Applied Energy 233–234 (2019) 244–253

Contents lists available at ScienceDirect

Applied Energy

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

Grading buildings on energy performance using city benchmarking data

Sokratis Papadopoulos, Constantine E. Kontokosta*

Department of Civil and Urban Engineering and Center for Urban Science and Progress, New York University, 370 Jay Street, 12th Floor, Brooklyn, NY 11201, United States

HIGHLIGHTS

- Create city-specific energy performance grading model using city disclosure data.
- Method to grade buildings that is interpretable, reproducible, scalable, and robust.
- Develop a clustering algorithm to grade buildings on performance.
- Validated approach demonstrates significant improvements over current methods.

ARTICLE INFO

Keywords:
Building energy performance
City-specific energy benchmarking
Energy efficiency labeling
Machine learning
Energy disclosure data
XGBoost

ABSTRACT

As the effects of anthropogenic climate change become more pronounced, local and federal governments are turning towards more aggressive policies to reduce energy use in existing buildings, a major global contributor of carbon emissions. Recently, several cities have enacted laws mandating owners of large buildings to publicly display an energy efficiency rating for their properties. While such transparency is necessary for market-driven energy reduction policies, the reliance on public-facing energy efficiency grades raises non-trivial questions about the robustness and reliability of methods used to measure and benchmark the energy performance of existing buildings. In this paper, we develop a building energy performance grading methodology using machine learning and city-specific energy use and building data. Leveraging the growing availability of data from city energy disclosure ordinances, we develop the GREEN grading system: a framework to facilitate more accurate, fair, and contextualized building energy benchmarks that account for variations in the expected and actual performance of individual buildings. When applied to approximately 7500 residential properties in New York City, our approach accounts for the differential impact of design, occupancy, use, and systems on energy performance, out-performing existing state-of-the-art methods. Our model and findings reinforce the need for more robust, localized approaches to building energy performance grading that can serve as the basis for data-driven urban energy efficiency and carbon reeduction policies.

1. Introduction

1.1. Background and motivation

The importance of “greening” existing buildings in cities cannot be overstated. The Intergovernmental Panel on Climate Change highlights that existing buildings are responsible for more than one-third of global primary energy consumption and greenhouse gas emissions [1]. At the same time, the building sector has the highest potential for cost-effective and long-term carbon reductions among all economic sectors [2]. While carbon reduction policies and targets have historically been adopted at the federal level, city leaders are increasingly taking action to reduce energy use and carbon emissions across urban areas. Along with climate change mitigation, fiscal and economic benefits associated with improved energy efficiency have prompted municipalities to focus policy initiatives on long-term sustainability [3]. In the United States, Europe, and Australia, market-driven policy tools to reduce energy use in buildings have centered on information disclosure and transparency. Building energy benchmarking, which refers to the process of assessing the energy performance of buildings compared to their peers, constitutes the basis for these initiatives, which are predicated on eliminating information asymmetries between the owners and users of buildings [4,5]. In the United States, more than 20 cities and local governments have passed energy benchmarking and disclosure laws as the foundation for city sustainability plans [6]. These benchmarking laws require building owners to annually report their energy

* Corresponding author.
E-mail address: ckontokosta@nyu.edu (C.E. Kontokosta).

https://doi.org/10.1016/j.apenergy.2018.10.053
Received 18 July 2018; Received in revised form 26 September 2018; Accepted 11 October 2018

consumption, adding transparency to real estate markets and clarity to energy-saving opportunities [3,7].

While energy data disclosure has been shown to drive energy use reductions in certain building types [8,9], cities are adopting more aggressive policy measures to further transform the energy efficiency market through economic incentives and competition [10]. From a top-down perspective, city decision-makers can identify poorly-performing buildings and promote more equitable and efficient regulations or incentive mechanisms to reduce emissions. From a bottom-up view, such schemes can help building owners understand their buildings' performance and expose them to competitive market pressures that (should) encourage greater energy efficiency. However, this type of performance grading relies on the ability to effectively and accurately establish expected and actual energy performance targets. Unlike grading in other industries, such as restaurant cleanliness grades or vehicle fuel efficiency ratings, building energy performance is influenced by a bundle of physical, mechanical, meteorological, and behavioral systems that interact to determine current and potential energy consumption patterns. Without understanding and capturing the dynamics and interactions of these systems, it becomes difficult to determine the optimal, real-world energy performance of a particular building, resulting in high levels of uncertainty when comparing energy efficiency. New York City recently enacted a law that requires large property owners to publicly display their energy efficiency grades (see Appendix A for further details). However, these grades, as in the case of Chicago, are based on the U.S. Environmental Protection Agency's EnergyStar score, an approach that has been heavily criticized in recent literature, mainly due to its high uncertainty, poor data quality, and model specification errors [10–13].

In this work, we critically assess the state-of-the-art in energy benchmarking and introduce the GREEN grading system; a new method for building energy performance measurement that accounts for the full range of factors that impact building energy use. Our method contextualizes these factors to the local metropolitan area, providing a needed city-specific performance baseline and grading system. We apply the proposed framework to New York City's large, multi-family residential building stock (approximately 7500 properties) and contrast it with its current EnergyStar-based grading. We find that our method explains more than 30% of the variance in energy use intensity, whereas the EnergyStar algorithm is not able to generalize at all when using city-specific data. This is one of the first large-scale studies to leverage city energy disclosure data to develop and test a contextualized building energy performance grading scheme. As more cities adopt the requirement to publicly display building energy performance grades, our method is directly relevant to local governments, energy policy makers, and building owners. Based on our findings, we catalyze the debate for revisiting building energy performance assessment by utilizing more robust approaches that provide greater certainty and flexibility in meeting individual city's policy needs and goals.

1.2. The need for a paradigm shift in energy benchmarking

The increasing availability of building energy data, enabled by the adoption of city energy disclosure ordinances and open data mandates, has led to a new interest in peer comparison and data analytics as tools to assess relative performance. Statistical benchmarking models have become quite popular, as these methods typically utilize machine learning algorithms that can map complex relationships between energy consumption and building characteristics using large samples of buildings [10,14–16]. Measured data, as opposed to modeled or simulated data that constitute the basis for physics-based building energy models, provide opportunities for researchers to bridge the "performance gap" encountered in engineering benchmarking methods [17,18]. Measured data provide a snapshot of a buildings "real-world" operational energy performance, rather than the "idealized" or hypothetical performance characteristics described by physical models and simulation software. However, there has yet to be consensus on the appropriate standard for statistical approaches to building energy benchmarking. Building energy consumption is a complex problem, with non-linear and sometimes unexpected interactions between architectural design, mechanical systems, occupant behavior, management quality, and the surrounding environment [19]. Additionally, a recent study that used energy disclosure data from various US cities found that the relationship between building characteristics, such as age and gross floor area, and energy use intensity (EUI) varies from city to city, beyond what could be accounted for by weather or climate variations alone [20].

