2023 European Conference on Computing in Construction
40th International CIB W78 Conference
Heraklion, Crete, Greece
July 10-12, 2023

FEATURE EXTRACTION FOR ENHANCING DATA-DRIVEN URBAN BUILDING ENERGY MODELS

Said Bolluk[1], Senem Seyis[1], and Reyhan Aydoğan[1]
[1]Özyeğin University, İstanbul, TURKEY

Abstract

Building energy demand assessment plays a crucial role in designing energy-efficient building stocks. However, most studies adopting a data-driven approach feel the deficiency of datasets with building-specific information in building energy consumption estimation. Hence, the research objective of this study is to extract new features within the climate, demographic, and building use type categories and increase the accuracy of a non-parametric regression model that estimates the energy consumption of a building stock in Seattle. The results show that adding new features to the original dataset from the building use type category increased the regression results with a 6.8% less error and a 30.8% higher R^2 Score. Therefore, this study shows that building energy consumption estimation can be enhanced via new feature extraction equipped with domain knowledge.

Introduction

Understanding the building stocks' energy demand in cities has great importance since the buildings accounted for 34% of the world's overall energy consumption in 2021 (Buildings – Analysis - IEA, 2022). Such an energy demand covering the embodied and operational consumption of the buildings brings along a massive carbon emission since the primary energy source of the built environment is fossil fuels (2022 Global Status Report for Buildings and Construction, 2022). Therefore, great attention should be given to assessing the energy performance of buildings. In this sense, data-driven models can provide accurate consumption estimations with computational efficiency (Hong et al., 2020). Data-driven models are statistical models that seek a correlation between energy-related features and energy consumption of buildings from historical data. However, statistical models are highly dependent on the existing data, and they might malfunction when the data is not correctly recorded or does not provide relevant information on building energy consumption patterns. Thus, data-driven models might require data enhancement.

Data enhancement can be interpreted in different ways. For example, incorporating new features into a dataset might improve the model training (Hancer, 2020). On the contrary, some features might be redundant or misleading in a dataset. Hence removing these features could benefit the efficiency of the model training (Granell et al., 2022). It is important to master the context and examine the dataset in detail for feeding data-driven models with the most efficient feature combinations. Adding new features to a dataset can be done by processing the existing features or extracting them from external datasets. Modelers search for the hidden relationships between the features to create new ones when the existing dataset is used for feature extraction. Furthermore, external datasets might be a good source of information when the existing dataset lacks information.

The existing datasets in urban building energy modeling practices frequently feel the deficiency of building-specific information considering the diversity of the large building stocks (Hong et al., 2020) and the practical and legal challenges in data collection (Cerezo Davila, Reinhart and Bemis, 2016). Hence, data-driven urban building energy models can benefit from feature extraction. It is important to comprehend the factors affecting building energy demand to successfully perform feature extraction. Thus, domain knowledge is a must when enhancing datasets with new energy-related information.

Using data-enhancing techniques, the research objective of this study is to improve the accuracy of data-driven models utilized in urban building energy consumption estimation. To that end, a building stock in Seattle is selected for a case study. The original dataset and external datasets are utilized in feature extraction. In this sense, climatic characteristics and demographic profiles of neighborhoods, and building use type information are used to derive new features. Various feature combinations are then assessed (Table 3), and the buildings' energy consumption is estimated using a nonparametric regression model. The contribution of the new features is analyzed using regression evaluation metrics. Finally, the achieved and potential improvements through feature extraction are discussed.

Methodology

The original dataset is a building energy benchmarking dataset available at Seattle Open Data (2020 Building Energy Benchmarking | City of Seattle Open Data portal, 2021). This dataset includes 3628 building instances with 42 features, such as the address, use type, and energy consumption details. Jupyter Notebook platform and Scikit-learn library were used for coding purposes and statistical analysis respectively. To enhance the quality of the original dataset, new features within the climate, demographics, and building use type categories will be derived using external data sources.

