solution to break the trade-off between model complexity and model interpretability. The paper is organized as follows: Section 2 provides a brief introduction on interpretable machine learning. The research methodology is described in Section 3. Research results are presented and analyzed in Section 4 and Conclusions are drawn in Section 5.

2. Basics of interpretable machine learning

2.1. Typical approaches to enhancing model interpretability

Interpretability refers to the ability to explain in understandable terms to a human [24]. The core idea is to provide explanations to the inference mechanisms or the logic of the prediction model.

In general, there are two approaches to enhancing model interpretability [24]. The first is to use algorithms with high transparency to create prediction models, e.g., linear regression models and decision trees. These models are relatively easy for human to interpret. For instance, the coefficient of linear regression models can be interpreted as the influence of a certain input variable to the output variable. In terms of decision tree models, the variables and their values used for node splitting can well describe the inference process. The main drawback of this approach is that the models developed are rather simple and hence, the resulting prediction accuracy may not be satisfactory.

The second is to adopt model-agnostic methods to gain insights into the inference mechanism learnt for predictions [24]. Model-agnostic interpretability methods can be applied to any prediction model, i.e., the explanation process is independent of the supervised learning algorithms used. The main advantage is that the users are free to use any supervised learning algorithms for predictive modeling. As a result, the data mining process is less affected by the trade-off between model accuracy and interpretability. As indicated in [25], a desirable model-agnostic interpretability system should have three key features: (1) the system is compatible with any supervised learning algorithm; (2) the system should not be tied with a certain type of explanations, such as linear formulas or rules; (3) the system has the ability to work with features in different representations than that used in the model to be explained.

2.2. Global and local model interpretations

A model can be explained at two levels, i.e., global and local levels. At the global level, interpretation is made based on a holistic view of the model architecture and parameters. In practice, it is very challenging to achieve accurate global model interpretations, especially when there are many correlated input variables [26]. In addition, global interpretations cannot explain why a certain prediction is made and therefore, they cannot fully justify the practical applicability of prediction models. Representative techniques for achieving global interpretations include partial dependency test, individual conditional expectations and feature importance [24]. Such techniques can be used to describe the influence of an input variable on the overall prediction accuracy. By contrast, local interpretations focus on each individual observation and investigate why a certain prediction is made for that observation. The underlying inference mechanism can be better explained by presenting the supporting and conflicting evidences towards a certain prediction. In such a case, domain expertise can be used to examine the inference mechanism learnt, which enables an alternative way to evaluate the model validity. The general approach to local interpretability is to build local surrogate models, where high-transparency algorithms are used to simulate the local relationships around that observation [24]. The state-of-the-art technique in this field is local interpretable model-agnostic explanations [25] and the details are described in the next subsection.

2.3. Local interpretable model-agnostic explanations (LIME)

LIME aims to explain why a certain prediction is made for an observation and what are the supports and conflicts towards the prediction. It is compatible with any supervised learning algorithms. The general idea is to develop a surrogate model based on interpretable representations. The local surrogate model is locally faithful to the complicated model developed. The surrogate models are interpretable models, such as linear regression models and decision trees. Previous studies have shown the power of LIME in explaining highly complicated black-box models and different data types, e.g., text and image data [24,25]. It is adopted as the basis for this study.

The general steps of LIME are summarized as follows: (1) A set of permutated samples is generated and used to get predictions from the

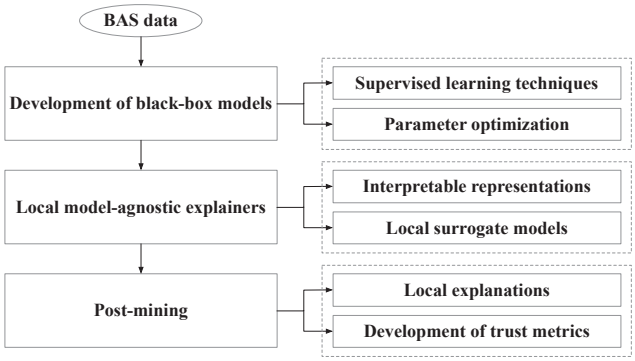

Fig. 1. Research outline.

black-box model; (2) For a given observation, its proximity to permutated samples are calculated and used as weights to represent the relative importance of each permutated sample; (3) The permutated data are transformed into interpretable representations, e.g., transforming numeric values into categorical values; (4) The interpretable representations are recoded and used to reveal the locally stable relationship with prediction outcomes; (5) Explanations are obtained by interpreting the local surrogate model developed, e.g., the coefficients of linear regression models. A detailed description on LIME can be found in [26,27].

3. Research methodology

3.1. Research outline

This paper aims to develop a unified and data-driven methodology to explain and evaluate building energy performance models. As shown in Fig. 1, the methodology contains three key steps, covering the topics on developing prediction models, local model-agnostic explainers, and knowledge post-mining. The first step is prediction model development. Several state-of-the-art supervised learning algorithms are used for constructing black-box models. The second step is to develop local model-agnostic explainers based on the concept of LIME. The third step is knowledge post-mining. It contains three sub-tasks: (1) Perform local explanations; (2) Develop a trust metric to automatically evaluate the reliability of individual predictions; (3) Apply the trust metric to evaluate the overall model performance.

3.2. Prediction model development

In this study, five supervised learning techniques are selected for prediction model development, i.e., generalized linear models, artificial neural networks, support vector machines, random forests and extreme gradient boosting trees. These techniques have been widely used for developing prediction models in the building field [28–30].

Generalized linear models (i.e., denoted as GLM) are developed as the performance benchmark, as they are primarily used to capture linear relationships [31]. The other four techniques are capable of modeling nonlinear and complicated relationships. The multi-layer perceptron (i.e., denoted as MLP) is selected as the architecture for developing artificial neural network models. In this study, the activation function and the number of hidden neurons at the hidden layers are optimized during model training. Support vector machine (i.e., denoted as SVM) is a classic and popular machine learning techniques developed in 1995 [32]. It can efficiently solve nonlinear problems by using kernel functions. In this study, the *C*-type support vector machine with a Gaussian radial basis kernel function is adopted for model training. The complexity parameter *C* is optimized through cross-validation. A larger *C* typically leads to more accurate predictions, yet with an increasing risk of over-fitting. The latter two techniques can be regarded as tree-based ensembles. Random forests (i.e., denoted as RF) adopt a parallel way to grow individual tree models and therefore, individual trees are independent from each other [33]. In this study, the total tree number is set as 500. The tree depth and the number of variables selected for node splitting are optimized during model training. The extreme gradient boosting trees (i.e., denoted as XGB) use a sequential approach to grow individual tree models. It is regarded as one of the most efficient and powerful boosting tree methods [34]. Similar to the development of random forests, the total tree number is set as 500. The tree depth and the learning rate are optimized during model development.

