Applied Energy Symposium 2021: Low carbon cities and urban energy systems
September 4-8, 2021, Matsue, Japan
Paper ID: 3

Synthetic minority oversampling based machine learning method for urban level building EUI prediction and benchmarking

Xiaoyu Jin[1], Fu Xiao[1*]

1 Department of Building Services Engineering, The Hong Kong Polytechnic University, Hong Kong

ABSTRACT

Machine learning holds a lot of promise for quickly and correctly assessing building energy performance at urban level. However, due to the lack of data for minority types of buildings, unfavorable results are produced sometimes. Therefore, this study proposes a concise approach to generate enough data for training machine learning models while avoiding overfitting. Superior results are obtained. The importance of variables is analyzed using urban open data sets, which are valuable to data collectors and publishers in decision-making.

Keywords: urban building energy data/ building energy performance/ machine learning/ data generation

1. INTRODUCTION

Urban areas contribute more than 70% of the global final energy usage as well as greenhouse gas emissions [1]. The building sector is responsible for significant proportion of energy consumption in cities, therefore, it offers great opportunities for energy conservation [2]. Building energy benchmarking in an effective way to provide supervision towards energy usage in buildings, building owners tend to take more energy efficient measures under benchmarking policies, city managers and policy makers could also gain more insight based on urban level benchmarking results. And the energy saving effect brought by building energy performance benchmarking is noteworthy, for example, according to an investigation by the US Environmental Protection Agency (EPA) in 2012, Energy Star benchmarks revealed 7% reduction in energy use of over 35,000 buildings during 4 years [3]; in Australia, the National Australian Built Environment Rating System (NABER) enabled large buildings to achieve an average of 33% energy use reduction within 10 years [4].

For city planners and policy makers, quickly scoring the building energy performance at urban level with high accuracy becomes an essential task. Some existing benchmarking approaches are based on comparisons between estimated energy consumption and empirical building operation data. For example, EnergyStar uses multiple linear regression to make prediction and compares the performance of buildings within the same peer groups [5]. However, as proved by many studies, the EnergyStar model is poor at explaining variability for city-level energy data sometimes [6]. Some benchmarking schemes such as Energy Performance of Buildings Directive (EPBD) in Europe, estimate energy consumption by simulation software, requiring for sophisticated model input and consumes lot of time to develop the model [7]. There are also benchmarking schemes that needs equipment examination, such as LEED [8] , which is also labor consuming.

Meanwhile, machine learning models are proved to be capable of making precise estimations on building energy consumption according to increasing numbers of studies. Progress has been made on developing novel benchmarking methods based on machine learning models. Urban building energy data sets are usually classified according to the types of buildings (i.e. residential building, office, school). However, many studies developed machine learning models with very limited (no more than 6) property types of the buildings [3, 9], even if there were over 20 property clusters claimed by the benchmarking frameworks. A handful of studies have a complete coverage of all the property

Selection and peer-review under responsibility of the scientific committee of CUE2021

types, while the models were not reliable for the buildings from minority types [10]. This issue occurs because of the high data-dependence of machine learning models. Some types of buildings only contain small number of samples that are not sufficient for the model to learn something meaningful. As an example, Table 1 shows an example from the dataset *New York City Energy and Water Disclosure for Local Law 84* (LAW84)[11], the building categories are seriously imbalanced. And this problems widely known as the long tail problem in machine learning [12] .

Table 1 Building Category Distribution of LAW84.

Building Category	Sample Number	Building Category	Sample Number
Residential	18462	Meeting	83
Office	2524	Indoor Sports	79
School	2179	Laboratory	24
Storage	754	Restaurant	24
Mall	651	Prison/Incarceration	21
Utility	492	Personal Services	6
Hospital	275	Ice/Curling Rink	5
Worship	170	Zoo	4
Recreation	164	Data Center	4
Service	138	Open Stadium	2

This paper leverages urban level open data to predict building Energy Use Intensity (EUI) using machine learning algorithms. A concise and effective solution of the long tail problem is proposed, which contributes to improve the universality of machine learning based benchmarking methods towards all kinds of buildings. In addition, the importance values of input features are ranked to provide insight of what kind of variables are more important in building energy performance benchmarking. In the following sections, the research methodology, case studies, results and discussions as well as conclusion of the study will be introduced.