Data Preprocessing

There are different building use types, such as residential buildings, public facilities, and hospitals, in the dataset. Some instances include more than one building in their facilities that directly affects the total energy

consumption. Therefore, only the instances with one building were selected as the first step of the preprocessing. There were 3211 instances after removing the missing and incorrect entries. Some columns were considered redundant and thus removed. For example, the city, state, and data year features are not necessary for the analysis since all instances have the same entries: Seattle, Washington, and 2020. Moreover, features, such as Compliance Status and Total GHG Emissions were discarded since they are the products of the total energy consumption of a building. The target variable is the normalized version of the building energy consumption with the gross floor area. It is called Site Energy Use Intensity (EUI) with a unit of kBtu/sf. The selected features and the target variable are presented in the original category in Table 1.

Table 1. Feature categories

Feature Category	Features
Original	Building Type Zip Code Latitude Longitude Year Built Number of Floors Property Gross Floor Area (sf) Site EUI (kBtu/sf)
Climate	Count of Parks Elevation (ft.) Rainfall (in.) Snowfall (in.)
Demographics	Median Home Cost ($) Median Household Income ($)
Building Use Type	ASHRAE Building Type Occupancy Density (people/m2) Lighting Density (W/m2) WWR (%) External Wall U-Value (W/m2-K) EUI by Building Type (kBtu/sf)

Deriving New Features

The features in the original dataset do not provide reasonable correlations with the target variable. This dataset mainly holds information about the final energy consumption of buildings rather than parameters affecting the energy consumption. This complicates deriving relationships between the consumption and available features, and thus creating data-driven models estimating building energy demand. Therefore, the datasets need more features that can help understanding patterns in building energy demand. This study aims to derive new features to improve the accuracy of building energy consumption estimation. In this sense, the original and external datasets were used to extract new features. All features were collected under four main categories: Original, Climate, Demographics, and Building Use Type (Table 1).

Climate

Climate conditions have a huge impact on building energy consumption since these conditions determine the demand type (e.g., cooling or heating) and the envelope properties of buildings (Zhou et al., 2020). Therefore, a set of features were derived using external sources. The first feature was the number of parks (green areas) in each district. Seattle Open Data shares the counts of parks by zip code (Seattle Parks and Recreation Park Addresses | City of Seattle Open Data portal, 2016). Buildings' zip codes were utilized to determine the count of parks. The next features were the elevation (ft.) and the annual rainfall (in.) and snowfall (in.) information by zip code. These features were obtained from an open-source website of a private organization called Sperling's Best Places (Seattle, Washington: 28 Zip Codes, 2023). The buildings were assigned these features using the zip code information.

Demographics

The geographical and demographic conditions of a neighborhood might help us to estimate the building energy consumption. For example, the wealth level of a neighborhood might hold valuable insights into the condition of buildings in that district. From here, buildings' age, renovation history, and maybe energy characteristics can be derived. The aim was to question whether the economic profiles of districts can be a good descriptor for building energy efficiency. In this sense, using an open-source mapping service called ZipDataMaps (Seattle Washington ZIP Code Map, 2023), the Median Home Cost ($) and Median Household Income ($) features were assigned to the buildings by their zip codes.

Building Use Type

There are two features related to building use types in the original dataset. These are the Building Type and the Primary Use of Property (EPA Property) features. However, these features are not categorized well. The Building Type feature includes only six classes, which are low-rise, mid-rise, and high-rise multifamily apartments, district schools, university campuses, and non-residentials buildings. This feature categorizes most of the instances in a very generalized way. For example, hospitals, office buildings, and shopping malls are all in the non-residential class. On the other hand, the Primary Use of Property has 65 classes, and more than twenty of these classes have only one or two entries. Such categorization might complicate understanding the similar patterns between the same building use types. Additionally, such a large class number poses an obstacle to statistical analysis. This is because each class of a categorical feature must be presented as a single feature with zero or

one entry in the analysis since most machine learning models can only work with numerical values. Considering the building use type features in the original dataset were either over-generalized or extra detailed, a new feature named ASHRAE Building Type was created.

The U.S. Department of Energy (DOE) created prototype commercial buildings across the country within a study called Commercial Reference Buildings (Commercial Reference Buildings, 2011). The prototype buildings were created using historical data, regional characteristics, and expert opinions in the study. These prototypes were exposed to detailed computer energy simulations using EnergyPlus with their energy-related properties, such as envelope properties and occupancy loads. There is a pilot city for each state to hold these energy-related properties of buildings in the DOE's study. Seattle is the selected city that represents the commercial buildings in Washington State. Hence, this study utilized the building templates of Seattle.