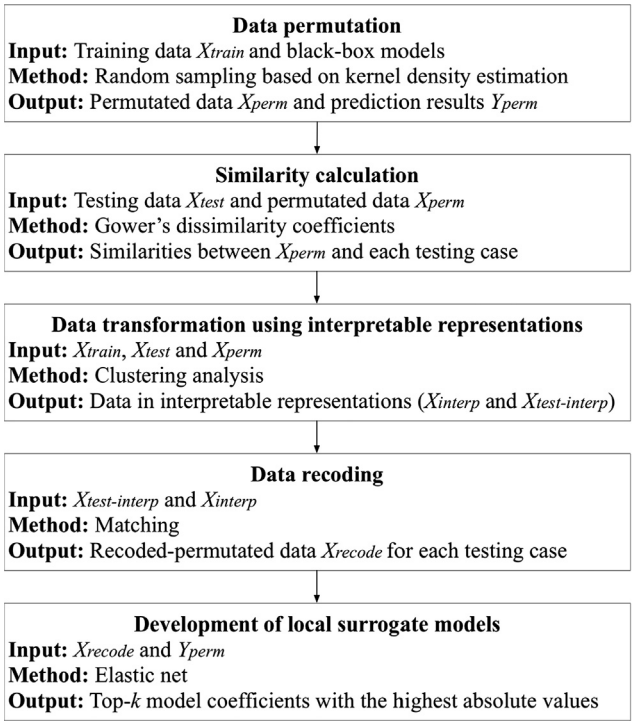

Fig. 2. Key steps in developing local model-agnostic explainers.

3.3. Development of local model-agnostic explainer

As introduced in Section 2.3, the LIME framework contains five main steps and the overall performance can be greatly affected by the methods used in each step. Considering the unique characteristics of building operational data, specific methods are developed to ensure the quality and reliability of local interpretations.

The methods proposed is depicted in Fig. 2. Firstly, data permutation is performed according to a user-define parameter *N*, which specifies the number of artificial observations to be generated. The permutated data is denoted as X_{perm}. Building operational data contain both numeric and categorical variables. In this study, if a variable is numeric, random sampling is performed based on the distribution identified using kernel density estimation. If a variable is categorical, the relative frequency of each possible level in the training data is calculated and used for random sampling. The data permutated are then used as inputs for the black-box model to be explained. The predictions obtained (i.e., denoted as Y_{perm}) are used for local surrogate model development.

Secondly, the similarities between permutated data and each observation in the testing data are calculated. These similarities are used as weights for developing local surrogate models. The rationale behind is that the more similar a permutated sample is to a testing observation, the more useful it is in deriving local explanations. In this study, the Gower's dissimilarity coefficient is used as it is compatible with mixed-type data. The Gower's distance is calculated as $d_{i,j} = \frac{1}{p}\sum_{f=1}^{p} d_{i,j}^{f}$, where

$d_{i,j}$ is the dissimilarity coefficient between the i^{th} and j^{th} observations, p is the total variable number, and $d_{i,j}^{f}$ is the dissimilarity coefficient between the i^{th} and j^{th} observation in terms of the f^{th} variable [35]. The $d_{i,j}$ ranges from 0 to 1 and can be converted into a similarity coefficient by using $1 - d_{i,j}$. If the f^{th} variable is categorical, then $d_{i,j}^{f}$ equals to 0 if $X_i^f = X_j^f$ and 1 otherwise, where X is the data input. If the f^{th} variable is numeric, then $d_{i,j}^{f} = \frac{|X_i^f - X_j^f|}{R_f}$ where R_f is the range of the f^{th} variable.

Thirdly, the data permutated are transformed into interpretable representations and the resulting data is denoted as X_{interp}. This step is typically needed for numeric variables, as categorical variables can be directly used for interpretation. A straightforward approach to numeric data transformation is to discretize the numeric data using equal-width or equal-frequency methods. Such approaches are easy to implement, yet the resulting representations may have little practical value. In this study, a clustering-based method is adopted for data discretization. The k-means algorithm is applied to identify the intrinsic clusters in univariate numeric data. The data in each cluster are represented as the cluster centroid. The representation errors can be calculated based on the total residual sum of squares, i.e., $RSS_{tot} = \sum_{i=1}^{k} RSS_i$ and $RSS_i = \sum_{m=1}^{n_i} (X_m - \bar{X}_i)^2$, where RSS_i is the representation error in the i^{th} cluster, n_i is the number of observations in the i^{th} cluster, and $\bar{X}_i$ is the cluster centroid for the i^{th} cluster. The optimal cluster number is identified by minimizing the total representation errors while limiting the cluster complexity, i.e., the number of clusters. It can be identified by finding the turning point where the decrease in RSS_{tot} is not significant by adding a new cluster.

Fourthly, a new data matrix X_{recode} is created by recoding the interpretable representations regarding each testing observation. The recoding is performed by comparing the interpretable representations in the permutated data and each testing observation. Fig. 3 presents an example for the recoding process. In this case, $N = 3$, X_1 and X_2 are numeric variables, X_3 is categorical and there are two observations in the testing data. Numeric variables X_1 and X_2 are transformed into interpretable representations based on the breakpoints of $\{0, 10, 20\}$ and $\{0, 25, 100\}$ respectively. Considering that there are two observations in the testing data, there will be two recoded data sets and the values are either 0 or 1.

Finally, a local surrogate model is developed to describe the locally stable relationship between the recoded input data X_{recode} and prediction results Y_{perm}. In this study, the elastic net technique is used for linear model development. Elastic net can effectively perform feature selection and reduce the negative effect of multi-collinearity. To ensure the model performance, the parameter α, which specifies the combination of Lasso and Ridge regression, is optimized through cross-validation.

3.4. Post-mining methods

The post-mining step contains three sub-tasks. The first is to gain local explanations based on local surrogate models. Visualization techniques are used to present the top-k model coefficients with the largest absolute values, which represent the primary evidences for each prediction. Such visualizations help to provide a straightforward impression of the major inference mechanisms used for individual predictions.

