UTILIZING EXTERNAL DATA FOR ENHANCING BUILDING ENERGY DEMAND ESTIMATION

Author(s): Bolluk, Said; Seyis, Senem; Aydogan, Reyhan

Presenting Author: Bolluk, Said

Presenter's E-Mail: said.bolluk@ozu.edu.tr

ABSTRACT

Building energy demand estimation plays a crucial role in constructing energy-efficient building stocks. However, most studies adopting a data-driven approach feel the deficiency of datasets with building-specific information in building energy estimation. Considering the great potential to enhance the quality of inference from datasets, the research objective of this study is to increase the accuracy of data-driven models by incorporating additional features obtained from external data sources, such as weather data, natural hazard risk maps, and demographic data. To that end, the original and external datasets are utilized in feature extraction, and the buildings' energy consumption is estimated using a nonparametric regression model. The results show that an 6% error reduction was achieved through the inclusion of new features, which indicates that feature extraction can be valuable in building energy demand estimation.

Keywords*: Machine Learning, Urban Building Energy Demand Prediction, Feature Extraction*

1. INTRODUCTION

Understanding the building stocks energy demand in cities has a great importance since the buildings accounted for the 34% of the world's overall energy consumption in 2021 (Buildings – Analysis - IEA, no date). Such an energy demand covering the embodied and operational consumption of the buildings brings along a huge carbon emission since the main energy source of the built environment is fossil fuels (2022 Global Status Report for Buildings and Construction, no date). Therefore, a great attention should be showed in assessing the building energy performance. In this sense, data-driven models provide accurate demand predictions with computational efficiency (Ang, Berzolla and Reinhart, 2020). Data-driven models are the statistical models that seek for 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 recorded properly or does not provide relevant information on building energy consumption patterns. Thus, data-driven models might require data enhancing.

Data enhancing can be interpreted in different ways. For example, incorporating new features to an existing dataset might improve the learning process (Hancer, 2020). On the contrary, some features might be redundant or misleading in an existing dataset so removing them could benefit the model accuracy (Granell et al., 2022). It is important to master the context and examine the dataset in detail for feeding statistical models with the most efficient feature combinations. Adding new features to a dataset can be done by processing the existing features or referring to an external dataset. Modelers search for the hidden relationships between the features to create new ones when the existing dataset is used for feature extraction. As unsupervised machine learning models, clustering models can be used in such practices to gather and label instances with similar patterns (Johari et al., 2020). Furthermore, external datasets might be a good source of information when the existing dataset lacks information. Modelers can utilize external datasets with specific or even more general information which the existing dataset lack.

Urban building energy models with data-driven approaches can benefit from feature extraction. The existing datasets in these practices 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). Therefore, the research objective of this study is to improve the accuracy of data-driven models utilized in urban building energy demand estimation using data-enhancing techniques. To that end, the residential houses in the city of Seattle are selected for a case study. The original dataset and external datasets are utilized in feature extraction. Various feature combinations are assessed, and the buildings' energy consumption is estimated using a nonparametric regression model. The contribution of the newly added features is analyzed using regression evaluation metrics. Finally, the achieved and potential improvements through feature extraction are discussed.

2. METHODOLOGY

The main 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 the address, use type and annual energy consumption information of more than three thousand buildings to determine the energy performance class of the city's building stock. It contains 3,628 building recordings (instances) with 42 features. Jupyter Notebook was used for coding purposes (Project Jupyter, no date), and the Scikit-learn library was used for statistical calculations in the study (Pedregosa et al., 2011).

2.1. Data Preprocessing

There are different use types, such as residential buildings, public facilities, and hospitals, in the dataset. This might be problematic in demand estimation since the diversity raised from the building use type complicates to derive similar consumption patterns. Moreover, some facilities have more than one building which directly affect the annual energy consumption. Therefore, only the residential buildings with one building in the facility were selected for the study. The annual energy consumption was recorded separately for different end uses, which are natural gas, electricity, and steam use. For simplicity, end use consumptions were summed and collected under the **Total Energy Use** column recorded in thousands of British thermal units (kBtu).