Building energy grades could provide an important signal to the market to convey a critical, yet unobservable, component of green buildings: their energy performance. Therefore, grading buildings on their energy performance to drive efficiency gains needs to account for the range of factors that may influence consumption patterns in day-to-day operations. A building grading system that can be implemented by cities and other government agencies, and accepted and adopted by building owners and investors, must meet several criteria:

1. It must be understood by all potential stakeholders and end users. While it is not necessary that all can directly interpret the specific algorithm used, it must be made clear how different building features are considered and accounted for in the model and the resultant grade. It should also be possible to understand what changes would be necessary to move from one grade band to another.
2. It must account for the multitude of characteristics - and their interactions - that influence building energy use, focusing on those attributes that can be readily changed. For instance, we would not want to grade buildings based on the age of installed heating systems; this would have the effect of penalizing buildings with more efficient systems by controlling for this feature in the model.
3. It must be scalable and generalizable so that it can be deployed across a range of climate and market-specific conditions. Current rating tools are national in scope; these have been criticized because of the non-trivial variations in building types, operational parameters, occupant behavior, and local environmental conditions across cities and regions, as well as their data quality and coverage. The model must be able to ingest data from any specific geography and produce reliable grading results. Moreover, it must be developed using representative, sufficient, and up-to-date data sources that are publicly-available to ensure reproducibility.
4. Differences in grade bands must be statistically significant with a high degree of confidence. The marketplace must be able to have trust in the model such that a building rated 'A' is known to be superior to a 'B'-rated building with certainty.

In machine learning problems, we encounter the so called "bias-variance tradeoff", which refers to the opposing relationship between a models complexity and its ability to generalize. In the context of statistical energy benchmarking, we are interested in mapping the relationships between energy consumption and its drivers, while minimizing the effect of noise in the data. Various machine learning algorithms have been applied to building energy benchmarking data, from simple linear regression models to complex artificial neural network architectures. Linear models have been tested extensively, and they appear to be insufficient in capturing the non-linear relationship between building energy consumption and explanatory variables [10,21,22]. Nevertheless, EnergyStar, the predominant energy benchmarking approach in the US, is built on a multivariate linear regression model [23]. (See Appendix B for a detailed description on the method.)

Table 1 summarizes the limitations of the EnergyStar scoring method, as well as our proposed improvements. More complex machine learning models, such as artificial neural networks or ensemble learning methods, exhibit promising results in modeling building energy performance [24–28]. However, a limitation of neural networks is their

Table 1
Limitations in EnergyStar benchmarking approach and potential improvements.

Current limitations	Potential improvements
Linear model	Non-linear machine learning algorithms
Nationwide data	City-specific data
Limited sample (surveys, reference buildings)	Extended samples (energy disclosure policies)
Limited features	Extended features (physical, operational, qualitative)
Continuous scale (0–100)	Letter grade (A-D)

black-box nature, that makes the interpretation of the model difficult. On the other hand, ensemble learning methods have not been widely adopted, yet initial applications to building energy benchmarking show potential [21,24,26,29]. Tree-based ensemble learning is suitable for modeling complex, non-linear data, while allowing for result interpretability to some extent [25,26] through the contribution of features in the tree-building process (i.e. feature importance). Furthermore, novel feature attribution approaches have been recently developed to further improve ensemble models' interpretability [30,31]. Statistical benchmarking requires large and representative building energy data samples to yield robust models [10]. EnergyStar utilizes data obtained through a nationwide survey. The survey includes relatively small data samples (i.e. 322 residential properties), across nine US census regions. The small sample size along with its nationwide coverage limit EnergyStar's ability to account for heterogeneity in local building stocks and provide contextualized estimations at a more granular spatial level, such as the city or metropolitan area [12,32]. Additionally, the EnergyStar model specification includes only five features, neglecting important aspects related to energy consumption. Finally, an important aspect of energy benchmarking is the communication of the outputs to a wide range of stakeholders (e.g. building owners, tenants, policy makers, etc.) [10]. The 1–100 scale used for EnergyStar scores presents a specific numerical rating that belies the inherent uncertainty in the model estimates [12]. Although we do not explicitly argue against continuous scale grading, evidence from energy performance certifications in the European Union suggest that letter grades can be reflected in increased stakeholder awareness and real estate premiums [33,34].

2. Data and methods

2.1. The GREEN grading methodology overview

To address the limitations of current building energy benchmarking models, we propose GREEN grading; an approach that integrates localized energy disclosure data and machine learning methods (Fig. 1). Overall, our methodology can be split into three parts: (a) data preprocessing, (b) model selection, and (c) building energy grading.

Our primary variable of interest is weather normalized source energy use intensity (hereafter EUI). We elect to use source, rather than site, EUI in our models to reflect a comprehensive assessment of a building's energy efficiency and carbon emissions, including fuel source and production, delivery and transmission losses. Due to its self-reported nature, energy disclosure data often contain missing, misreported, or anomalous entries that need to be removed prior to analysis. The cleaning steps applied in both EUI and occupancy-related features are detailed in the following subsection.

The relationship between a building's EUI and its physical or occupancy characteristics has been often found to be non-linear (e.g. building age [35,36], weekly operating hours [10], number of occupants [37]). Given this condition, there is little rationale to support the use of linear models in energy benchmarking applications, beyond the benefits of interpretability of such methods. On the other hand, deep neural networks, a popular non-linear approach, have demonstrated high predictive accuracy. From the pool of non-linear algorithms, deep neural networks and ensemble learning methods demonstrate superior performance when copmared to other popular machine learning algorithms [21,26,29,38,39], although the former is computationally expensive, often requires large datasets for training, and has limited interpretability [21]. Hence, for data modeling, we use XGBoost, an ensemble learning algorithm based on gradient tree boosting, with proven capabilities for handling nonlinear datasets in a computationally efficient manner [40]. To maximize the model's predictive power and generalizability, we fine-tune its parameters via cross-validation. Having identified the model's optimal parameters, we fit the data using jackknife sampling and estimate the model residuals for each individual building. The reasoning behind the use of jackknife sampling is to avoid training bias and treat each building as "test set" in the residual calculation.

From the residuals, we calculate the energy performance ratio, defined as the reported EUI divided by the model-predicted EUI, to quantify a building's relative performance. In the last step, we use unsupervised learning to cluster the energy performance ratios and assign a letter-grade to each building, according to their respective group membership.