Secondly, a novel metric, i.e., trust, is developed to evaluate the reliability of each prediction. It is formulated with the following two considerations: (1) among the top-k model coefficients, the larger the number of positive coefficients, the more reliable the prediction is; (2) the larger the absolute values of the positive coefficients, the more reliable the prediction is, as they represent the strengths of supporting evidences. To summarize, the trust value for each prediction is calculated as $T_{ind} = (1 - e^{-\frac{N_s+1}{N_c+1}}) \times \frac{\sum_{i=1}^{N_s} \theta_{N_s,i}}{\sum_{i=1}^{N_s} \theta_{N_s,i} + \sum_{i=1}^{N_c} |\theta_{N_c,i}|}$, where N_s and N_c represent the numbers of supports and conflicts for each prediction, $\theta_{N_s,i}$ and $\theta_{N_c,i}$ refer to the values of the i^{th} positive and negative coefficients respectively. More specifically, the first part of the trust metric, i.e.,

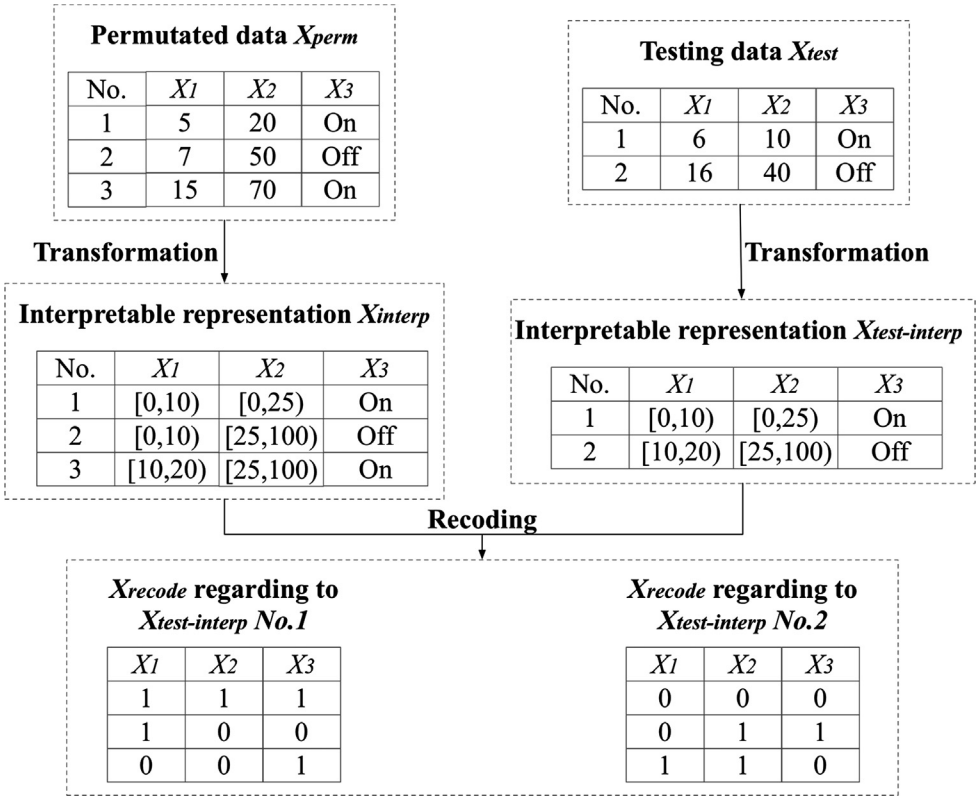

Fig. 3. An example of data transformation and recoding.

$1 - e^{-\frac{N_s+1}{N_c+1}}$, represents the influence of the numbers of supporting and conflicting evidences on prediction reliability. The plus-one Laplace smoothing is used to prevent the denominator from being zero. This part ranges from 0 to 1, and is closer to 1 with the increase of the ratio between the numbers of supports and conflicts. The second part, i.e., $\frac{\sum_{i=1}^{N_s} \theta_{N_s,i}}{\sum_{i=1}^{N_s} \theta_{N_s,i} + \sum_{i=1}^{N_c} |\theta_{N_c,i}|}$, represents the relative strengths of supporting and conflicting evidences. It also ranges from 0 to 1, and is closer to 1 when the strengths of supporting evidences becomes larger. As a result, the whole trust metric for each individual prediction ranges from 0 to 1. The higher the trust value, the more trustworthy or reliable the prediction is. It should be mentioned that since multiplication operations are used for combining these two effects, the trust values calculated can be relatively low. The relative orders of the trust values, rather than their absolute values, are more useful for practical applications. For instance, a certain proportion can be manually specified as the threshold to automatically identify the most unreliable or untrustworthy predictions.

Thirdly, given the same testing data set, the distributions or the summarizing statistics (e.g., mean and median) of trust values can be used to indicate the overall performance of different supervised learning algorithms. Similarly, the relative orders of the summarizing statistics, rather than their absolute values, should be used for performance comparison.

4. Applications on real-world BAS data

4.1. Description of the building and BAS data

The data retrieved from an educational building in Hong Kong were adopted for analysis. It is a fourteen-story building and consists of classrooms for students, offices for university staffs, and a data center for computing devices. The gross floor area is around 11,000 m^2 and 8500 m^2 are air-conditioned. The building cooling load is handled by a complicated central air conditioning system. Fig. 4 presents the schematic of the water-side HVAC system. The chiller plant has a total cooling capacity of 6336 kW. It consists of four water-cooled chillers. Three of them have a cooling capacity of 1932 kW and denoted as *Type A*, while the other one is denoted as *Type B* with a cooling capacity of 540 kW. The set-point of supplied chilled water temperature is 7 °C, and the returned chilled water temperature typically ranges from 11.5 °C to 14.5 °C according to different chiller sequencing control strategies. The heat rejection system contains four cooling towers, i.e., three identical cooling towers with two fans and one with a single fan. The supplied condenser water temperature ranges from 23 °C to 25 °C under different cooling tower sequencing control strategies. The chilled water is circulated using six constant-speed primary chilled water pumps and six variable-speed secondary pumps. The condensing water is circulated using six constant-speed water pumps.

The whole year building operational data in 2015 were retrieved for analysis. The collection time interval was 30-minute. The variables can be divided into four general categories: (1) time variables (*Month*, *Day*, *Hour*, *Minute* and *Day type*); (2) outdoor variables (outdoor dry-bulb temperature and relative humidity); (3) operating parameters of the chiller plant (e.g., the temperatures and flow-rates of chilled water and condenser water, the on-off status of different components); (4) energy related variables, such as the total building cooling load (calculated from the chiller water flow rate and temperature difference) and the total electricity consumption of the chiller plant. The system coefficient of performance (COP) can be calculated based on the building cooling load and the total power consumption of the chiller plant. It should be mentioned that most of the operating parameters are numeric, while a set of variables are categorical variables, e.g., the on-off status of different equipment.