A great attention is needed in data imputation if the aim is recovering the missing or incorrect entries. However, this is not the objective of this study. Plus, the remaining part of the dataset after removing the missing, incorrect, and outlier values is enough to make a reasonable estimation. Therefore, data imputation part was kept simple. For example, instances with zero total energy consumption were removed since it is impossible for a building to have a zero-energy demand. Some columns were considered redundant and thus removed. For example, the city, state, and data year features are not necessary in 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 do not have a prior knowledge about the total energy consumption of a building. These features were directly derived from the total consumption value. The latitude and longitude features were hold as they might provide valuable insight into the micro-climate effects of the city's different districts, such as the thermal interaction between the neighboring buildings. Finally, the dataset contains 1,854 building recordings (instances) with nine features. The selected features for the rest of the study are presented in Table 1.

Table 1. Original Feature List

Feature Name	**Data Type**
Building Type	Categorical
Neighborhood	Categorical
Council District Code	Categorical
Latitude	Numerical
Longitude	Numerical
Year Built	Numerical
Number of Floors	Numerical
Property GFA Building(s)	Numerical
Total Consumption (kBtu)	Numerical

2.2. Feature Extraction

Original dataset and external datasets were used to derive new features that can affect the prediction accuracy. These features were collected under three main features: Energy Use Type, Climate Type, and Energy Efficiency Class.

2.2.1. Energy Use Type

Buildings in the dataset have different consumption rates in terms of end uses. For example, some buildings require electricity more than natural gas or otherwise. Such patterns might tell us about the main energy source of the buildings. Therefore, a new feature called **Energy Use Type** were created using the end use consumptions. In this sense, if any of the end uses forms more than 80% of the total energy consumption for a building, the building was labeled as the specific end use (e.g., Electricity or Natural Gas). If the previous condition was not met for any of the end uses, the building was labeled as Mixed. The steam use was not the determinant end use for any of the buildings, thus there is no such label.

2.2.2. Climate Type

Climate conditions have a huge impact on buildings' energy demand 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 to create the second main feature Climate Type. It should be noted that the climate can be the determinant of the building energy use type. This is because the local climate affects the type of building energy demand whether it is cooling or heating demand. Since the end-uses of Natural Gas and Electricity are correlated with cooling and heating demand, the Energy Use Type feature was evaluated in creating the Climate Type feature.

The first sub-feature is called **Landmark**. This feature specifies the geographic conditions of the buildings, such as Coastal, Green, and Terrain. The latitude and longitude features were used to detect the major coastal and forest regions on Google Maps. The instances located in these regions were labeled accordingly. The rest were assumed to be in the terrain regions.

The next sub-features are related with the weather conditions. It is very important to observe the weather to determine the climate condition of a region. The annual weather statistics, including the **Wind Speed**, **Daytime Temperature**, and **Nighttime Temperature**, were obtained from an online platform providing weather forecast with real measurement (Windfinder.com, no date). The latitude, longitude and neighborhood features from the original dataset were utilized to detect the nearest weather station. Following this, the climate type for the buildings were determined using an unsupervised machine learning technique, clustering. Clustering investigates similar patterns in a dataset to create similar clusters or groups. These clusters are the combinations of the feature density distributions (Alpaydin, 2020). They include all the information coming from every single feature but also, they might provide more information with combining features. A clustering model called K-Means Clustering was used to determine the climate type of the instances. K-Means Clustering is a model that group instances under a predefined number of clusters (2.3. Clustering, no date). For example, if x is an instance with multiple features, k is the number of clusters, m is the cluster center located in the feature space of the dataset, and E is the error function:

$$E(m_i|X) = \sum_t \sum_i b_i^t \|x^t - m_i\|^2 \qquad \text{▪ } X = \{x^t, y^t\}_{t=1}^N \quad \text{▪ } i = 1, \dots, k \tag{1}$$

Then, the aim is to minimize the error function that calculates the distance between instances and cluster centers (Equation 1). The function b takes the value of one if the distance is the minimum among all the clusters (Alpaydin, 2020). Otherwise, it takes the value of zero. The model is initiated with random cluster centers and for each iteration, the cluster centers are changed according to the gradient of the total error. This way, the model approaches to convergence where each cluster become homogeneous at a certain level.