The DOE classified buildings under 16 different use types, including office buildings, schools, hospitals, and residential apartments. These different use types were used as a reference in creating the new feature ASHRAE Building Type. In this sense, the original features Building Type and Primary Use of Property were utilized to first create main types. For example, entries with office, financial office, and medical office were gathered under the Office Class. The entries with hospitals, clinics, and physical therapy centers were named hospitals. Similarly, multifamily houses, lodging facilities, and residential care facilities were collected under the Residential Class. After labeling each building with a main type, the age, floor number, and gross floor area features were used for classifying the building according to its final use type. For example, floor numbers were used to classify office buildings under small, medium, and large offices. Hospitals and outpatient healthcare facilities were separated using the gross floor area information.

The aim was to place buildings into the classes in a way that obtains the most similar characteristics in each class. To that end, using the DOE's study and considering the original building types, 17 building use types were created within the ASHRAE Building Type. There is a difference in the number of use types between the original and extracted use types. This is because there are no instances in the original dataset that can be labeled as quick service restaurants, which exists in the DOE's study. Moreover, the DOE's study labels all residential apartments as mid-rise apartments, where the number of floors varies between 1 and 76 in the original dataset. Since it was impossible to label each residential building as a mid-rise apartment, three classes were created for the residential apartments: Low-rise, Mid-rise, and High-rise. Table 2 illustrates the mean EUI of the buildings within each ASHRE building type.

Using the DOE's templates once again, five more features were extracted covering the buildings' occupancy schedules, envelope properties, and consumption details. The occupancy schedules were the Occupancy Density (people/m2) and Lighting Density (W/m2). The features Window-to-Wall Ratio (WWR (%)) and External Wall U-Value (W/m2-K), which is the thermal transmittance value of the external walls, were the envelope properties. The final feature was EUI by Building Type (kBtu/sf) which represents the annual energy consumption density of prototype buildings. The DOE's study determined such densities by computer energy simulations with the energy-related properties of the prototype buildings. Therefore, each building in this study was assigned these five features according to the ASHRAE Building Type feature.

Table 2. EUI statistics by ASHRAE Building Type

ASHRAE Building Type	Mean of EUI	Std. of EUI	Number of Observations
High-rise Apartment	46	15.3	120
Hospital	196	14.1	3
Large Hotel	54.9	22.1	34
Large Office	51.9	25.1	92
Low-rise Apartment	34.1	14.8	989
Medium Office	64	67.9	227
Mid-rise Apartment	36.9	14.6	649
Outpatient Health Care	131.8	74	5
Primary School	36.4	15	105
Restaurant	153.9	91.5	8
Secondary School	33.1	7.8	30
Small Hotel	50.2	20.4	50
Small Office	62.7	52.7	207
Stand Alone Retail	50.8	29.8	80
Strip Mall	70.8	53.8	15
Supermarket	201.4	94.3	37
Warehouse	32.6	27.3	183

Model Development

Estimating the energy consumption of buildings forms a regression task with a continuous target variable called Site Energy Use Intensity (EUI). The aim here is to estimate the building energy consumption by analyzing the available features and their relationships with the target variable. However, the existing and generated features do not provide a great correlation with the target feature. This is because the existing dataset lacks some of the most essential building energy-related parameters, such as thermal transmittance value, air infiltration rate, and properties of the mechanical systems (Wang et al., 2020). Therefore, linear models might be insufficient to understand the energy consumption patterns of the

buildings. Moreover, the features in the dataset are not perfectly Gaussian or do not have a certain probability distribution. In such cases, parametric models fail to satisfy accurate estimations. This is because parametric models assume a certain distribution for the features of a dataset and estimate the parameters of that distribution when making estimations (Alpaydin, 2020). On the other hand, non-parametric models do not need a certain probability distribution to make estimations (Alpaydin, 2020). These models analyze a small subset of instances rather than the whole dataset and derive more complex patterns from the data. Therefore, the non-linearity provided by such models can represent the correlation between the features and the target. It should be noted that non-parametric models are computationally exhaustive and prone to be overfitted since they aim to understand complex patterns more than generalize the correlation (Alpaydin, 2020). However, considering the low correlation between the features and the target, the non-parametric models can provide more effective regression results than the linear models.