4.2. Prediction models of COP and their accuracy

In this study, the HVAC water-side system COP was set as the model output. It represents the ratio between total cooling load supplied and the total power consumption of chillers, water pumps and cooling towers. The prediction problem is formulated as a classification problem and the system COP was categorized into two levels, i.e., *Low* and *High*. A data-driven approach, which is based on the use of *k*-means clustering analysis, was adopted to determine the optimal cut-off value for COP discretization.

To fully capture the variations in system COPs, three types of variables were adopted as model inputs. The first describes the outdoor environment and contains the outdoor dry-bulb temperature and relative humidity. The second describes the operating parameters of major components, e.g., the flow rates, supplied and returned temperatures of chilled and condenser water at the main pipe, the chilled water bypass flow rate, the supplied and returned temperatures of chilled and condenser water for each chiller, and the on-off status of different chillers and water pumps. In addition, indoor occupancy has profound influence on building energy performance [36,37]. However, such data are typically not available in practice. Considering that indoor occupancy of large buildings is relatively fixed according to time, time variables have been successfully used as proxies for data analysis

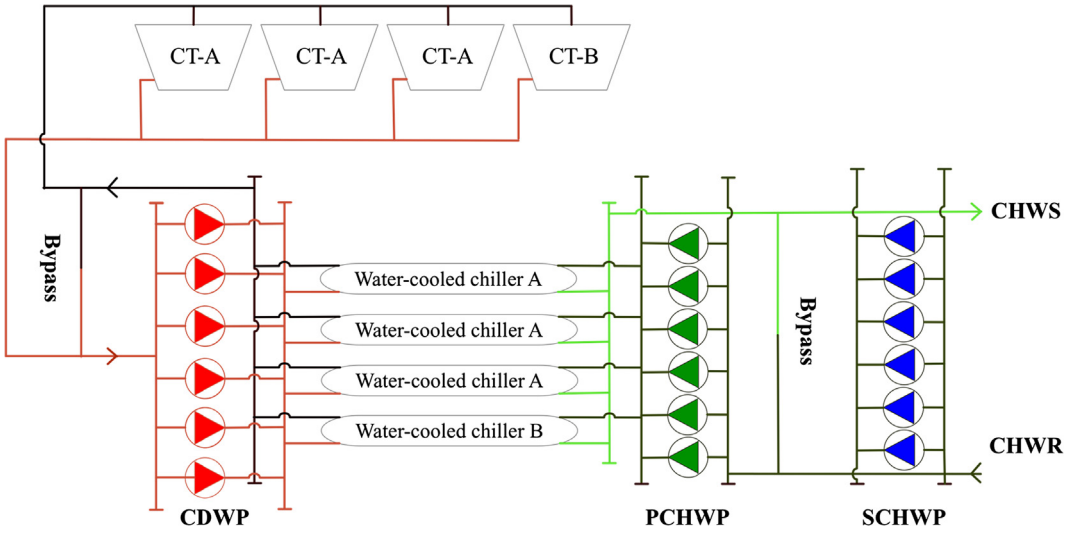

Fig. 4. Schematic of the multi-chiller system.

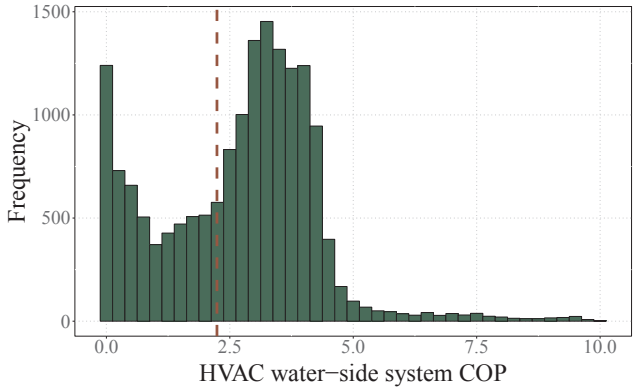

Fig. 5. The overall distribution of system COPs.

[38,39]. Therefore, the third type of input variables are time variables, e.g., *Month*, *Hour* and *Day type*.

The overall distribution of system COPs are shown in Fig. 5. The cut-off value, which is shown as the red dashed vertical line, was identified using the *k*-means clustering analysis and *k* was set as two. The clustering results are summarized in Table 1. It should be mentioned that in practice, domain expertise can be used to determine the cut-off values. However, data-driven approaches are more flexible to use, especially when building operation staffs do not have in-depth understandings on the actual building operation conditions.

In this study, the number of *Low* COPs is smaller than that of *High* COPs. The prediction task is therefore an imbalanced binary classification problem. In general, an imbalanced classification problem is more difficult to deal with, as the prediction models developed may lead to biased predictions towards the majority class and misleading accuracies [40]. One common practice is to apply special data treatments or techniques when the ratio between the majority and minority classes is larger than ten to one [41,42]. There are two main approaches for handling severely imbalanced classification problems [43]. The first is the data-level approach, where over-sampling and under-sampling techniques are used to reorganize the data. The second is the algorithm-level approach, which adopts ensembling methods or cost-sensitive learning to avoid biased predictions.

Considering that the imbalanced ratio in the case study is rather small, i.e., less than 1.5 and hence, no special treatment is needed for model development [40,43]. Random stratified sampling was used for data partitioning, which ensures similar proportions of *Low* and *High* COPs in both training and testing data sets. Five supervised learning algorithms were then utilized to develop prediction models based on training data. The 3-fold cross-validation was used to optimize model parameters. The performance in terms of accuracy metrics was evaluated based on the testing data set. Three accuracy metrics, i.e., classification accuracy, sensitivity (or true positive rate) and specificity (true negative rate), are used to reflect the prediction performance. It should be noted that the *Low* and *High* COPs are recoded as 0 and 1 for modeling. Therefore, the sensitivity refers to the proportion of actual *High* COPs being correctly classified as *High*, while the specificity refers to the proportion of correctly predicted *Low* COPs out of all actual *Low* cases. The testing data have 1644 *Low* and 3334 *High* COPs. The benchmark for classification accuracy is therefore $\frac{3334}{3334+1644} = 67.0\%$. It can be used as an initial assessment of model performance in terms of classification accuracy. For instance, the model developed may not learn meaningful mathematical mapping functions if the classification accuracy is close to the benchmark.

Table 1
Summary on the categorization of system COPs.