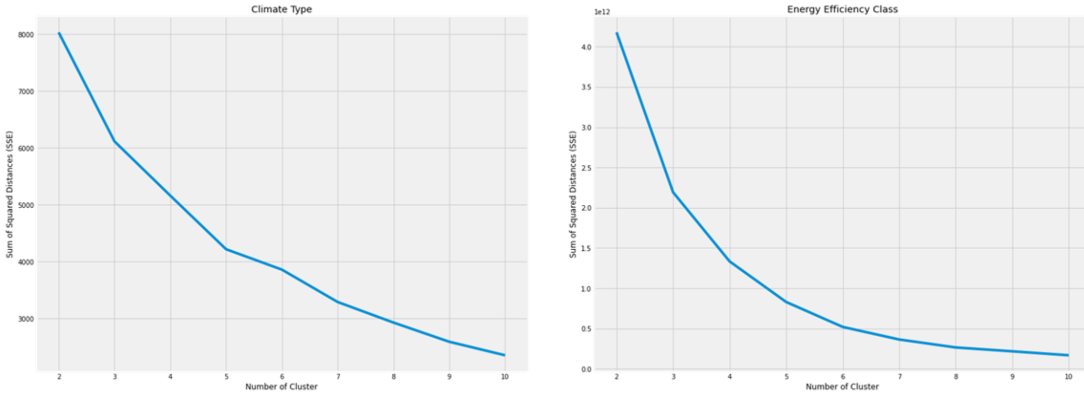

Figure 1. Elbow point graphs for k-means clusterings: Climate Type (left) and Energy Efficiency Class (right)

The features used in clustering to obtain the Climate Type feature are represented in Table 2. The optimal cluster number was five according to the elbow point in Figure 1. The elbow point graph illustrates the error reduction as the cluster number increases. The pattern here is understandable since the more clusters mean the more homogenous clusters and thus less error. However, after some point, increasing the cluster number does not provide a notable error reduction but weakens the generalization capacity of the model. Therefore, a trade-off should be made among the error and the model complexity. After training the clustering model with five clusters, which are the four different climate types, the instances were labeled according to the climate type.

2.2.3. Energy Efficiency Class

The geographical and demographical conditions of a district might help us to predict the building energy demand. For example, natural disaster risks can be determinant on the structural quality as well was the envelope properties of buildings (Shen, Zhou and Shrestha, 2021). In this sense, from external datasets, the earthquake risk map (Emergency Management - Hazards - Earthquake, no date) and the flood risk map (Emergency Management - Hazards - Floods, no date) of Seattle were used to determine the risk groups for the instances. The Neighborhood feature was consulted when labeling the instances with the natural hazard risk groups.

Similarly, the demographical profile of a district can be a great descriptor for the energy consumption patterns of buildings. For example, the number and the type of the nearby facilities, such as malls, hospitals, and factories, can hold valuable insights into the structural quality of a building and thus its energy efficiency. To that end, the number of the commercial, industrial, and public facilities were derived from Seattle Open Data (Office of Planning & Community Development - Land Use Zoning and Permitting, no date). The next step was to perform clustering to determine the energy efficiency class of the instances using a set of feature combinations. (Table 2). The optimal cluster number was five according to the elbow point graph in Figure 1. The instances were labeled according to the energy efficiency class after training the K-Means

Clustering model with five clusters. The datasets with different feature combinations are presented in Table 3.

Table 2. Features used in clustering

Climate Type	**Energy Efficiency Class**
Neighborhood	Building Type
Council District Code	Neighborhood
Latitude	Council District Code
Longitude	Year Built
Energy Use	Number of Floors
Landmark	Property GFA Building(s)
Wind Speed (kts)	Earthquake Risk
Daytime Temperature (C)	Flood Risk
Nighttime Temperature (C)	Commercial
-	Industrial
-	Public Utilities

2.3 Model Selection

This study forms a regression task with a continuous target variable called Total Energy Consumption. Here, the aim is to predict the buildings energy demand by analyzing at the available features and analyzing their relationships with the target variable. However, the existing and generated features does not provide a great correlation with the target feature in this study. This is because the existing dataset lacks some of the most important 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 (Alpaydin, 2020). In such cases, parametric models fail to satisfy accurate predictions. This is because parametric models assume a certain distribution for the features of a dataset and estimate the parameters of that distribution to make predictions. On the other hand, non-parametric models do not need a certain probability distribution to make predictions (Alpaydin, 2020). These models analyze a small sub-set of instances rather than the whole dataset and derive more complex patterns from the data. Therefore, the non-linearity provided from such models can be sufficient to 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 target to understand complex patterns rather than generalizing the correlation. However, considering the low correlation between the features and the target, the non-parametric models can provide effective regression results.