2. RESEARCH METHODOLOGY

The research methodology consists of five steps, as shown in Fig 1. Firstly, the raw data is preprocessed. Then, the Random Forests (RF) and Adaptive Boosting (Adaboost) algorithms are preliminary trained on the processed datasets, with the input of building information and output of EUI. The objective of this step is to obtain the optimal parameters for these algorithms. Next, to solve the long tail problem caused by minor types of buildings, the Synthetic Minority Oversample Technique (SMOTE) is used for generating more data of the minor buildings. Cross validation is conducted to determine the optimal number of samples to be generated. The cleaned datasets, machine learning algorithms with optimal parameters, together with the generated data, can predict EUI in step 4, which is the basis for building energy performance benchmarking. Finally, it can determine which variables are essential for EUI prediction and building energy benchmarking according to feature importance aggregation.

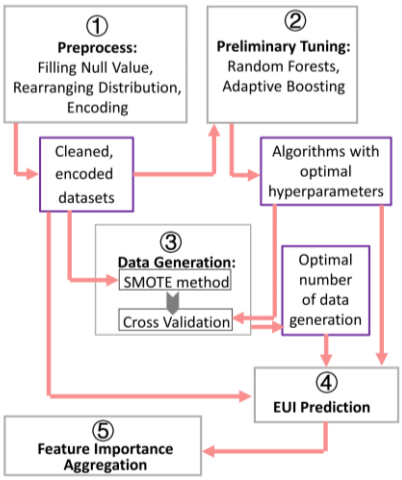

Fig 1 Research flowchart

2.1 Data Preprocessing

The first step of data cleaning is to deal with the null values. Since the target variable is EUI, the samples without EUI values are deleted. Then the columns with too many (>40%) null values are also deleted. And for variables with self-reporting values, the statistics null values can be filled with self-reported records. If without self-reporting values, the null values are filled by K-Nearest Neighbor (KNN) imputation, a handy method to estimate the value based on other closest samples. Outliers in the samples are deleted. For example, the samples with negative “building age” values are deleted.

The next step is the rearrangement of variable distribution, as some of the variables in the datasets have serious left-leaning problems. Variables with heavily skewed distributions are replaced by their natural logarithm values.

Then, the categorical variables are encoded. For the variables with only limited categories (i.e., the “AC inspection condition” only has values of “yes” and “no”), one hot encoding is conducted to change them into dummy variables. While for variables containing great number of categories (i.e., over 100 community districts are recorded in one dataset), which generally also have imbalance problem, one hot coding method makes input matrix extremely sparse. This issue will seriously affect the performance of machine learning

models. Therefore, mean encoding is adopted for this kind of variables.

2.2 Preliminary training and tuning of machine learning algorithms

The machine learning models require most suitable hyperparameters for the best results. During this step, the RF and Adaboost algorithms are tuned. The reasons of choosing RF algorithm include its excellent performance of predicting EUI value shown in similar studies [3] [13], and its visibility which makes the model more explainable. Adaboost is an optimization of RF algorithm. A brief summary of the principle of these algorithms is illustrated as follows.

Random Forests work by creating a large number of decision trees during training. For regression problems in this study, the mean or average forecast of the individual trees is returned. Assuming that the training data set used for modeling contains N samples, P independent variables and 1 dependent variable, first use the Boostrap sampling method to extract N samples from the original training set with replacement to construct a single decision tree. Then randomly select p fields from the P independent variables for the field selection of the decision tree node, and grow an unpruned decision tree according to the MSE. Finally, through multiple rounds of sampling, k data sets are generated, and then assembled into a random forest containing k trees.