Table 3. Training scenarios

Scenario	Combination of Feature Categories
S1	Original
S2	Original + Climate
S3	Original + Demographics
S4	Original + Building Use Type
S5	Original + Climate + Demographics
S6	Original + Climate + Building Use Type
S7	Original + Demographics + Building Use Type
S8	Original + Climate + Demographics + Building Use Type

In this sense, Random Forest Regressor was selected to perform the building energy consumption estimation. Random Forest Regressor is an ensemble learning method utilizing many random estimators called Decision Trees (1.11. Ensemble methods, 2023). A decision tree is an algorithm that adopts a hierarchical order by splitting the data with yes-no questions according to the features (Alpaydin, 2020). Each split forms a binary decision node, and, thus, sub-trees (Figure 1). For example, if a random feature selected for the first split is the Year Built and the arbitrary question is whether the buildings are constructed before 1980, there will be two decision nodes with buildings constructed before 1980 and the buildings constructed after 1980. After each split, the nodes are stretched and become more homogenous. Once a certain homogeneity is achieved for a decision node, the model stops separating the instances and creates the leaf nodes with explicit estimation parameters. Assume that we would like to estimate the error for a node m in a decision tree regression. If X is the whole dataset with features and targets, x is an instance with multiple features, N_m is the number of instances reaching the node m, y is the target variable (output), g is the model's estimation which is the average output of the instances on node m, and E is the error function for node m, then the error is calculated using Equation 1 (Alpaydin, 2020):

$$E_m = \frac{1}{N_m}\sum_t |y^t - g_m| b_m x^t$$

$$X = \{x^t, y^t\}_{t=1}^N \qquad (1)$$

$$g_m = \frac{\sum_t b_m x^t y^t}{\sum_t b_m x^t}$$

The objective of the tree model is to minimize the overall absolute difference between the targets and estimations (Equation 1) covering each node. In Equation 1, the term b takes the value of one if the instance x reaches the node m and it takes zero otherwise. The split strategy is based on a random process. The number of possible splits is exponential with the feature numbers. Therefore, there might be great variance between the results of the tree models initiated with a different random seed. By averaging the performance of these trees, however, the Random Forest algorithm can converge fast and decrease the variance in different tree estimations (Shalev-Shwartz and Ben-David, 2014).

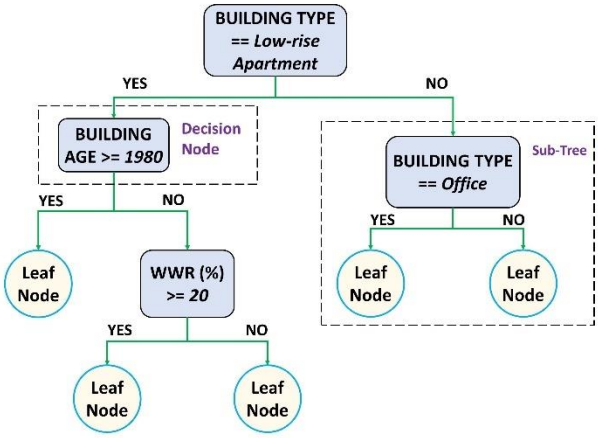

Figure 1. A decision tree structure with maximum tree depth of three (3).

The Random Forest algorithm used in this study has a hundred different Decision Tree Regressors with a random splitting strategy. The forest regressor might avoid overfitting since each tree within the forest utilizes a different subset of data for training and test purposes. The error function that the algorithm optimizes is the mean absolute error. Another hyperparameter is the maximum tree depth among all the trees in the forest regressor. The maximum tree depth refers to the maximum decision nodes of the branch that is extended

most in the whole tree (see, e.g., Figure 1). A validation curve was used for determining the optimal tree depth. In machine learning practices, the aim is to generalize the correlation between the features and the target rather than memorizing the existing dataset. This is because if a machine learning model has a generalization capacity, it can represent the population but not only the sample space, which is the available dataset. Herein, the validation curve helps us determine the optimal hyperparameters of a model by observing the error or accuracy change over different hyperparameter settings on both the training and test data. Once the optimal hyperparameters are selected, the model is trained using the training set, and the model's performance is evaluated using the test set.