2.2. Data description

The main data source used to train our model is NYC's Local Law 84 (LL84) energy benchmarking data. LL84 covers all buildings with gross floor area that exceeds 50000 square feet, and mandates their owners to annually report energy and water usage, along with other building characteristics [10,24]. To include additional features in our model and better capture the drivers of energy consumption, we merge the latest LL84 database (i.e. 2016) with land use data from the Primary Land Use Tax Lot Output (PLUTO) data provided by the NYC Department of City Planning. We merge the two datasets on the Building Block Lot (BBL) number, a unique identifier for NYC properties, to avoid inconsistencies during the merging process. The merged sample consists of 13137 properties, 9611 of which are residential buildings. Prior to modeling, we perform substantial data cleaning to remove errors and outliers resulting from the self-reported nature of the data, similar to other research done on energy disclosure data [10,41–43]. First, we drop all properties with one or more of the following characteristics: missing/zero weather normalized EUI, missing/zero gross floor area, or missing or duplicated building identification numbers. Given the log-normal distribution of EUI, we then apply a logarithmic transformation to the EUI values, and filter out observations falling outside the threshold of two standard deviations from the sample's mean as outliers [9,41]. We do this independently for the sample of residential buildings to account for the variations in the EUI distributions across property types. Finally, we remove values above the 99th percentile and below the 1st percentile for occupancy-related features, such as unit density and number of bedrooms. The resultant cleaned data set includes 7487 residential buildings. The dropped observations can mainly be attributed to the manual data collection process used for the LL84 data, highlighting the issue of data entry errors by non-expert users and those unaccustomed to tracking building energy use.

2.3. Statistical learning algorithm

To capture the non-linear relationships in the data, we choose XGBooost [40], a scalable version of gradient tree boosting [44]. Unlike linear regression models previously used in building energy benchmarking [10,14–16,23], boosted trees are capable of fitting highly nonlinear data by learning higher interactions between features, while requiring minimum data preprocessing [45,46]. In building energy-related datasets, specifically, gradient boosting has shown great potential in modeling both granular (i.e. hourly) building energy consumption [21], as well as annual energy demand for urban scale applications

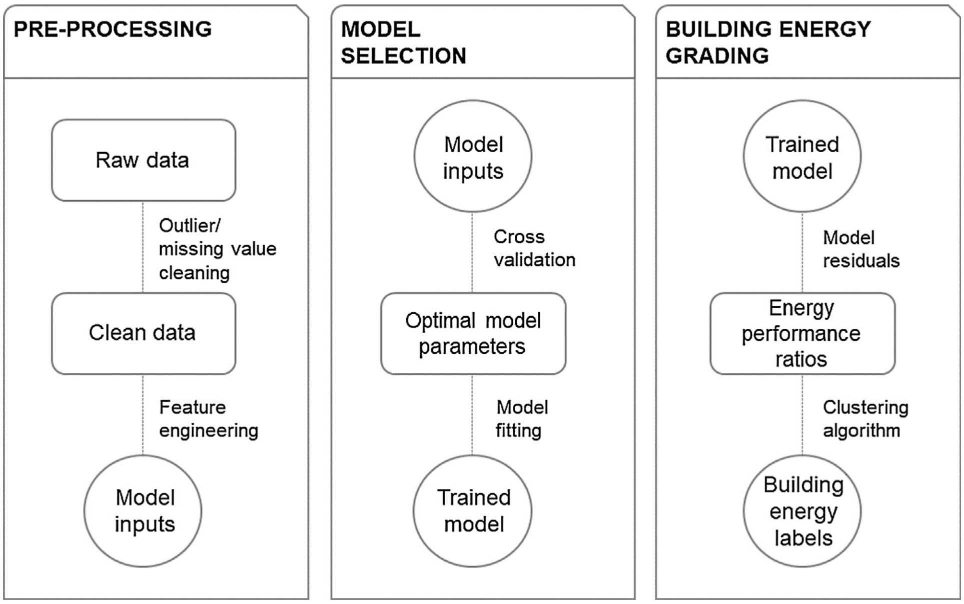

Fig. 1. GREEN grading methodology.

when compared with linear models [26,29].

2.3.1. Decision trees

Gradient boosting is an *ensemble learning* algorithm, consisting of multiple decision trees (i.e. base models) [44]. Each decision tree provides a different solution to the problem and their outputs are combined to yield the final output [45]. Decision trees partition the feature space in a set of regions using a series of hierarchical rules to approximate a simple function in each region (i.e. constant). In Fig. D.6, we illustrate the decision tree algorithm in a 2-dimensional feature space.

2.3.2. Gradient tree boosting - XGBoost

The gradient boosting algorithm trains sequential weak base models,[1] where each consecutive tree learns from the errors of the preceding ones (Eq. (1)). The algorithm strategically re-samples observations that were hard to predict by the previous models in order to provide useful information in the next model.

$$f(x) = f_0(x) + \sum_{m=1}^{M} \gamma_m f_m(x) \tag{1}$$

where $f_0(x)$ is the first learner, $f_m(x)$ the base model at boosting iteration m, and γ_m the weight of the m-th iteration.

XGBoost determines the gradient by solving Eq. (2).

$$\frac{\partial L(y, f^{m-1}(x) + f_m(x))}{\partial f_m(x)} = 0 \tag{2}$$

where L is the loss function to optimize and y is the ground truth for the target variable. Essentially, XGBoost is a more stochastic and regularized variant of gradient boosting. With the introduction of L1 and L2 regularization on top of the existing tree complexity regularization terms, XGBoost controls for overfitting, a commonly encountered drawback of gradient boosting [46]. Additionally, XGBoost allows for feature subsampling in both tree and split level to introduce additional randomness in the learning process.

2.4. Model selection

In the machine learning context, model selection (or hyper-parameter tuning) is the process of identifying the model parameters that maximize a learning algorithm's performance on a given dataset. Overtuning an algorithm's parameters results in a complex model and *overfitting*, meaning that the model learns the noise in the training data. On the other hand, a simple model might not be able to learn the patterns in the data, resulting in poor performance and (See Fig. D.7, [47] for an illustration of overfitting and underfitting on dummy data.) In the context of energy benchmarking, models should be complex enough to explain variations in energy consumption between different buildings, but not so complex as to capture the noise that by default exists in the energy data [10].

We elect to tune eight hyper-parameters, associated with both the general nature of the algorithm and each individual tree.

- The number of boosting iterations.
- The learning rate, referring to feature weight shrinkage in each boosting iteration.
- The maximum depth of each tree, controlling the complexity of the algorithm.
- Fraction of examples used to train each tree.
- Degree of purity in leaf node.
- Regularization weights.
- Fraction of features used to perform each node split.
- Fraction of features used to train each tree.