In this sense, Random Forest Regressor was selected to perform the building energy demand estimation. Random Forest Regressor is an ensemble learning method utilizing many random estimators called Decision Trees (1.11. Ensemble methods, no date). A decision tree is an algorithm that adopt a hierarchical order by splitting the data according to the features (Alpaydin, 2020). Each split forms a decision node. For example, if a random feature selected for the first split is Energy Use Type, there will be tree decision nodes since there are three different use types: Electricity, Natural Gas, and Mixed. After each split, the nodes are stretched and become homogenous. For example, if x is an instance with multiple features, m is the node number, N is the number of instances reaching to the node m, y is the target variable, g is the estimation, and E is the error function for node m:

$$E_m = \frac{1}{N_m} \sum_t |y^t - g_m| b_m x^t \qquad (2)$$

- $X = \{x^t, y^t\}_{t=1}^{N}$
- $g_m = \frac{\sum_t b_m x^t y^t}{b_m x^t}$

Then, the objective is to minimize the overall absolute difference between the targets and predictions (Equation 2). The function b takes the value of one if the x reaches to 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 a huge variance between the results of the tree models initiated with different random seed. By averaging the performance of these trees, however, Random Forest algorithm can converge fast and decrease the variance in estimations (Shalev-Shwartz and Ben-David, 2014).

The Random Forest algorithm used in this study have hundred different Decision Tree Regressors with random splitting strategy. The error function that the algorithm optimizes is the mean absolute error. Another hyperparameter of the forest regressor is the maximum depth of the nodes of each tree. The depth refers to the maximum split for each random splitting. A validation curve was used to determine 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 id the model has a generalization capacity, it can represent the population but not only the sample space, the existing dataset. Validation curve helps us to determine the optimal hyperparameters of predictive models by observing the error or accuracy change over different hyperparameter setting on both the training and test data. Therefore, different tree depths ranging between two and twenty were evaluated. However, before moving with the validation curve, the regression evaluation metrics should be elaborated. In this study, two different regression metrics were employed: Coefficient of determination (R2) and Mean Absolute Percentage Error (MAPE). R2 is a metric that examines to what extend the estimator outperforms the mean estimator (Ross, 2020). This metric evaluates how much the estimator explains the variance in target values

(Equation 3). The MAPE proportions the absolute difference between the target and estimations (Equation 4). R2 is a metric assessing the model's generalization capacity whereas MAPE analyzes the model's accuracy. For example, if y is the target, g is the estimation and r is the mean of the target:

$$R^2(y,g) = 1 - \frac{\sum_1^t (y^t - g^t)}{\sum_1^t (y^t - r)} \qquad ▪ \quad X = \{x^t, y^t\}_{t=1}^N \tag{3}$$

$$MAPE(y,g) = \frac{1}{N}\sum\nolimits_1^t \frac{|y^t - g^t|}{|y^t|} \qquad ▪ \quad X = \{x^t, y^t\}_{t=1}^N \tag{4}$$

The validation curve was utilized over the original dataset (Table 3) for the consistency of the result comparison. Because the hyperparameters should be defined for the original data and fixed for the rest of the datasets with different feature combinations so that the contribution of the generated features can be assessed. The optimal tree depth was four according to the validation curves in Figure 2.