Different tree depths ranging between one and twenty were evaluated within the validation curve. However, the regression evaluation metrics utilized in the validation curve should be first discussed to understand the hyperparameter selection procedure. In this study, two different regression metrics were employed: Coefficient of Determination (R^2) and Mean Absolute Percentage Error (MAPE). R^2 Score is a metric that examines how much the model outperforms the mean estimator (Equation 2). This metric evaluates to what extent the model explains the variance in target values using the available features (Ross, 2020). MAPE is the mean absolute difference between each actual (target) value and the model's estimation divided by the target value (Equation 3). In brief, R^2 Score examines the model's generalization capacity, whereas MAPE analyzes the model's accuracy. Equation 2 and 3 denotes how to calculate R^2 Score and MAPE, respectively, where X is the whole dataset with features and targets, N is the total number of instances, y is the target value, g is the model's estimation according to the instance x, and r is the mean of the target values:

$$R^2(y,g) = 1 - \frac{\sum_1^t (y^t - g^t)}{\sum_1^t (y^t - r)} \qquad (2)$$
$$X = \{x^t, y^t\}_{t=1}^N$$

$$MAPE(y,g) = \frac{1}{N}\sum_1^t \frac{|y^t - g^t|}{|y^t|} \qquad (3)$$
$$X = \{x^t, y^t\}_{t=1}^N$$

The validation curve was utilized over Scenario-1, having only the original features (Table 3), for the consistency of the model evaluation. This is because if the hyperparameters are defined for Scenario-1 and fixed for the rest of the scenarios with different feature combinations, then the contribution of the generated features can be assessed. Figure 2 shows that increasing the model complexity with larger tree depths could not reduce the error after some point. This is the point where the maximum tree depth is equal to six (6). Even though the R^2 Score slightly increases after the depth of six, using more complex models should be avoided. This is because the model performs well on the training data since it starts recognizing the training data with complex structures. However, this behavior decreases the accuracy of the model on the test set after a certain depth even with a minor increase in the R^2 Score. Hence, the optimal value for the maximum tree depth was defined as six (6) according to the validation curves in Figure 2.

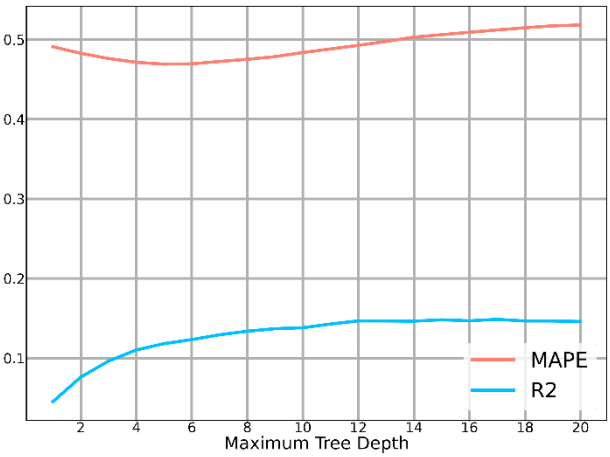

Figure 2. Validation curves for the test sets of Scenario-1: R^2 Score vs. Maximum Tree Depth (blue curve) and MAPE vs. Maximum Tree Depth (red curve).

To better illustrate the trade-off between the complexity and generalization power of a regressor, several Random Forest Regressors with different tree depths and a Linear Regression model were trained, and their regression performances were analyzed in Figure 3. The distribution of the feature (Number of Floors) and the target variable (Gross Floor Area in square feet) was scattered, and different regressors were trained and tested over these instances without a train-test-split in Figure 3.

The non-parametric regressors, which are the Random Forest Regressors, distinguish easily from the Linear Regression model as they fit the data points in more detail according to Figure 3. However, as the tree depth increases, these models tend to memorize the data and thus lose their generalization capacity. This is problematic because when new instances are introduced, the model might not be able to make accurate estimations since it recognizes the patterns in the training data rather than generalizing them. Therefore, a random forest regressor with a maximum tree depth of six (6) was used to assess the scenarios (Table 3), including different feature combinations according to the validation curves in Figure 2. The datasets were split into training and test sets to prevent overfitting with 25% of the instances belonging to the test set. This splitting strategy was repeated five times to make sure the model utilizes each instance in both the

training and test sets. This helps obtain consistent and general results because the model is trained over multiple instance combinations rather than a single random combination.