Prior to tuning the hyper-parameters, we split the data into training (80%) and testing (20%) sets. We specify a parameter grid to evaluate outputs through a 5-fold cross-validation (CV). For each parameter combination, we train the model using 4 folds of training and one fold for validation. We repeat the process 5 times, until each individual fold is used as a validation set and then average the scores. We use the root mean squared error as the CV performance metric. After cross-validating our model the optimal XGBoost hyper-parameter are as follows: *number of estimators:* 667, *subsample ratio:* 0.75, *learning rate:* 0.01, *maximum tree depth:* 8, *minimum sum of instance weight needed in a leaf node:* 5, *regularization:* 100, *fraction of features used to split:* 0.75, and *fraction of features used in each tree:* 0.75.

[1] A base learner is defined as "weak" when it performs slightly better than random guessing.

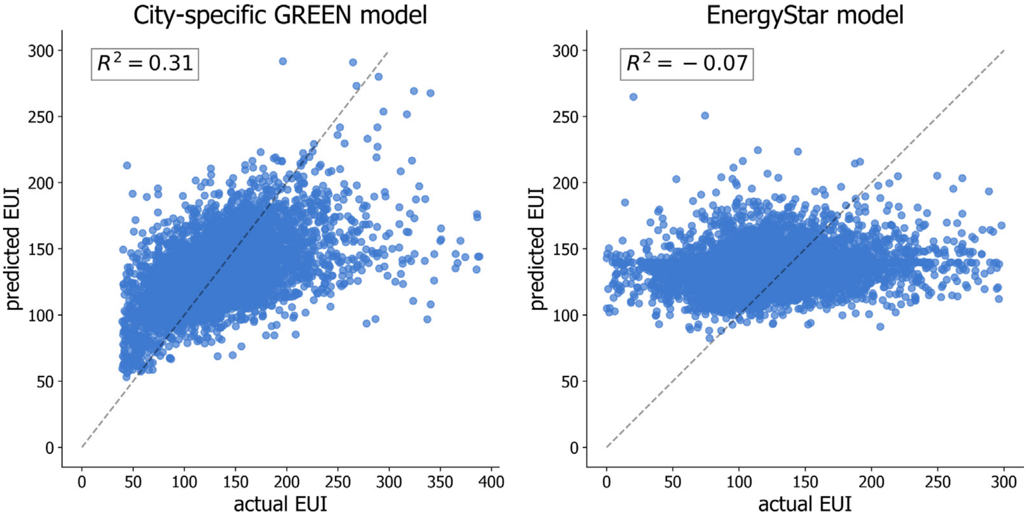

Fig. 2. Explained variance. The city-specific GREEN model (left) and national EnergyStar model (right).

2.5. Model interpretation

To interpret the importance of individual model features and understand drivers of EUI prediction, we compute the SHAP (SHapley Additive exPlanation) values, proposed by Lundberg, Erion and Lee [31]. Although XGBoost supports traditional feature importance reporting, these values can be inconsistent and not individualized for each prediction. SHAP builds on ideas from game theory [48] and local explanations [49], and unlike other popular feature attribution methods, such as gain or split count, SHAP values are individualized to each prediction and consistent. As an additive feature attribution method, SHAP develops an explanation model g that is a function of binary features:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i' \tag{3}$$

where $z_i' = 1$ if the feature is observed and $z_i' = 0$ otherwise, ϕ_i are the feature attribution values, and M is the number of features in the model. To calculate the feature attribution values, SHAP uses the traditional Shapley values [48] along with conditional expectations as follows:

$$\phi_i = \sum_{S \subseteq z' \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!} \left[f_x\left(S \cup \{i\} \right) - f_x(S) \right] \tag{4}$$

where S is the set of non-zero z' indexes and $f_x(S) = E[f(x)|x_S]$ is the expected value of the model f conditioned on S.

For more detailed description of the SHAP methodology, we refer interested readers in the work of [30,31].

2.6. Score calculation

Once we obtain the optimal model parameters, we train the XGBoost model and get the predicted EUI for each building. For each building in the dataset n, we train a model with $n-1$ samples, leaving one building out as a test set and then use the trained model to predict the particular building's EUI (EUI_{pred}), repeating the process for each building in the dataset. By applying this "jackknife prediction" scheme, we avoid biased scores, since the building to be graded is not used to train the model. Unlike other sampling techniques (e.g. random or latin hypercube sampling), jackknife sampling systematically leaves one observation out and guarantees that each building grade will result from a different model.

We quantify relative building energy performance as the ratio of the building's reported EUI (EUI_{actual}) to EUI_{pred}:

$$Energy\ Performance\ Ratio = EUI_{actual}/EUI_{pred} \tag{5}$$

Values lower than 1 indicate that a building is consuming less energy than the model predicts, thus suggesting better performance. On the other hand, values greater than 1 indicate that a building is consuming more energy than expected compared to its peers. Although interpretable, an energy performance ratio is not as intuitive for end-user engagement as more familiar grading schemes, such as letter-grade scoring [50].

With this in mind, we use a *K-means* clustering algorithm to assign grades to buildings based on their energy performance ratios. K-means is a partitioning unsupervised learning algorithm that aims to split the data into K groups, by minimizing the variance within the clusters and maximizing the variance among different clusters [51]. We elect to use clustering over equal frequency rating procedures since it is most suitable for this particular problem. Equal frequency rating would assign grades based on each building's frequency distribution and by considering an equal number of buildings in each class. We argue that building energy performance follows a normal-like distribution, with the majority of buildings demonstrating average performance and fewer buildings showing extremely high or low energy consumption. For a more detailed comparison between unsupervised learning and equal frequency rating procedures for energy classification, please see [52].

2.7. Implementation

The implementation of the methodology is in Python 2.7, using the following packages: pandas, numpy (*data pre-processing*), XGBoost (*XGBoost algorithm*), scikit-learn (*model selection, K-means clustering*), shap (*model interpretation*). We use Python's matplotlib, seaborn, and plotly libraries for visualizations.

3. Results

In this section we summarize the key findings from applying the developed GREEN grading system to the residential building stock of NYC.