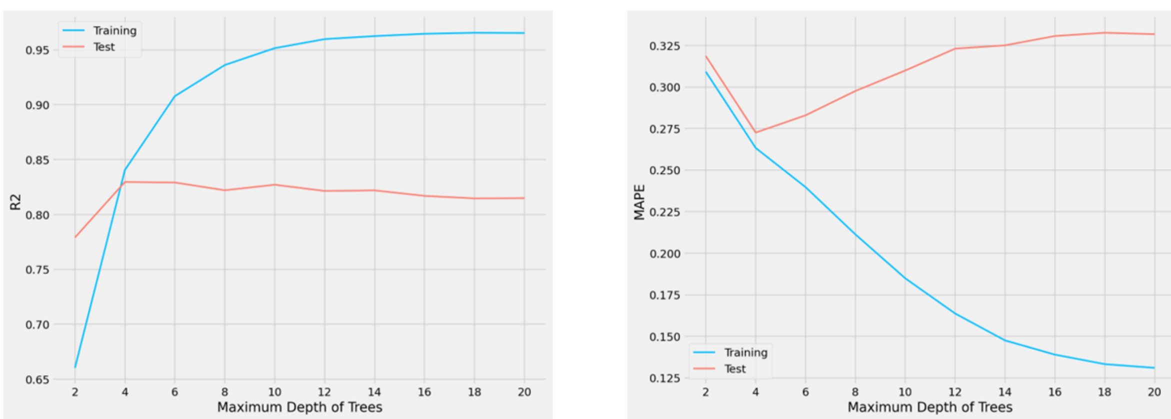

Figure 2. Validation curves for the original dataset: R2 vs. Maximum Tree Depth (left) and MAPE vs. Maximum Tree Depth (right)

As it is seen in Figure 2 that the training data performs well as the maximum tree depth increases. This is because the model starts recognizing the training data as the complexity increases. However, this behavior of the model decreases the accuracy for the test set after a certain depth. It is evident from the validation curves above that increasing the model complexity caused the model losing its generalization power. Several Random Forest Regressors with different tree depths and a Linear Regression model were trained and their performance in making estimation analyzed in

Figure 3 to better illustrate model the trade-off between model's complexity and generalization. The distribution of the feature Total Ground Floor Area (GFA) in square feet and the target variable Total Energy Consumption was scattered and different regressors were fitted in Figure 3.

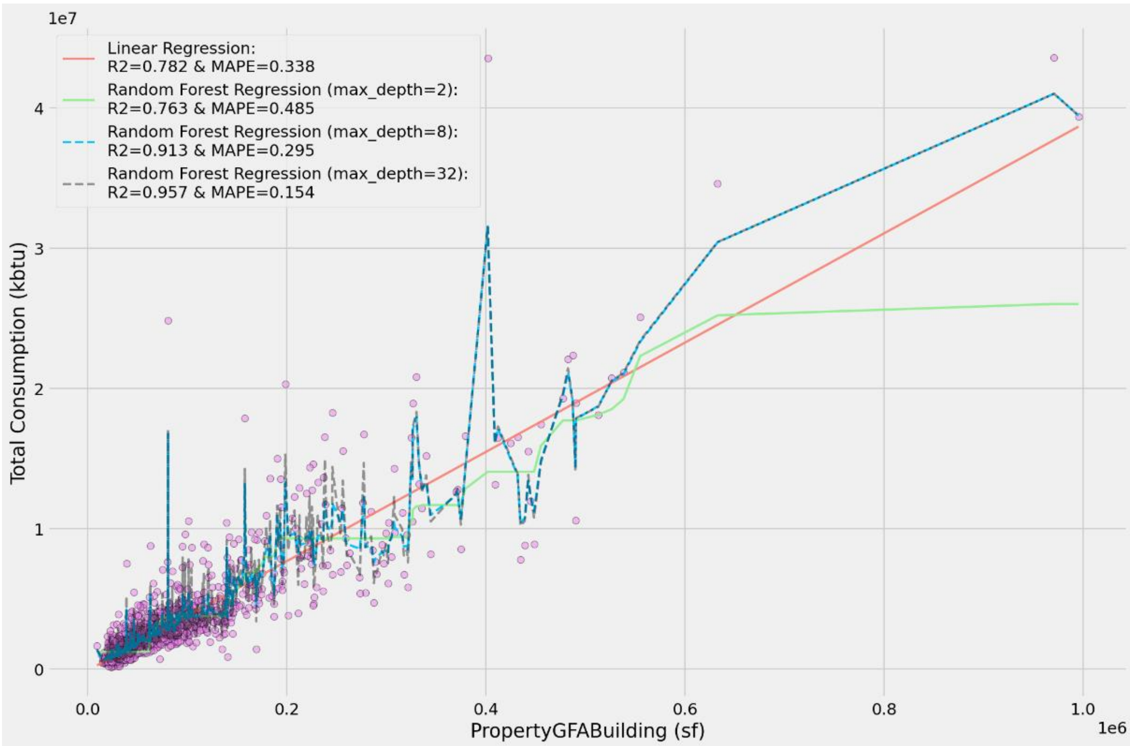

Figure 3. Model performance vs. complexity

The non-parametric regressor, which are the Random Forest Regressors, separate 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, the models tend to memorize the data and thus loses its generalization capacity. This is problematic because when the new instances are introduced, the model is not able to make accurate estimations. The Random Forest regressor with maximum tree depth of four were used to assess the datasets with different feature combinations and their results were evaluated. The datasets were split into training and test sets to prevent overfitting with 20% of the instances belonging to the test set.