Cluster No.	Numerical range	Category	No. of observations
1	[0, 2.24)	Low	5641
2	[2.24, 9.8)	High	10,951

Table 2
Prediction performance in terms of accuracy.

Models	Accuracy	Sensitivity	Specificity
GLM	0.905	0.922	0.871
MLP	0.923	0.948	0.873
SVM	0.930	0.943	0.905
RF	0.954	0.972	0.916
XGB	0.953	0.966	0.927

The resulting prediction accuracy are summarized in Table 2. The classification accuracies obtained range from 90.5% to 95.4%, which are significantly larger than the performance benchmark, i.e., 67.0%. It is observed that the classification accuracy, sensitivity and specificity are positively correlated. GLM model has the worst performance. This is expected due to its limitation in modeling nonlinear relationships. The tree-based ensembles result in the best classification accuracy. Based on the accuracy metrics, it is rather confident to conclude that all the models developed can well handle the imbalanced classification problem. Global explanation techniques, such as the variable importance for tree-based ensembles, can be used to identify the most significant input variables for prediction. Nevertheless, the insights obtained are rather constrained and they cannot assess the validity of individual predictions.

4.3. Assessment of overall prediction model performance based on trust metrics

As described in Section 3.3, local surrogate models were developed to explain the inference mechanism of individual predictions. The number of permutated observations, i.e., *N*, was set as 5000 to ensure the reliability in developing local surrogate models. The *k*-means clustering algorithm was applied to transform numerical variables into categorical variables. The elastic net algorithm was then adopted to tackle the potential problems of variable selection and multi-collinearity. The number of variables used for interpretation was set as 10. The trust metrics proposed in Section 3.4 were calculated for performance evaluation. The summarizing statistics of trust values for each prediction model are shown in Table 3 and the distributions of trust values are presented in Fig. 6. As described in Section 3.4, the trust metric considers both the number of supporting evidences and their strengths. Since multiplication operations are used to combine these two effects, the final trust values can be relatively low even though each of the two parts has acceptable scores, e.g., a final trust value is only 0.49 when both parts have scores of 0.70. It should be mentioned that the relative orders of the final trust values, rather than their absolute values, are more useful for practical applications.

In terms of the supervised learning algorithms, it is found out that the predictions generated by SVM and MLP models are more reliable in terms of trust metrics. By contrast, the trust values of tree-based ensembles are relatively low (especially the RF model), even though they have the best prediction performance in terms of accuracy metrics. It indicates that the tree-based models may have learnt either some local

Table 3
Prediction performance in terms of trust.

Models	Minimum	Median	Mean	Maximum
GLM	0.079	0.379	0.408	0.965
MLP	0.085	0.400	0.417	0.910
SVM	0.076	0.427	0.440	0.916
RF	0.004	0.338	0.374	0.967
XGB	0.004	0.355	0.365	1.000

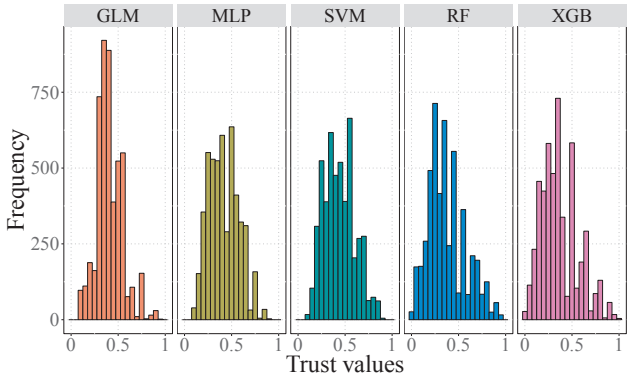

Fig. 6. Distributions of trust values for different prediction models.

characteristics in the data sets or contradicting rules for inference, which may not be helpful for generalization. In such a case, building professionals should investigate the inference mechanism for each individual prediction, especially those with low trust values.

4.4. Evaluation and explanation of individual predictions

The validity of individual predictions can be assessed based on the trust values. Visualization techniques were used to present the top supporting and contradicting evidences for each prediction. The supporting and conflicting evidences are shown in green and red bars respectively and their intensities are reflected by the bar height (i.e., the coefficients of local surrogate model developed). The variables are shown in their abbreviations, i.e., *WCC* refers to the water-cooled chiller, *CHW* and *CDW* refer to chilled water and condenser water, *Main* indicates the main pipe, *CHWP* and *CDWP* represent the chilled water and condenser water pumps, *ST* and *RT* represent the supplied and returned temperatures.

Figs. 7 and 8 present the local explanations for the most trustworthy and untrustworthy predictions generated by the GLM model respectively. Fig. 7 shows that the 2980th testing observation has the most trustworthy prediction. The system COP is predicted as *High* with a trust value of 0.965. The top-5 supporting evidences are the returned chilled water temperature, the supplied chilled water temperature, the cooling load supplied, the total chilled water flow rate, and the on-off status of condenser water pump No. 5. The result meets domain expertise as the intervals identified for numerical variables indicate that the whole system was operating at relatively high part-load ratios, e.g., the cooling load is in the highest interval. By contrast, Fig. 8 shows the 4583th testing observation has the most untrustworthy prediction. The prediction made by the GLM model is *Low*, while there are three major contradicting evidences related to the returned chilled water temperature, the supplied chilled water temperature and the on-off status of condenser water pump No. 6. Further investigations show that when the returned chilled water temperature is in [13.2 °C, 24 °C), 68.5% of the testing observations have *High* COPs. Approximately 67.4% of the testing observations have *High* COPs when the condenser water pump No.6 is switched off.

Figs. 9 and 10 present the most trustworthy and untrustworthy predictions generated by the MLP model respectively. The 2614th testing observation is predicted to have a *High* COP with a trust of 0.91. All the top-five evidences are supporting this prediction. For instance, the temperature difference between the supplied and returned chilled water was larger than 5 °C. The cooling load supplied and chilled water flow rate were in their largest categories, indicating that the system was operating at near-full capacity. By contrast, in Fig. 10, only one out of the top five evidences, i.e., the supplied chilled water temperature in [7.98 °C, 10.2 °C), is supporting the prediction of *High* COP. The cooling load supplied is in the lowest category. Since 53.7% of the testing instances under this category have *Low* COPs, it should have negative impact in predicting *High* COP. Similarly, only 32.0% of the testing observations have *High* COPs when the returned condenser water temperature of the water-cooled chiller No.1 is between 13.5 °C and 22.5 °C. The chilled water flow rate is in the second lowest category and 60.7% of the testing observations have *High* COPs. Considering that the testing data are unbalanced and 67.0% of the labels are *High*, the chilled water flow rate in this category should has a negative impact on predicting *High* COPs.