3.1. Modeling energy performance

In Fig. 2, we show the goodness of fit for the city-specific GREEN model, based on XGBoost algorithm, and contrast it with the EnergyStar score methodology. It is apparent that the EnergyStar model, based on a linear algorithm and trained on a relatively small, national sample, is

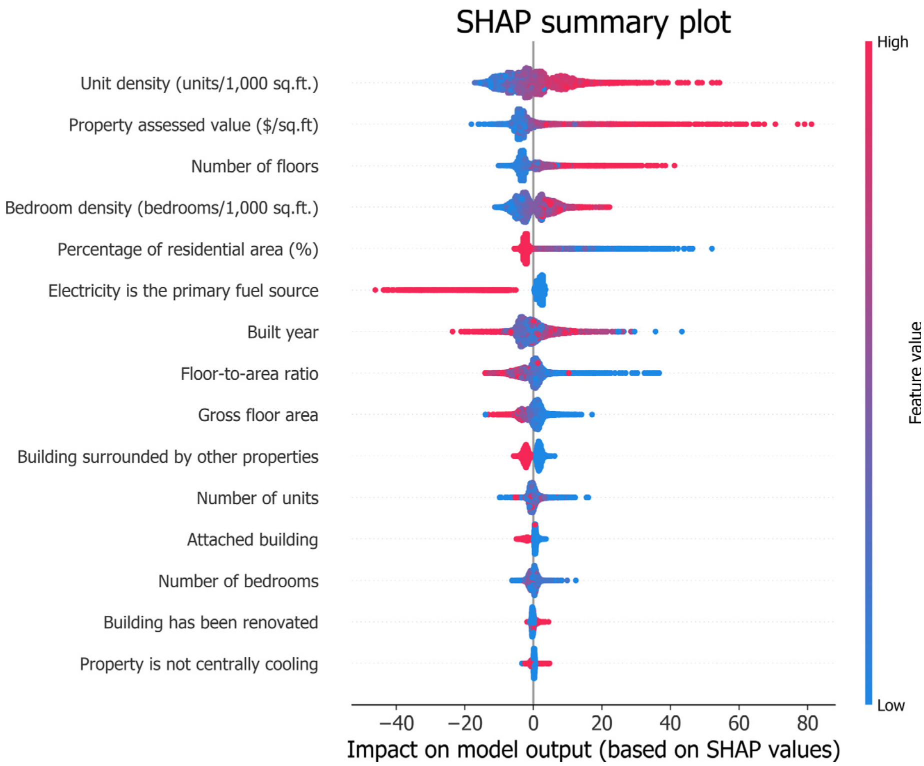

Fig. 3. Feature attributions.

not sufficient to explain *any* of the variability in the city-specific energy data, hence the negative R^2 value. On the other hand, the XGBoost algorithm is able to explain 31% of the variance in the data. As a robustness check, in Fig. D.8 we show a scatter plot of the residuals against the observations and the residual distribution.

Interpretability should be an important aspect of benchmarking models, in addition to model accuracy. Understanding which factors drive energy consumption is essential to remove ambiguity in the resultant benchmarks (Appendix C). Also, it can help cities refine future data collection processes and, eventually, improve the quality of current benchmarking methods. Fig. 3 illustrates the contribution of each feature in the model obtained from the SHAP, as described above. The attributions are sorted based on their global impact $\sum_{x=1}^{N} |\phi_i^x|$ on the model output (y-axis) and color-coded according to the feature value. The SHAP values (x-axis) are related to the feature's impact on the model's output. The vertical stacking corresponds to the feature value frequency in the dataset. Unit density is the strongest EUI predictor, with higher densities yielding higher EUI values, as expected. Similarly, properties with higher assessed value and higher number of floors tend to consume more energy per square foot. Electricity being the building's primary fuel source, although not commonly encountered (see density of red dots), is associated with lower EUI predictions. Although it is not the most important feature, in certain buildings its effect on EUI can be stronger than unit density, given the longer tail in the summary plot. The pattern of positive skewness is observed in several features (e.g. unit density, building estimated value, number of floors, floor-area ratio), reflecting the observation that extremes feature values tend to be associated with high EUI values.

3.2. *Energy performance grade assignment*

We calculate the energy performance ratio for each building, as described in Section 2.6, and cluster the ratios to form four energy performance categories, similar to the forthcoming law passed by NYC. The cluster assignment shows a distinct definition of groups with different energy performance levels (Fig. 4). Cluster A consists of buildings with energy performance ratios significantly lower than 1. Cluster B buildings perform close to the model's prediction, whereas Cluster C buildings perform 25% worse, on average, than the expected performance based on the model output. Cluster D performs on average approximately 85% worse, and is comprised of the poorest performing properties with EUI values up to 3.5 times higher than their peers. The majority of buildings are assigned to Cluster B, followed by buildings in Cluster C and A. Less than 500 properties obtain the D grade, representing the worse performing properties in the building stock. Given such transparent classification, stakeholders are aware of both the current status of a building's energy performance and the magnitude of improvements needed to move from a grade band to another. The letter-grade classification also provides a clear differentiation between the median EUI levels of each performance grade, unlike the [0–100] EnergyStar score scale (Fig. D.9).

Fig. 5 is a Sankey diagram mapping the interaction between the GREEN grading and the EnergyStar-based scoring for NYC's large residential properties. There are significant differences between the two, with 42% of the properties receiving different grades between the two grading schemes, reinforcing the limitations of existing methods. In particular, buildings with mid-tier performance (i.e. "20–49" and "50–89") are split between "A", "B", and "C" grades. We also note a few extreme cases where buildings with low EnergyStar scores receive an "A" grade, and where high EnergyStar scoring buildings received low GREEN grades (See Appendix C for a detailed comparison between individual buildings.)

4. Discussion and policy implications

Although energy disclosure is being widely adopted by cities across the US and the world, the use of these data for market transformation and data-driven policy is in its nascent stage. As Allcott and Mullainathan [53] argue, the behavioral component is as crucial as the technological in energy efficiency adoption and, ultimately, to achieving city-wide carbon reduction goals. Research has shown that peer-pressure [54] and public energy performance information

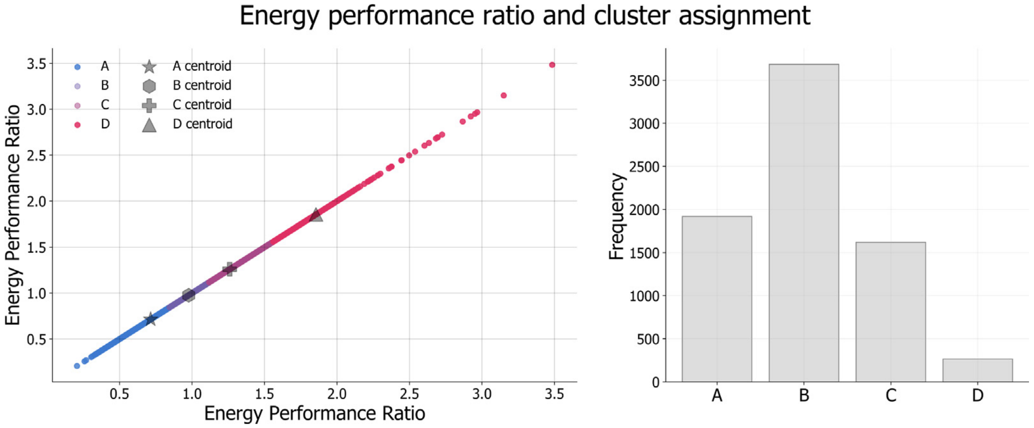

Fig. 4. Energy performance ratios, color-coded based on cluster assignment, and cluster centroids (left). Distribution of buildings assigned in each cluster (right).

disclosure [55] can motivate energy efficiency actions and investments. However, information transparency alone is not sufficient to have large-scale impacts on energy use behavior, nor on the integration of energy performance into property valuation and locational decisions [9].