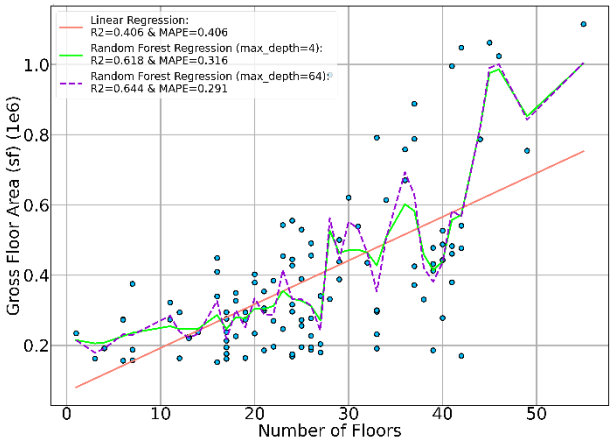

Figure 3. Model performance vs. model complexity.

Results and Discussion

The complete regression results are given in Table 4. Overall, the results show that the linear regression models performed worse than the Random Forest Regressors for each scenario in terms of both MAPE and R^2 Score. This indicates the power of the nonparametric models in handling datasets with low feature-target correlation and features that contain irregular probability distributions. The results suggest that Scenario-7 has the least regression error and the highest R^2 Score. Scenario-7 provided 6.8% less MAPE and 30.8% higher R^2 Score than Scenario-1, which is the baseline dataset with the original features. Such improvement was provided by the newly extracted features from the building use type and demographics categories. However, it is evident that the three best scenarios (Scenario-7, Scenario-4, and Scenario-8) contain features from the building use type category. Therefore, each feature's contribution to the regression results should be elaborated to better understand the improvements achieved in the accuracy and generalization power of the model through new feature extraction. To that end, a Permutation Feature Importance (PFI) function, which detects the singular importance of features when training a regressor, was used.

The PFI first calculates the score of a machine learning model (main score) trained by a certain feature combination. Then, it shuffles the instances of a selected feature and re-calculates the model score according to the selected feature (relative score). Finally, it determines the feature's importance by subtracting the relative score from the main score. In brief, PFI determines the importance of a feature by analyzing the effect of a change in the feature's order on the overall score (4.2. Permutation feature importance, 2023). The relative score can be determined by applying multiple shuffles. This ensures the PFI function covers many combinations of the feature so the results can be more valid and general.

Table 4. Regression results

Model	Scenario	MAPE	R^2 Score
Random Forest Regressor	S7	0.343	0.436
Random Forest Regressor	S4	0.343	0.431
Random Forest Regressor	S8	0.344	0.436
Random Forest Regressor	S6	0.344	0.435
Random Forest Regressor	S1	0.411	0.128
Random Forest Regressor	S3	0.412	0.135
Random Forest Regressor	S5	0.412	0.135
Random Forest Regressor	S2	0.413	0.130
Linear Regression	S7	0.439	0.422
Linear Regression	S4	0.440	0.420
Linear Regression	S6	0.441	0.422
Linear Regression	S8	0.441	0.418
Linear Regression	S2	0.534	0.155
Linear Regression	S3	0.535	0.154
Linear Regression	S1	0.537	0.155
Linear Regression	S5	0.538	0.153

For example, if the main score of the model is s, the relative score of a random feature j for each iteration k is $s_{k,j}$, then the importance of the feature i_j is calculated using Equation 4 (4.2. Permutation feature importance, 2023):

$$i_j = s - \frac{1}{K}\sum_{k=1}^{K} s_{k,j} \qquad (4)$$

MAPE (Equation-3) was selected as the scoring metric for the PFI calculations. Scenario-7 was used for the PFI evaluation since it outperformed the other scenarios. Each feature of Scenario-7 was shuffled thirty (30) times to obtain a general and reasonable understanding of its importance. The permutation feature importance values are given in Table 5. According to Table 5, the feature EUI by Building Type (kBtu/sf) had a remarkable impact on the regression results with a 19.2% change in the main score. The rest of the features did not or barely impact the regression results. For example, the features that exist in the original category cumulatively resulted in a 10.6% change in the main score. The other extracted features