Table 3. Datasets with different feature combinations

Feature Name	Normal Data	Climate Data	Efficiency Data	Mixed Data
Building Type	+	+	+	+
Neighborhood	+	+	+	+
Council District Code	+	+	+	+
Latitude	+	+	+	+
Longitude	+	+	+	+
Year Built	+	+	+	+
Number of Floors	+	+	+	+
Property GFA Building	+	+	+	+
Energy Use	-	+	-	+
Climate Type	-	+	-	+
Energy Efficiency Class	-	-	+	+

3. RESULTS AND DISCUSSION

The complete regression results are given in Table 4. It should be noted that the linear regression model performed worse than the Random Forest Regressor for each dataset. This indicates the power of the nonparametric models in handling datasets with low correlation and features that contains irregular probability distributions. The results in Table 4 show that Climate Data and Mixed Data have the least regression error whereas the Climate Data has the highest R2 Score. It is evident that the Climate Data outperformed the other datasets in terms of accuracy and generalization power. Even though the accuracy of the Mixed Data is higher than the Normal data, the contribution here comes from the Climate Type feature rather than the Energy Efficiency Class feature.

Table 4. Regression results

Model	Dataset	R2 Score	Mean Absolute Percentage Error
Random Forest Regressor	Original Data	0.685	0.281
Random Forest Regressor	Climate Data	0.694	0.265
Random Forest Regressor	Efficiency Data	0.684	0.281
Random Forest Regressor	Mixed Data	0.693	0.265
Linear Regression	Original Data	0.675	0.338
Linear Regression	Climate Data	0.706	0.313
Linear Regression	Efficiency Data	0.675	0.337
Linear Regression	Mixed Data	0.706	0.315

It is seen that the Random Forest regression model performed 6% better in terms of accuracy when the new feature, Climate Type, is included. Plus, the models' generalization power was slightly increased thanks to the features in the Climate Data. These improvements are relative and very small. However, considering that the features in the original dataset have a very small correlation with the target, it is not rational to expect a huge error reduction through the inclusion of new features. Table 5 illustrates the ten most important features affecting the regression results.

A Decision Tree model aims to split instances to obtain the purest clusters in each branch of the tree (Alpaydin, 2020). In each iteration, the model seeks for the optimal split that decrease overall the impurity in the branches most. However, the impurity term is used in classification problems. Therefore, a reasonable error evaluation metric for a regression task should be defined. In this study, the split quality of the Decision Trees and thus the Random Forest Regressors were evaluated using the absolute error (Equation 2). According to Table 5, only the feature Property GFA Building has a remarkable impact on the regression results with 85% decrease in impurity (absolute error). Energy Use Type and Energy Efficiency Class are the only generated features ranked among the top ten. The rest of the features do not or barely impact the regression results.

Table 5. Mean decrease in absolute error per feature

Feature	Mean Decrease in Absolute Error
Property GFA Building	0.884
Number of Floors	0.046
Energy Use: Natural Gas	0.022
Energy Use: Electricity	0.019
Year Built	0.008
Building Type: Multifamily HR (10+)	0.006
Latitude	0.003
Energy Efficiency Class: 0	0.003
Longitude	0.002
Neighborhood: DOWNTOWN	0.001

This study has some limitations. The first one is the deficiency of the original dataset to make reasonable estimations or derive valuable inference that might improve the accuracy of regression. The dataset is an energy benchmarking dataset with almost no building energy-related parameters, such as air infiltration ratio, structural material, and window-to-wall ratio (Sokol, Cerezo Davila and Reinhart, 2017). This complicated obtaining accurate regression results. Furthermore, since the dataset does not represent the building envelope properties or the mechanical system details, it is hard to gather such information from external datasets. The model selection part involves another limitation of this study. The regression model is determined as a non-parametric model due to the low correlation raised from the original dataset. However, some parametric models (e.g., Multilayer Perceptron Regressor) can be valuable to cope with such a low correlated data if extensive hyperparameter tuning is performed. Nevertheless, the computational capacity of this study did now allow us to perform a detailed model selection procedure. Despite all these difficulties, however, this study showed a potential in increasing the regression accuracy with the help of adding new features. Either from the original datasets or using external datasets, the building energy demand prediction can be enhanced via data manipulation integrated with domain knowledge.