In NYC, performance measurement has extended across several agencies and industries, and a similar grading concept has been applied to the sanitary conditions of the City's restaurants. Recent research found that the public disclosure of a restaurant's grade resulted in improved sanitary conditions [56]. In the building sector, the expectation is that publicly-available building energy grades will have a similar effect, encouraging competition among building owners and allowing tenants to factor energy efficiency into their leasing decisions. The implications of these market shifts would be real estate pricing and asset valuations that account for energy performance and other energy-related risks, such as obsolescence or regulatory exposure [57,7]. In addition to changes in market behavior, building grading provides the basis for city energy policies that are performance-based and data-driven, utilizing prescriptive targets for energy performance that allow the market to find the optimal solutions to achieving required goals.

Our results reinforce the need to reconsider how energy benchmarking data and information transparency can be used to motivate city-wide energy and carbon emissions reductions. First, the demand profiles and energy behavior of regional building stocks differ significantly from national samples. New York City and San Francisco, for instance, are high-density urban environments characterized by older buildings with unique architectural, zoning, and structural attributes that make them distinct from small- and medium-sized cities developed later in the 21st century [20]. Since drivers of energy performance depend on the city's particular characteristics and environment, our model factors in a comprehensive, and localized, set of physical, operational, and qualitative features to explain differences in EUI. In terms of model complexity, we argue that linear techniques are not suitable for statistical energy benchmarking. We demonstrate that a non-linear algorithm, such as XGBoost, with the appropriate features, is able explain more than 30% of the variability in the benchmarking data, whereas *none* of the variability can be explained by current state-of-the-art linear techniques. Since energy benchmarking is based on deviations between expected energy intensity (predictive model outputs) and actual energy intensity (as reported), non-linear methods can be used to increase the reliability of resultant building energy

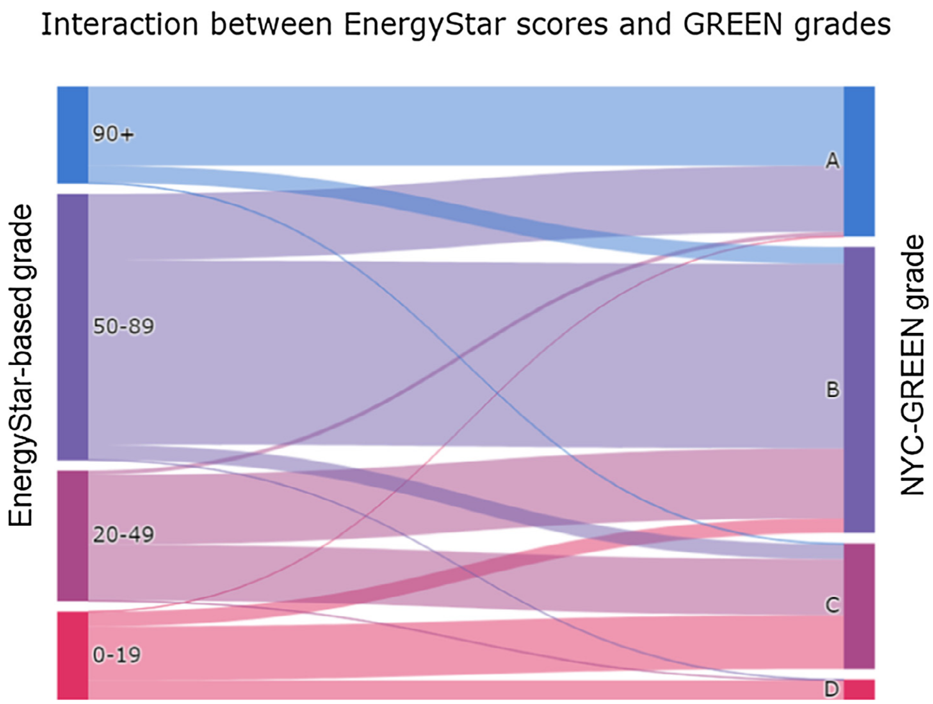

Fig. 5. Sankey diagram with flows from the EnergyStar-based grades (left node) to the proposed NYC-specific GREEN grading (right node).

performance grades by capturing complex interactions in the independent variables. Finally, there is a need for easily interpretable model outputs that can be adopted by non-expert end users. Our energy performance grade assignment using K-Means clustering yields a 4-grade energy performance classification. Unlike the $(0, 100)$ range that is the current EnergyStar scoring standard, we propose a binned classification to highlight the differences between energy performance grades, account for uncertainties in the underlying model, and inform behavior change in the marketplace. Such transparency can help building owners understand their energy performance against their peers and motivate the adoption of energy conservation measures.

Although we use NYC residential building stock as a case study for this particular research, the GREEN grading system's dynamic nature allows for application to different building typologies (e.g. commercial, retail, etc.) and any city with enacted energy disclosure laws (e.g. Austin, Boston, Chicago, San Francisco, and Seattle to name a few).

5. Conclusion

Cities across the US and globally are turning to energy disclosure as a means to better understand their building stocks' energy performance, and use that information to develop more effective, data-driven policies. In this work, we propose a novel methodology to develop a city-specific energy performance grading system for New York City's multifamily residential building stock 7500. Specifically, we use XGBoost, a variant of gradient tree boosting, to model building EUI using an optimally-selected set of physical, operational, and qualitative features. For each building in the data set, we calculate its energy performance ratio by comparing actual EUI to model-predicted EUI. Using a clustering algorithm, we then partition these ratios into a 4-grade energy performance classification scheme. We contrast the proposed method with the EnergyStar scoring model, which has gained widespread market adoption in the building industry. We show that EnergyStar has limited predictive power when using NYC building energy data, and is not a suitable methodology for reliably comparing building energy performance.