from the Building Use Type and Demographics categories provided only a 2.1% change in the main score in total. The expected contribution of the other features from the building use type category (e.g., External Wall U-Value (W/m2-K) and Occupancy Density (people/m2)) was much higher since these features should represent the patterns in building energy demand in theory. However, it is understandable that there is not any decent variation in the values of these features comprising each building with its specific characteristics. To conclude, the 6.8% error reduction in MAPE with a 30.8% increase in R^2 Score was achieved after equipping the original dataset with mainly the EUI by Building Type (kBtu/sf).

Table 5. Permutation Feature Importance (PFI)

Feature Name	PFI
EUI by Building Type (kBtu/sf)	0.192
Year Built	0.037
Building Type	0.015
Longitude	0.013
Property Gross Floor Area (sf)	0.012
Latitude	0.011
Zip Code	0.009
Number of Floors	0.009
Occupancy Density (people/m2)	0.006
WWR (%)	0.004
ASHRAE Building Type	0.003
Median Home Cost ($)	0.003
Median Household Income ($)	0.003
Lighting Density (W/m2)	0.002
External Wall U-Value (W/m2-K)	0.001

This study has some limitations. The first one is the deficiency of the original dataset to make reasonable estimations or derive new features that might improve the accuracy of regression. The original dataset is an energy benchmarking dataset with almost no building energy-related parameters, such as air infiltration ratio, structural material, and occupancy features. Even though some valuable features on the building use type category (Table 1) were added to the original dataset, the values here were not specific to the recorded buildings, but they are rather based on certain building types. Moreover, there is a great variance in the energy consumption of buildings within the same use type. Table 2 shows the mean and standard deviation of the Site EUI by the classes in the feature ASHRAE Building Type. Most of the building types have a great variation in energy consumption in the original dataset. Considering that the EUI is the normalized form of the building energy consumption by the gross floor area, standard deviations close to the mean or greater than the mean indicate a massive diversity in the dataset. Plus, the distribution of the instances by ASHRAE Building Type is highly imbalanced. For example, most instances are low-rise apartments with 989 observations, whereas there are only five outpatient healthcare facilities. Such an imbalanced distribution hinders the model from adequately generalizing each building type. This is because the model is inclined to learn the patterns in the building types with many observations (e.g., mid-rise apartments). When testing the model's performance, it is likely to confront a low accuracy for the building types with fewer observations (e.g., restaurants) since there is not much training nor test instances.

All these dataset characteristics complicate obtaining homogenous nodes with the same building types and thus attaining similar consumption patterns. The building use type is a feature that holds uncertainties about the occupancy, age, envelope, and mechanical systems of buildings. Well-organized and comprehensive databases are the keys to reducing these uncertainties and discovering the building energy characteristics.

The selection of the machine learning model forms another limitation of this study. A non-parametric regression model was selected due to the low correlation raised from the original dataset and the irregular density distribution of the features. However, the optimal hyperparameters were defined by testing only Scenario-1 due to the computational capacity of the study. An extensive hyperparameter tuning over each scenario might yield better results in the future. Despite all these difficulties, however, this study showed the potential of increasing the accuracy of building energy consumption estimation with the help of new feature extraction utilizing internal and external data sources and domain knowledge.

Conclusion

A building energy benchmarking dataset was used and enhanced with new feature extraction in this study to estimate the annual energy consumption of a building stock in Seattle. Using the original dataset and external sources of information, new features within the Climate, Demographics, and Building Use Type categories were created. A non-parametric machine learning model called Random Forest Regressor was then trained over eight scenarios with different feature combinations. The results showed that Scenario-7 outperformed Scenario-1, which is the baseline scenario, with a 6.8% decrease in MAPE and a 30.8% increase in the R2 Score. Such improvements were achieved through the integration of the new features from the Demographics and Building Use Type categories. After analyzing the features' effect on the building energy consumption, it was seen that the feature EUI by Building Type (kBtu/sf) is the most critical feature in the model's estimation process. This outcome underlines the importance of archetypes that represent

buildings with similar energy performance characteristics in analyzing the energy demand of urban building stocks.