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

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

Among them,$F(x)$is the final boosting tree composed of M basic Decision Trees, $F_{m-1}(x)$ represents the lifting tree after $m-1$ rounds of iteration, α_m is the corresponding weight of the m^{th} basic decision tree, $f_m(x)$ is the m^{th} basic decision tree. In addition, the generation of each basic decision tree is not like a random forest. Instead, different weights are set for the sample points based on the classification result of the previous basic decision tree. If the prediction is wrong, the weight of the sample in the next decision tree will be increased, and vice versa, and the next basic decision tree will be constructed.

In order to determine the best sets of parameters for the algorithms, random search with cross validation test is applied. With these optimal parameters, further analysis can be conducted.

2.3 Data Generation

To solve the long tail problem of the essential variable "property type (or building type)", the method of SMOTE is applied. SMOTE works by selecting neighbor examples in the feature space, drawing lines between them and interpolating to get new samples at points along the line with randomly chosen ratios.

A proper data generation number requires to be defined. A plain assumption is that with the increase of generated sample number, the performance of models will become better, but when there are too many created samples, features of the original data will be overwhelmed. Therefore, there should be an optimal number of samples to be generated.

To determine this number, a cross-validation method is applied. As shown in Fig 2, the samples from minority group are divided into five folds, four of them are used to generate data by SMOTE. The generated data is combined with training set from the majority group to make up the training set for the model. One of the folds is maintained and combined with the test set from majority group to constitute of the testing set. In this way, the characteristics of original minority group can be preserved. And the R-squared value (r^2) of the maintained data is calculated separately as an important performance index, which represents the proportion of the dependent variable's variation that is explained by independent variables.

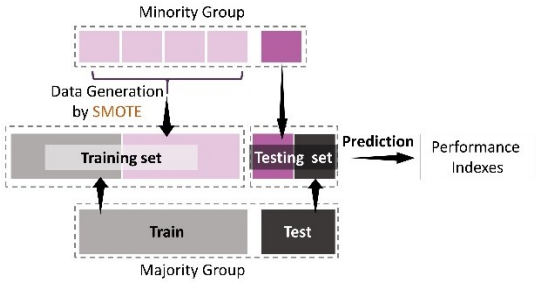

Fig 2 Performance index calculation for one round.

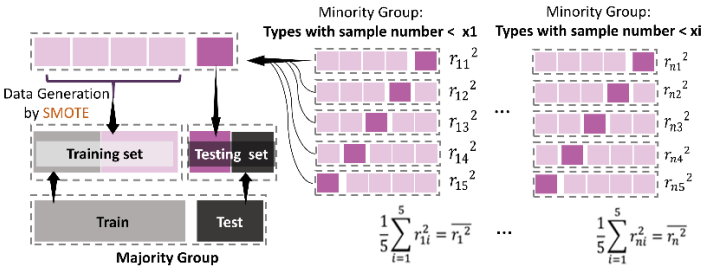

Fig 3 Cross validation for different data generation number.

Similarly, changing the maintained fold gets the r^2 value for other folds (i.e.,$r_{12}^2 \sim r_{15}^2$), as shown in Fig 3. After a 5-fold calculation, the mean value of r^2 can be determined as the performance index ($\bar{r}_1^2$) of a certain data

generation number $x1$. Increase data generation number, the corresponding r^2 can be calculated ($\bar{r}_n^2$).

2.4 EUI Prediction

As shown in Fig 1, there are three key components for EUI prediction models: cleaned dataset obtained from step 1, machine learning algorithms with optimal parameters, derived from step 2, and generated data from step 3. Information of buildings is extracted from the datasets as inputs, which will be presented in 3.1.

Even though the Adaboost algorithm is theoretically more advanced than RF, the performance of the models may rely on the nature of data sometimes. Therefore, both models are applied in the cases, to give more reliable results after comparison.