Figs. 11 and 12 describe the local explanations obtained for the most trustworthy and untrustworthy predictions generated by the RF model respectively. The cooling load has the largest impact in predicting the 514th testing observation to have a *Low* COP. It is in accordance with domain expertise as the cooling load was in its lowest category. By contrast, the 4638th testing observation is predicted to have a *High* COP, even though the cooling load was in the lowest category. Such prediction is very suspicious and domain expertise should be involved for further investigation.

5. Conclusions

Predictive modeling is closely related to typical tasks in building energy management, such as building energy performance modeling and model-based anomaly detection. Conventional analytics, such as physical principle-based and statistical methods, are neither efficient

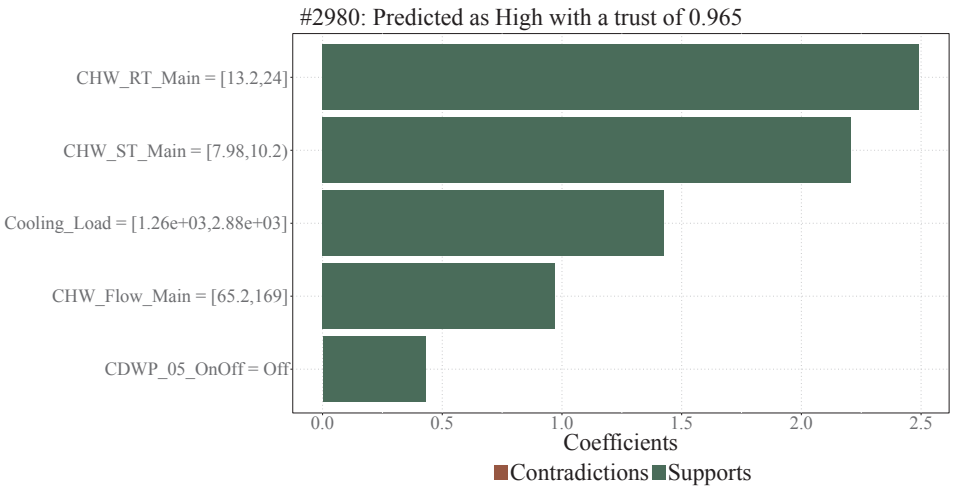

Fig. 7. The most trustworthy prediction generated by the GLM model.

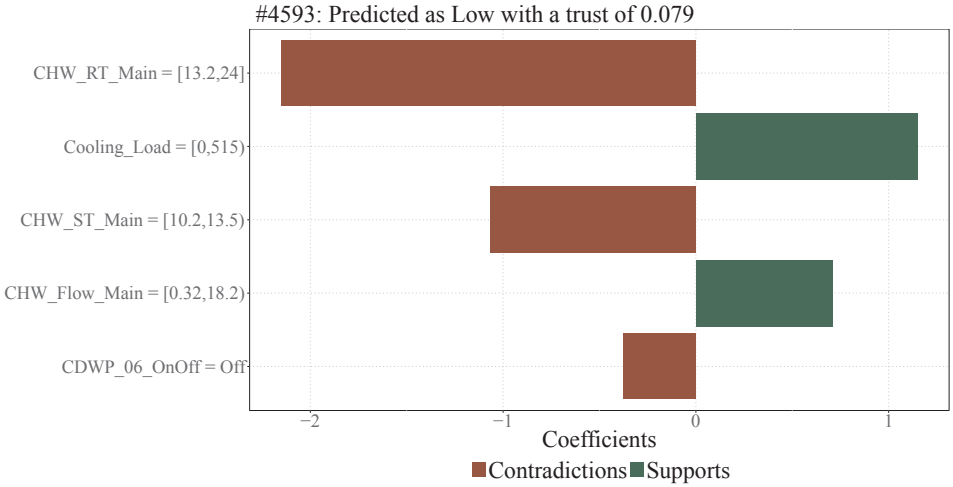

Fig. 8. The most untrustworthy prediction generated by the GLM model.

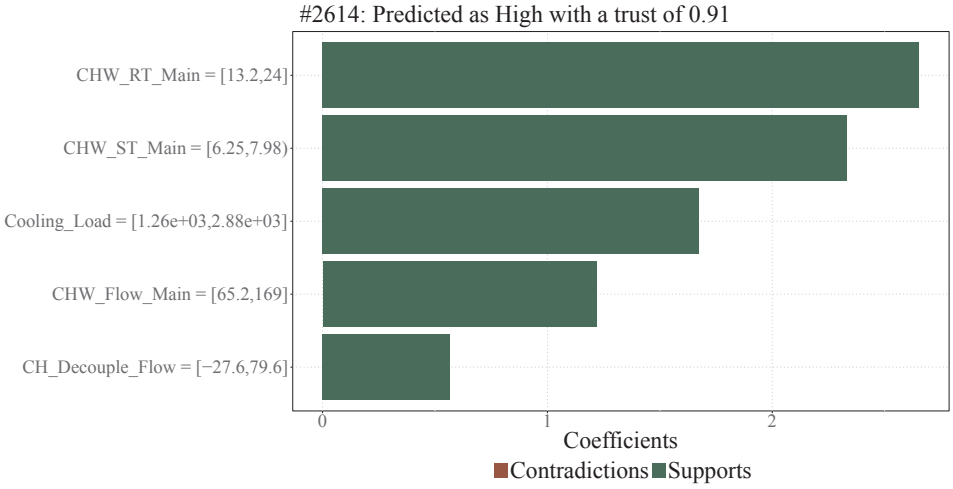

Fig. 9. The most trustworthy prediction generated by the MLP model.