This study showed that feature extraction can be a good choice for urban planners in estimating the energy demand of building stocks in cities when the available datasets do not allow for performing accurate estimation. Similarly, this study addresses the abundance of external information sources that can be associated with the parameters affecting building energy performance. For example, datasets regarding the building envelope properties, climatic and geospatial data, and demographic structure of the districts might preserve valuable insights into the building materials or microclimate effect in the neighborhoods. To that end, it might be possible to integrate original datasets with highly correlated features and improve the accuracy of the models estimating the energy consumption of urban building stocks. A possible future work could be collaborating with the municipalities that can provide valuable datasets. In this way, the performance of data-driven urban building energy models can be enhanced.

References

1.11. Ensemble methods (2023). Available at: https://scikit-learn.org/stable/modules/ensemble.html#forest.

2020 Building Energy Benchmarking | City of Seattle Open Data portal (2021). Available at: https://data.seattle.gov/dataset/2020-Building-Energy-Benchmarking/auez-gz8p.

2022 Global Status Report for Buildings and Construction (2022). Available at: https://www.unep.org/resources/publication/2022-global-status-report-buildings-and-construction.

4.2. Permutation feature importance (2023). Available at: https://scikit-learn.org/stable/modules/permutation_importance.html.

Alpaydin, E. (2020) Introduction to Machine Learning, fourth edition. Amsterdam, Netherlands: Amsterdam University Press.

Cerezo Davila, C., Reinhart, C.F. and Bemis, J.L. (2016) "Modeling Boston: A workflow for the efficient generation and maintenance of urban building energy models from existing geospatial datasets," Energy, 117, pp. 237–250. Available at: https://doi.org/10.1016/j.energy.2016.10.057.

Commercial Reference Buildings (2011). Available at: https://www.energy.gov/eere/buildings/commercial-reference-buildings.

Granell, R. et al. (2022) "A reduced-dimension feature extraction method to represent retail store electricity profiles," Energy and Buildings, 276, p. 112508. Available at: https://doi.org/10.1016/j.enbuild.2022.112508.

Hancer, E. (2020) "A new multi-objective differential evolution approach for simultaneous clustering and feature selection," Engineering Applications of Artificial Intelligence, 87, p. 103307. Available at: https://doi.org/10.1016/j.engappai.2019.103307.

Hong, T. et al. (2020) "Ten questions on urban building energy modeling," Building and Environment, 168, p. 106508. Available at: https://doi.org/10.1016/j.buildenv.2019.106508.

IEA (2022), Buildings, IEA, Paris https://www.iea.org/reports/buildings, License: CC BY 4.0

Ross, S. (2020) Introduction to Probability and Statistics for Engineers and Scientists. Maarssen, Netherlands: Elsevier Gezondheidszorg.

Seattle Parks And Recreation Park Addresses | City of Seattle Open Data portal (2016). Available at: https://data.seattle.gov/Parks-and-Recreation/Seattle-Parks-And-Recreation-Park-Addresses/v5tj-kqhc.

Seattle, Washington: 28 Zip Codes (2023). Available at: https://www.bestplaces.net/find/zip.aspx?st=wa&city=5363000.

Seattle Washington ZIP Code Map (2023). Available at: https://www.zipdatamaps.com/en/us/zip-maps/wa/city/borders/seattle-zip-code-map.

Shalev-Shwartz, S. and Ben-David, S. (2014) Understanding Machine Learning: From Theory to Algorithms. 1st edn. Cambridge University Press.

Wang, C.-K. et al. (2020) "Bayesian calibration at the urban scale: a case study on a large residential heating demand application in Amsterdam," Journal of Building Performance Simulation, 13(3), pp. 347–361. Available at: https://doi.org/10.1080/19401493.2020.1729862.

Zhou, Y. et al. (2020) "Passive and active phase change materials integrated building energy systems with advanced machine-learning based climate-adaptive designs, intelligent operations, uncertainty-based analysis and optimisations: A state-of-the-art review," Renewable and Sustainable Energy Reviews, 130, p. 109889. Available at: https://doi.org/10.1016/j.rser.2020.109889.