2.5 Feature importance aggregation

In order to measure the importance of the feature, the value of the feature is replaced in the training data, and on this basis, the out-of-bag error perturbation data set is calculated again. The importance score of each feature is calculated by averaging the difference of out-of-pocket errors before and after the arrangement of all trees. The score is normalized by the standard deviation of these differences.

Since there are various cases in this study, which don't have the same variables as well as features importance ranking, the aggregated feature importance of one variable is defined as the summation of the importance values with the same variable name divided by the occurrence times of this variable (i.e., "Property Type" occurs in each case, so the denominator should be 3). In order to balance the deviation caused by the different number of variables in each case, there is a step before aggregation. The values of importance from each model are multiplied by a certain weight, which is the total number of variables in this case. In this way, all of the variables could be aggregated and compared.

3. CASE STUDIES AND RESULTS

A total of three cases are included, using datasets from open sources. This section will introduce the information of input data, sample generation result of SMOTE, prediction results and feature importance rank.

3.1 Description of cases

The first case uses the building benchmarking data from LAW84 [11], and a geographical database PLUTO [14] is also used to provide additional features of these buildings. The model inputs contain 16 variables, concerning property type, land use purpose, location information (from council to community), building physical information (gross floor area, frontage, etc.), building age, occupancy rate, assessed value of building, proportion of commercial and residential area.

Case 2 uses data of *Chicago Energy Benchmarking* [15], which has a smaller number of samples (2716 buildings). For the model, there are only six input variables: property type, community area, gross floor area, building age, number of buildings, natural gas usage. This case is to verify the feasibility of using small amount of data to conduct EUI prediction.

Case 3 leverages an extraction of benchmarking data from *Energy Performance of Buildings in England and Wales* [16], 45,230 buildings are included. A total of 14 variables are adopted as inputs, related with building type, city and district, AC power, AC installation and inspection, floor area, renewable energy usage, HVAC type, occupancy, fuel type, district heating and grid supplied electricity. This case is to verify the applicability of the methodology framework on building under benchmarking schemes in different countries.

3.2 Data generation results using SMOTE

The cross validation for data generation is processed by optimal algorithms obtained from step 2. The preliminary tuning shows both case 1 and case 3 perform better with Adaboost, in case 2 RF outperforms Adaboost. The models' performances during cross validation of SMOTE are quantified by r^2 of all testing data (R2_model), majority types (R2_major) and the minority types (R2_minor), as shown in Fig 4.

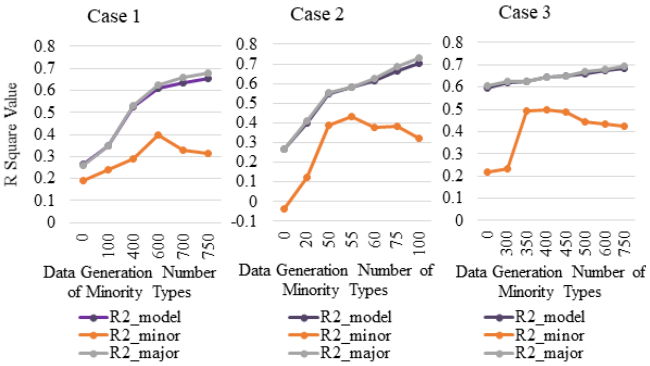

Fig 4 Cross validation results for SMOTE

All cases have similar trend for all types of r^2. It can be seen that the model performances on whole dataset and the majority types keep increasing with more data generated for the minority type. The r^2 values of majority types are always slightly higher than the whole dataset. This is because initially the minority type samples were so few that they are more like noise to the

datasets. With sample number increasing, the model will be more capable of distinguishing these types from other. When there are too many generated samples, there is a risk of overfitting hidden behind the high r^2 values of model.