Our GREEN building energy grading method is driven by the identified principles that it be understandable and reproducible, robust amd reliable, and scalable and generalizable. We accomplish this in several ways. First, our approach is city-specific, able to identify peer buildings and establish objective comparisons between them. Second, we employ a non-linear data modeling algorithm to capture the complex relationships between the variables that influence energy performance, and select the features that best explain variations in EUI. Finally, our methodology is dynamic, so that it can be updated with the most recent data streams or applied to cities with heterogeneous characteristics. Our methodology provides the foundation for continued research on contextual city-specific energy performance metrics, and the appropriate standards for building energy grading. In future work, we intend to expand our methodology using data from other US cities, identify the drivers of energy performance, and quantify their relationship with characteristics such as urban morphology, existing regulations, and demographics, among others.

Acknowledgments

The authors would like to thank the NYC Mayor's Office of Sustainability for access to relevant datasets, and for comments on early versions of the methodology. We would also like to thank two anonymous reviewers and the editors of Applied Energy for their constructive feedback. This material is based on work supported, in part, by the National Science Foundation under grant No. 1653772.

Appendix A. The NYC law on disclosure of energy efficiency scores and grades

In late 2017, NYC enacted a law requiring large building owners to post their energy performance grades near building entrances [58]. Building on the existing Local Law 84, the new legislation aims to further raise the awareness of building energy performance among tenants, investors, and the public and increase competition among owners to make their buildings more efficient [59]. Under the new law, buildings will be graded based on their EnergyStar scores as follows:

- A: 90 or above
- B: 50–89
- C: 20–49
- D: 0–19
- F: Buildings that do not submit benchmarking information.
- N: Buildings exempted from benchmarking.

Nevertheless, the new law has already received heavy criticism, mainly due to the exclusion of several important drivers of energy consumption in the Energy Star scoring calculation, and to significiant financial implications of receiving a low grade based on unreliablem, and potentially flawed, models [60–62].

Appendix B. EnergyStar grading method

The EnergyStar grading method for multifamily buildings consists of a linear regression model, trained on 322 sample buildings across the U.S. The model specification is as follows:

$$\widehat{EUI} = 140.8 + 52.57{*}cUnitDensity + 24.45{*}cBedroomPerUnit - 18.76{*}LowRise + 0.009617{*}cHDD + 0.01616{*}cCDD \tag{B.1}$$

where $\widehat{EUI}$ is the predicted energy use intensity, *UnitDensity* is the number of units per 1,000 ft^2, *BedroomPerUnit* is the number of bedrooms per unit, *LowRise* is a dummy variable being 1 if the building is lower that five floors tall and 0 otherwise, *HDD* and *CDD* are the annual heating and cooling degree days respectively. Prefix *c* denotes that the values are centered on the sample's mean value.

Based on the model's output the energy efficiency ratio $\frac{actualEUI}{predictedEUI}$ is defined and the 0–100 EnergyStar score is calculated based on energy efficiency ratios' distribution. For further details on the grading method we refer interested readers in EnergyStar's Technical Ref. [23].

Appendix C. Examples of individual building comparisons

Here we present two examples of buildings that received widely divergent scores between our proposed grading scheme and those resulting from the EnergyStar model. To maintain the anonymity of the individual buildings, we refer to them using their respective index identifier in the dataset.

C.1. High GREEN grade, low EnergyStar score

Building #4826 receives an 'A' GREEN grade, but an EnergyStar score of 15. The building's reported EUI is 152 kBtu/ft^2, whereas the median EUI for the entire dataset is 124.2 kBtu/ft^2 (Table D.2). Although this may suggest the low EnergyStar score is justified, a more comprehensive analysis of the building's characteristics demonstrate that its unit density, assessed value, and height are significantly higher than the sample's median (i.e. unit density 1.19, assessed value of $128.60 per square foot, number of floors 32). Sub-setting the sample with buildings higher than 25 floors, with unit density greater than 1.1, and assessed value above $100 per square foot, we observe a median EUI in this sample of 176.4 kBtu/ft^2 that reinforces the assigned 'A' grade, while highlighting the inadequacy of EnergyStar score in establishing fair peer-to-peer comparisons based on a the full range of attributes that impact energy use.

C.2. High EnergyStar score, low GREEN grade

Building #3701 is assigned a 'D' GREEN grade and receives an EnergyStar score of 74. The building's EUI is 126 kBtu/ft^2, which is close to the sample's median. The property's unit density, assessed value, and number of floors are close to the sample's median values as well, however in this particular case electricity is the building's primary fuel source. Similar properties report a median EUI of 62.2 kBtu/ft^2, hence the building's grade of 'D'. The example highlights our framework's ability to factor in features beyond physical and occupancy characteristics to establish more contextualized benchmarks.

Appendix D. Supplementary material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.apenergy.2018.10.053.

References

[1] IPCC. Climate change 2014: mitigation of climate change vol. 3. Cambridge University Press; 2015.

[2] UNEP. Buildings and climate change: summary for decision-makers. Paris: United Nations Environmental Programme, Sustainable Buildings and Climate Initiative; 2009. p. 1–62.

[3] Ribeiro D. Developments in local energy efficiency policy: a review of recent progress and research. Curr Sustain/Renew Energy Rep 2018;5:109–15.

[4] Zhao H, Magoulès F. A review on the prediction of building energy consumption. Renew Sustain Energy Rev 2012;16:3586–92.

[5] Pérez-Lombard L, Ortiz J, Pout C. A review on buildings energy consumption information. Energy Build 2008;40:394–8.

[6] Palmer K, Walls M. Using information to close the energy efficiency gap: a review of benchmarking and disclosure ordinances. Energy Efficiency 2017;10:673–91.

[7] Kontokosta CE. Energy disclosure, market behavior, and the building data ecosystem. Ann New York Acad Sci 2013;1295:34–43.

[8] Meng T, Hsu D, Han A. Estimating energy savings from benchmarking policies in New York city. Energy 2017;133:415–23.

[9] Papadopoulos S, Bonczak B, Kontokosta CE. Pattern recognition in building energy performance over time using energy benchmarking data. Appl Energy 2018;221:576–86.

[10] Kontokosta CE. A market-specific methodology for a commercial building energy performance index. J Real Estate Finan Econ 2015;51:288–316.

[11] Scofield JH. ENERGY STAR building benchmarking scores: good idea, bad science. *Oberlin College study for American Council for an Energy Efficient Economy (ACEEE)*; 2014.

[12] Hsu D. Improving energy benchmarking with self-reported data. Build Res Inform 2014;42:641–56.

[13] Gao X, Malkawi A. A new methodology for building energy performance benchmarking: an approach based on intelligent clustering algorithm. Energy Build 2014;84:607–16.

[14] Xuchao W, Priyadarsini R, Eang LS. Benchmarking energy use and greenhouse gas emissions in Singapores hotel industry. Energy Policy 2010;38:4520–7.