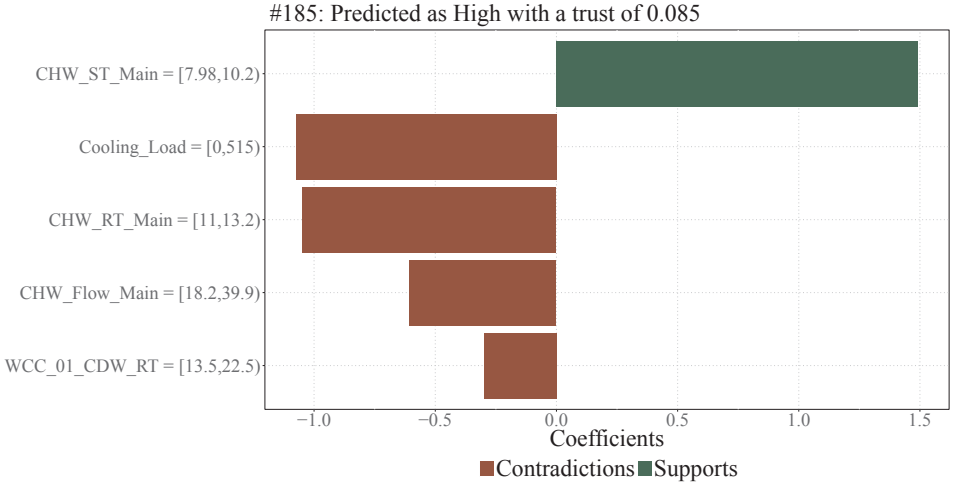

Fig. 10. The most untrustworthy prediction generated by the MLP model.

nor effective in analyzing large building operational data. As a promising solution, advanced machine learning techniques can be used to develop more accurate and reliable prediction models. In practice, there are two main challenges in fully realizing the potential of advanced machine learning techniques in building energy performance modeling. Firstly, there is an intrinsic trade-off between model complexity and model interpretability. Machine learning models can provide more accurate predictions, yet the inference mechanisms learnt from big data are not easy for human interpretation. Secondly, existing studies mainly rely on accuracy metrics to evaluate prediction model performance.

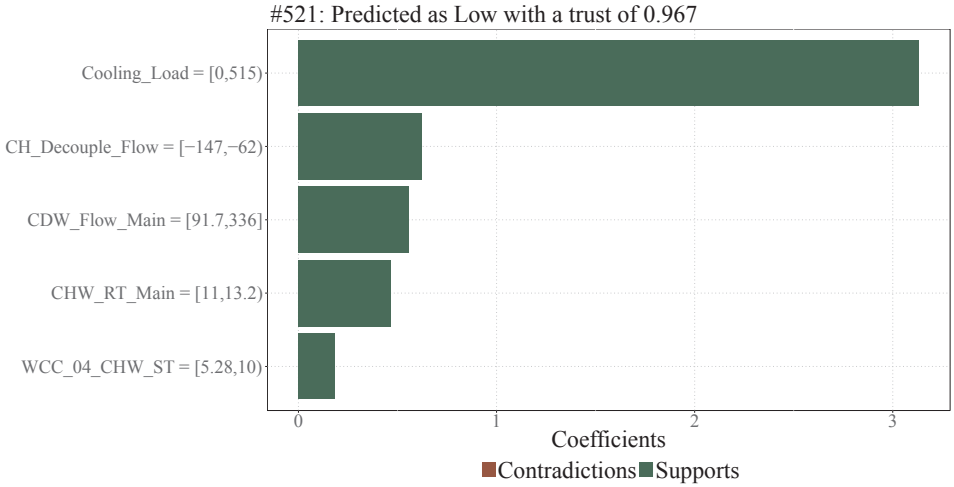

Fig. 11. The most trustworthy prediction generated by the RF model.

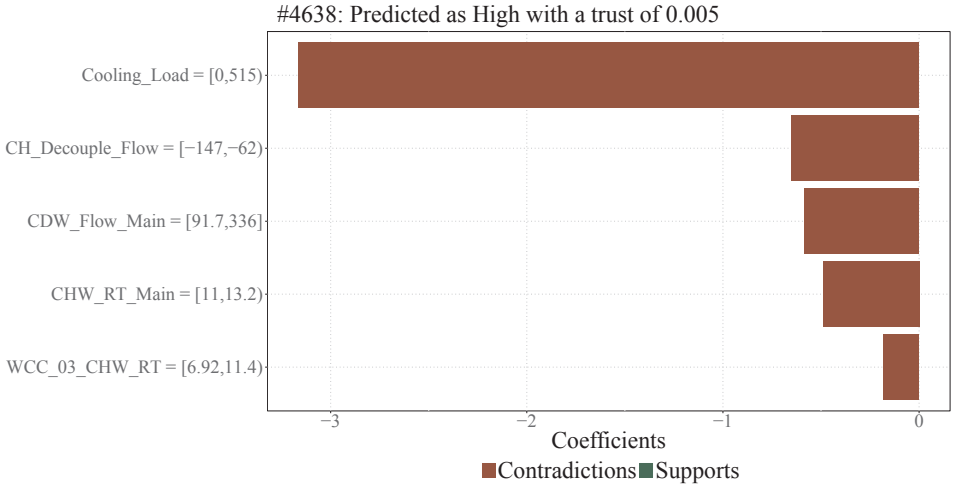

Fig. 12. The most untrustworthy prediction generated by the RF model.

Such metrics are typically used to evaluate model performance from a holistic view. There is a lack of metrics to assess the reliability of each individual predictions.

To tackle the abovementioned challenges, this study proposes a novel methodology to explain and evaluate building energy performance models. The methodology is developed based on the conceptual framework of local interpretable model-agnostic explanations. All the research work was performed using the open-source programming language *R* [44]. Various techniques, such as clustering analysis and Gower's dissimilarity coefficients, are integrated to ensure the quality of data analysis results. Local interpretable models are developed to explain the inference mechanisms of individual predictions. A novel performance evaluation metric, i.e., trust, is developed to quantitively assess the validity of each prediction.

The methodology has been applied to facilitate the prediction of HVAC system COP using real-world building operational data. The results obtained are promising from two perspectives. Firstly, the methodology proposed can be used to explain the local inference mechanisms on individual predictions, no matter how complicated the original prediction model is. It therefore helps to break the trade-off between model complexity and model interpretability. Secondly, the trust metric proposed in this study can be used as an alternative approach to model performance evaluation. As shown in the research results, models with higher prediction accuracies may result in less trustworthy predictions. Such kind of insights are especially useful for building professionals, as the desired outcomes from advanced analytics are not only a single prediction, but also the supporting and conflicting evidences towards the prediction. It can help users to better understand the intrinsic data characteristics and identify potential reasons for model failure. Future studies will be performed to improve the quality of local interpretable models, especially on how to devise interpretable and meaningful data representations based on the original data.

Acknowledgement

The authors gratefully acknowledge the support of this research by the National Natural Science Foundation of China (Grant No. 71772125 and 51708287), the Natural Science Foundation of Guangdong Province, China (Grant No. 2018A030310543), the Research Grants Council of the Hong Kong SAR, China (152181/14E) and the Natural Science Foundation of Shenzhen University, China (Grant No. 2017061).