On the contrast, r^2 values of the minority types are always lower than the other two, in all the three cases they increase to peaks before dropping. This phenomenon is as expected: model's performance on minority types improves as the number of generated samples grows, but with too much generated data, the original data's features of minority types will be overpowered. Therefore, the optimal point can be defined as the peak point of orange line in Fig 4, where the corresponding data generation number is at the peak point of r^2 of the minority types. Compared with the starting points, all the models' performances are significantly improved.

The optimal data generation numbers are determined as 600, 55 and 400 for case 1, case 2 and case 3 respectively, as listed in Table 2. Divide this number by the sample number, the proportion of the tail data can be calculated, this value for the first two cases are both around two percent, while it's 0.88% for the third case.

Table 2 Optimal data generation number of three cases.

	Sample Number	Input Variable Number	Optimal Data Generation Number of Minor Types	Proportion of Tail Data
Case 1	28,807	16	600	2.08%
Case 2	2,716	6	55	2.03%
Case 3	45,230	14	400	0.88%

3.3 Prediction results

The outputs of these cases are predicted ln (EUI). The performances of models are represented by the indexes of Mean Square Error (MSE), r^2 and accuracy. Unlike the prediction step in 3.2 for cross validation, the prediction at this step doesn't maintain any data by purpose. Table 3 has listed the results, all cases have the accuracy beyond 93%, indicating good performances.

Table 3 Prediction results of three cases.

	MSE	r^2	Accuracy
Case 1	0.209	0.600	93.00%
Case 2	0.093	0.585	95.40%
Case 3	0.094	0.646	96.31%

3.4 Feature importance aggregation

Then the variable importance is calculated for each case. As mentioned in 2.4, the variable importance values from all cases are aggregated by their averages after multiplied by the total input variable number of each case. The ultimate rank is shown in Fig 5, and the associated importance levels are assigned.

The variable "Property Type" has so overwhelming advantage that it's the only feature at the level of "Very Important". The "Assessed value of building/land" also shows a surprisingly high level of importance. "Important" level contains gross floor area, estimated AC power, building depth and frontage, lot area, floor-area ratio and community district. Some of the "Important" variables (building depth, frontage, floor-area ratio) are from the geographical dataset PLUTO, this phenomenon reveals the contribution from external database.

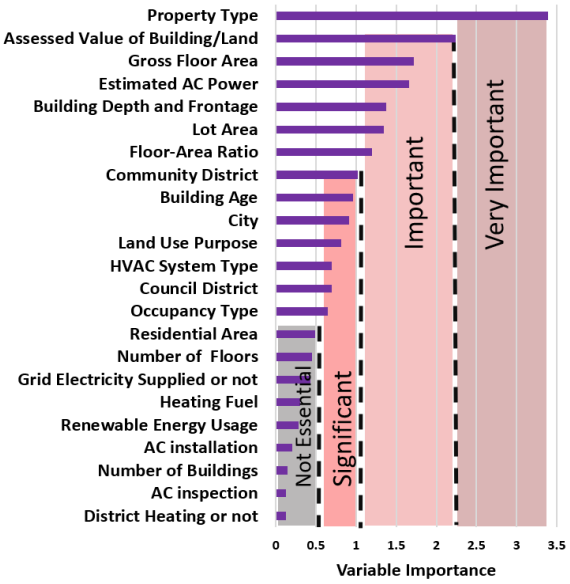

Fig 5 Feature importance ranking.

Moreover, there are some interesting implications from variables standing for similar information. Variables about area are all at "Important" level, the ranking sequence is: Gross Floor area > Lot Area > Floor Area Ratio. For variables about location, the ranking sequence is: Community District > Council District > City. For variables about energy systems, the numerical variable "Estimated AC power" is far more important than the categorical variables such as "HVAC type", "AC installation " and "AC inspection".

4. CONCLUSION

Machine learning is an effective tool to benchmark large number of buildings' energy performance efficiently. However, machine learning methods may derive undesirable results on some minority buildings due to the lack of data.