4. CONCLUSION

A building energy benchmarking dataset was used and enhanced with feature extraction in this study to estimate the annual energy demand of Seattle's building stock. Using the original dataset and external sources of information, Energy Use Type, Climate Type, and Energy Efficiency features were created. The datasets with different feature combinations were then trained using a non-parametric machine learning model called Random Forest Regressor. The results suggested that an 6% relative decrease in the mean absolute percentage error was introduced by the addition of the Energy Use Type feature. It was evident that the features' correlation with the target variable obstructed obtaining accurate regression results. Moreover, the irrelevant or missing information that the original dataset contains about the building energy characteristics restrained extracting valuable information from external data sources. Thus, the original dataset was not enhanced adequately through feature extraction. However, this study addresses the abundance of external sources of information that can be associated with the parameters affecting building energy performance. For example, datasets regarding the natural hazard risk, climatic and geospatial data, and demographic structure of the districts might reserve valuable insights into the building materials or microclimate effect of the neighborhoods. In this context, it is possible to equip original datasets with highly correlated features and improve the accuracy of the models estimating the energy demand of urban building stocks. A possible future work can be collaborating with the municipalities which can provide valuable datasets and improve the performance of the models assessing the energy demand of the urban building stocks. Therefore, this study shows that building energy demand prediction can be enhanced via feature extraction integrated with domain knowledge.

REFERENCES

1.11. Ensemble methods (no date). Available at: https://scikit-learn.org/stable/modules/ensemble.html#forest.

2.3. Clustering (no date). Available at: https://scikit-learn.org/stable/modules/clustering.html#k-means.

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 (no date). Available at: https://www.unep.org/resources/publication/2022-global-status-report-buildings-and-construction.

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

Ang, Y.Q., Berzolla, Z.M. and Reinhart, C.F. (2020) "From concept to application: A review of use cases in urban building energy modeling," Applied Energy, 279, p. 115738. Available at: https://doi.org/10.1016/j.apenergy.2020.115738.

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.

Emergency Management - Hazards - Earthquake (no date). Available at: https://www.seattle.gov/emergency-management/hazards/earthquake.

Emergency Management - Hazards - Floods (no date). Available at: https://www.seattle.gov/emergency-management/hazards/floods.

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

Johari, F. et al. (2020) "Urban building energy modeling: State of the art and future prospects," Renewable and Sustainable Energy Reviews, 128, p. 109902. Available at: https://doi.org/10.1016/j.rser.2020.109902.

Office of Planning & Community Development - Land Use Zoning and Permitting (no date). Available at: https://www.seattle.gov/opcd/population-and-demographics/about-seattle#landusezoningandpermitting.

Pedregosa, F. et al. (2011) "Scikit-learn: Machine Learning in Python," Le Centre Pour La Communication Scientifique Directe - HAL - memSIC [Preprint].

Project Jupyter (no date). Available at: https://jupyter.org/.

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

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

Shen, Z., Zhou, H. and Shrestha, S. (2021) "LCC-based framework for building envelope and structure co-design considering energy efficiency and natural hazard performance," Journal of Building Engineering, 35, p. 102061. Available at: https://doi.org/10.1016/j.jobe.2020.102061.

Sokol, J., Cerezo Davila, C. and Reinhart, C.F. (2017) "Validation of a Bayesian-based method for defining residential archetypes in urban building energy models," Energy and Buildings, 134, pp. 11–24. Available at: https://doi.org/10.1016/j.enbuild.2016.10.050.

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.

Windfinder.com (no date) Windfinder.com - Most popular locations in Washington / United States of America. Available at: https://www.windfinder.com/locations/united_states/washington.

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.