References

[1] Ramesh T, Prakash R, Shukla KK. Life cycle energy analysis of buildings: an overview. Energy Build 2010;42:1592–600.
[2] Cao XD, Dai XL, Liu JJ. Building energy-consumption status worldwide and the state-of-the-art technologies for zero-energy buildings during the past decade. Energy Build 2016;128:198–213.
[3] Liao SH, Chu PH, Hsiao PY. Data mining techniques and applications-A decade review from 2000 to 2011. Expert Syst Appl 2012;39:11303–11.

[4] Molina-Solana M, Ros M, Ruiz MD, Gomez-Romero J, Martin-Bautista MJ. Data science for building energy management: a review. Renew Sustain Energy Rev 2017;70:598–609.
[5] Miller C, Nagy Z, Schlueter A. A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings. Renew Sustain Energy Rev 2018;81:1365–77.
[6] Fan C, Xiao F, Li ZD, Wang JY. Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: a review. Energy Build 2018;159:296–308.
[7] Amasyali K, El-Gohary NM. A review of data-driven building energy consumption prediction studies. Renew Sustain Energy Rev 2018;81:1192–205.
[8] Yu Z, Haghighat F, Fung CM. Advances and challenges in building engineering and data mining applications for energy-efficient communities. Sustain Cities Soc 2016;25:33–8.
[9] Ding Y, Zhang Q, Yuan TH. Research on short-term and ultra-short-term cooling load prediction models for office buildings. Energy Build 2017;154:254–67.
[10] Fan C, Xiao F, Zhao Y. A short-term building cooling load prediction method using deep learning algorithms. Appl Energy 2017;195:222–33.
[11] Rahman A, Srikumar V, Smith AD. Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. Appl Energy 2018;212:372–85.
[12] Chen YB, Tan HW. Short-term prediction of electric demand in building sector via hybrid support vector regression. Appl Energy 2017;204:1363–74.
[13] Afroz Z, Shafiullah GM, Urmee T, Higgins G. Prediction of indoor temperature in an institutional building. Energy Procedia 2017;142:1860–6.
[14] Geronazzo A, Brager G, Manu S. Making sense of building data: New analysis methods for understanding indoor climate. Build Environ 2018;128:260–71.
[15] Yu Z, Haghighat F, Fung CM, Yoshino H. A decision tree method for building energy demand modeling. Energy Build 2010;42:1637–46.
[16] Chou JS, Hsu YC, Lin LT. Smart meter monitoring and data mining techniques for predicting refrigeration system performance. Expert Syst Appl 2014;41:2144–56.
[17] Rafe Biswas MA, Robinson MD, Fumo N. Prediction of residential building energy consumption: a neural network approach. Energy 2016;117:84–92.
[18] Deb C, Eang LS, Yang JJ, Santamouris M. Forecasting diurnal cooling energy load for institutional buildings using artificial neural networks. Energy Build 2016;121:284–97.
[19] Dong B, Cao C, Lee SE. Applying support vector machines to predict building energy consumption in tropical region. Energy Build 2005;37:545–53.
[20] Fan C, Xiao F, Wang SW. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Appl Energy 2014;127:1–10.
[21] Wang ZY, Wang YR, Srinivasan RS. A novel ensemble learning approach to support building energy use prediction. Energy Build 2018;159:109–22.
[22] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning; March 2017. arXiv:1702.08608v2.
[23] Lipton ZC. The mythos of model interpretability. ICML Workshop on Human Interpretability in Machine Learning, New York, USA; 2016. arXiv: 1606.03490.
[24] Molnar C. Interpretable machine learning: A guide for making black box models explainable. 2018. https://christopha.github.io/interpretable-ml-book/.
[25] Ribero MT, Singh S, Guestrin C. Why should I trust you: explaining the predictions of any classifier; 2016. arXiv:1602.04938v3.
[26] Ribeiro MT, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. ICML Workshop on Human Interpretability in Machine Learning, New York, USA; 2016. arXiv: 1606.05386.
[27] Pedersen TL, Benesty M. LIME: local interpretable model-agnostic explanations; 2018. https://CRAN.R-project.org/package=lime.
[28] Touzani S, Granderson J, Fernandes S. Gradient boosting machine for modeling the energy consumption of commercial buildings. Energy Build 2018;158:1533–43.
[29] Guo YB, Wang JY, Chen HX, Li GN, Liu JY, Xu CL, et al. Machine learning-based thermal response time ahead energy demand prediction for building heating systems. Appl Energy 2018;221:16–27.
[30] Wei YX, Zhang XX, Shi Y, Xia L, Pan S, Wu JS, et al. A review of data-driven approaches for prediction and classification of building energy consumption. Renew Sustain Energy Rev 2018;82:1027–47.
[31] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Springer series in statistics. 2nd ed. Springer; 2016.
[32] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97.
[33] Breiman L. Random forests. Mach Learn 2001;45:5–32.
[34] Chen TQ, Guestrin C. XGBoost: a scalable tree boosting systemThe 22nd ACM SIGKDD international conference on knowledge discovery and data mining, New York, USA. 2016. p. 785–94.
[35] Gower JC. Metric and Euclidean properties of dissimilarity coefficients. J Classification 1986;3:5–48.
[36] Ding Y, Zhang Q, Yuan TH, Yang F. Effect of input variables on cooling load prediction accuracy of an office building. Appl Therm Eng 2018;128:225–34.
[37] Kim YS, Heidaringejad M, Dahlhausen M, Srebric J. Building energy model calibration with schedules derived from electricity use data. Appl Energy 2017;190:997–1007.
[38] Vaghefi A, Jafari MA, Bisse E, Lu Y, Brouwer J. Modelling and forecasting of cooling and electricity load demand. Appl Energy 2014;136:186–96.
[39] Guo YB, Wang JY, Chen HX, et al. Machine learning-based thermal response time ahead energy demand prediction for building heating systems. Appl Energy 2018;221:16–27.
[40] Lopez V, Fernandez A, Garcia S, Palade V, Herrera F. An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf Sci 2013;250:113–41.
[41] Anand A, Pugalenthi G, Fogel GB, Suganthan PN. An approach for classification of highly imbalanced data using weighting and under-sampling. Amino Acids 2010;39:1385–91.
[42] Ali A, Shamsuddin SM, Ralescu AL. Classification with class imbalance problem: a review. Int J Adv Soft Compu Appl 2015;5(3).
[43] Zheng ZY, Cai YP, Li Y. Oversampling method for imbalanced classification. Comput Inform 2015;34:1017–37.
[44] R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing; 2008, Vienna, Austria, ISBN 3-900051-07-0. http://www.R-project.org.