In this study, the universality of machine learning based prediction on EUI for all kinds of

buildings is proved, by using the concise method of SMOTE. Models' performance has been greatly improved after data generation for minority groups, the risk of overfitting has also been eliminated by cross validation.

Cases presented in this study also show the feasibility of the methodology proposed in this paper. Though Case 2 contains about a tenfold reduction in the number of samples compared to Case 1, and there are not many input variables, it still shows good performance. These cases also demonstrate the generalization capability of the methodology in various nations.

From the aggregated feature importance ranking, the significance of the variable "Property Type" shows that generating data for various types of buildings is effective. Moreover, some variables of high importance are from the external database PLUTO, implying the merits from the combination of two data sources.

Implications are also derived from this study to provide reference for the data publishers. According to data generation results, the proportions of optimal generated tail data are determined to be 0.08%~2.08%, so it is recommended that the number of buildings in the minor type group should be at least 2.5% of the whole dataset. Publishers can also refer to the aggregated feature importance ranks, to determine the priorities of variables in data collection and disclosure.

Future work will focus on develop more novel benchmarking schemes using advanced machine learning algorithms with higher generalization capability to all kinds of buildings in different cities, making more contributions to the urban building energy saving.

Acknowledgements

The authors gratefully acknowledge the support of this research by the Research Grant Council of the Hong Kong SAR (152133/19E).

REFERENCE

[1] Johari F, Peronato G, Sadeghian P, Zhao X,Widén J, Urban building energy modeling: State of the art and future prospects. Renewable and Sustainable Energy Reviews, 2020. 128: p. 109902.

[2] Wei Z, Xu W, Wang D, Li L, Niu L, Wang W, et al., A study of city-level building energy efficiency benchmarking system for China. Energy and Buildings, 2018. 179: p. 1-14.

[3] Arjunan P, Poolla K,Miller C, EnergyStar++: Towards more accurate and explanatory building energy benchmarking. Applied Energy, 2020. 276: p. 115413.

[4] FLORES C, APEC Workshop on Energy Intensity Reduction in the APEC Regions, in Asia-Pacific Economic Cooperation (APEC) Workshop, Y.P.K. Wong, Editor. 2021.

[5] Benchmark Your Building Using ENERGY STAR Portfolio Manager. Available from: https://www.energystar.gov/buildings/benchmark?testEnv=false

[6] Papadopoulos S,Kontokosta C, Grading buildings on energy performance using city benchmarking data. Applied Energy, 2019. 233: p. 244-253.

[7] ENERGY PERFORMANCE OF BUILDINGS DIRECTIVE (EPBD). Available from: http://www.estif.org/policies/epbd0/.

[8] LEED Official. Available from: https://www.usgbc.org.

[9] Kontokosta CE,Tull C, A data-driven predictive model of city-scale energy use in buildings. Applied Energy, 2017. 197: p. 303-317.

[10] Robinson C, Dilkina B, Hubbs J, Zhang W, Guhathakurta S, Brown MA, et al., Machine learning approaches for estimating commercial building energy consumption. Applied Energy, 2017. 208: p. 889-904.

[11] Energy and Water Data Disclosure for Local Law 84 2020 (Data for Calendar Year 2019). Available from: https://data.cityofnewyork.us/Environment/Energy-and-Water-Data-Disclosure-for-Local-Law-84-/qb3v-bbre.

[12] Van Horn G,Perona P, The devil is in the tails: Fine-grained classification in the wild. arXiv preprint arXiv:1709.01450, 2017.

[13] Lokhandwala M, Nateghi RJSP,Consumption, Leveraging advanced predictive analytics to assess commercial cooling load in the US. 2018. 14: p. 66-81.

[14] PLUTO and MapPLUTO. Available from: https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page.

[15] Chicago Energy Benchmarking. Available from: https://data.cityofchicago.org/Environment-Sustainable-Development/Chicago-Energy-Benchmarking/xq83-jr8c.

[16] Energy Performance of Buildings Data:England and Wales Available from: https://epc.opendatacommunities.org/login