[15] Chung W. Using the fuzzy linear regression method to benchmark the energy efficiency of commercial buildings. Appl Energy 2012;95:45–9.

[16] Olofsson T, Meier A, Lamberts R. Rating the energy performance of buildings. Int J Low Energy Sustain Build 2004;3.

[17] De Wilde P. The gap between predicted and measured energy performance of buildings: a framework for investigation. Automat Construct 2014;41:40–9.

[18] van Dronkelaar C, Dowson M, Burman E, Spataru C, Mumovic D. A review of the energy performance gap and its underlying causes in non-domestic buildings. Front Mech Eng 2016;1:17.

[19] Borgstein E, Lamberts R, Hensen J. Evaluating energy performance in non-domestic buildings: a review. Energy Build 2016;128:734–55.

[20] Papadopoulos S, Bonczak B, Kontokosta CE. Spatial and geographic patterns of building energy performance: a cross-city comparative analysis of large-scale data. In: International conference on sustainable infrastructure; 2017. p. 336–48.

[21] Fan C, Xiao F, Zhao Y. A short-term building cooling load prediction method using deep learning algorithms. Appl Energy 2017;195:222–33.

[22] Wei Y, Zhang X, Shi Y, Xia L, Pan S, Wu J, et al. A review of data-driven approaches for prediction and classification of building energy consumption. Renew Sustain Energy Rev 2018;82:1027–47.

[23] EPA. Energy star®performance ratings technical methodology. Environmental Protection Agency; 2011.

[24] Kontokosta CE, Tull C. A data-driven predictive model of city-scale energy use in buildings. Appl Energy 2017;197:303–17.

[25] Khayatian F, Sarto L, et al. Building energy retrofit index for policy making and decision support at regional and national scales. Appl Energy 2017;206:1062–75.

[26] Papadopoulos S, Azar E, Woon W-L, Kontokosta CE. Evaluation of tree-based ensemble learning algorithms for building energy performance estimation. J Build Perform Simul 2017:1–11.

[27] Melo A, Cóstola D, Lamberts R, Hensen J. Development of surrogate models using artificial neural network for building shell energy labelling. Energy Policy 2014;69:457–66.

[28] Yalcintas M. An energy benchmarking model based on artificial neural network method with a case example for tropical climates. Int J Energy Res 2006;30:1158–74.

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

[30] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Advances in neural information processing systems. p. 4768–77.

[31] Lundberg SM, Erion GG, Lee S-I. Consistent individualized feature attribution for tree ensembles; 2018. Available from: arXiv preprint arXiv:1802.03888.

[32] Yang Z, Roth J, Jain RK. Due-b: data-driven urban energy benchmarking of buildings using recursive partitioning and stochastic frontier analysis. Energy Build 2017.

[33] Fuerst F, McAllister P, Nanda A, Wyatt P. Does energy efficiency matter to homebuyers? An investigation of epc ratings and transaction prices in England. Energy Econ 2015;48:145–56.

[34] Fuerst F, McAllister P, Nanda A, Wyatt P. Energy performance ratings and house prices in wales: an empirical study. Energy Policy 2016;92:20–33.

[35] Brounen D, Kok N, Quigley JM. Residential energy use and conservation: economics and demographics. Eur Econ Rev 2012;56:931–45.

[36] Wyatt P. A dwelling-level investigation into the physical and socio-economic drivers of domestic energy consumption in England. Energy Policy 2013;60:540–9.

[37] Kavousian A, Rajagopal R, Fischer M. Determinants of residential electricity consumption: using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. Energy 2013;55:184–94.

[38] Mocanu E, Nguyen PH, Gibescu M, Kling WL. Deep learning for estimating building energy consumption. Sustain Energy, Grids Networks 2016;6:91–9.

[39] Li C, Ding Z, Zhao D, Yi J, Zhang G. Building energy consumption prediction: an extreme deep learning approach. Energies 2017;10:1525.

[40] Chen T, Guestrin C. Xgboost: a scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM; 2016. p. 785–94.

[41] Kontokosta C, Bonczak B, Duer-Balkind M. Dataiqa machine learning approach to anomaly detection for energy performance data quality and reliability. In: Proceedings of the ACEEE.

[42] Hsu D. Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data. Appl Energy 2015;160:153–63.

[43] Hsu D. Identifying key variables and interactions in statistical models of building energy consumption using regularization. Energy 2015;83:144–55.

[44] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann

Stat 2001:1189–232.

[45] Zhang C, Ma Y. Ensemble machine learning: methods and applications. Springer; 2012.

[46] Friedman J, Hastie T, Tibshirani R. The elements of statistical learning vol. 1. Springer Series in Statistics New York; 2001.

[47] Scikit-learn. Undefitting vs. Overfitting; 2018. < http://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html > [accessed 2018-1-26].

[48] Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowl Inform Syst 2014;41:647–65.

[49] Ribeiro MT, Singh S, Guestrin C. Why should i trust you?: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM; 2016. p. 1135–44.

[50] Harbaugh R, Rasmusen E. Coarse grades: informing the public by withholding information. Am Econ J: Microecon 2018;10:210–35.

[51] Jain AK. Data clustering: 50 years beyond k-means. Pattern Recogn Lett 2010;31:651–66.

[52] Santamouris M, Mihalakakou G, Patargias P, Gaitani N, Sfakianaki K, Papaglastra M, et al. Using intelligent clustering techniques to classify the energy performance of school buildings. Energy Build 2007;39:45–51.

[53] Allcott H, Mullainathan S. Behavior and energy policy. Science 2010;327:1204–5.

[54] Allcott H. Social norms and energy conservation. J Public Econ 2011;95:1082–95.

[55] Delmas MA, Lessem N. Saving power to conserve your reputation? The effectiveness of private versus public information. J Environ Econ Manage 2014;67:353–70.

[56] Wong MR, McKelvey W, Ito K, Schiff C, Jacobson JB, Kass D. Impact of a letter-grade program on restaurant sanitary conditions and diner behavior in New York city. Am J Public Health 2015;105:e81–7.

[57] Walls M, Gerarden T, Palmer K, Bak XF. Is energy efficiency capitalized into home prices? Evidence from three us cities. J Environ Econ Manage 2017;82:104–24.

[58] The New York City Council. Int 1632-2017: A Local Law to amend the administrative code of the city of New York, in relation to energy efficiency scores and grades for certain buildings; 2017. < http://legistar.council.nyc.gov/LegislationDetail.aspx?ID = 3066694&GUID = A4E3E696-2927-4A44-BD39-4C2DCC8CAADD > [accessed 2018-5-26].

[59] Institute for Market Transformation. Coming to NYC: Building Energy Grades; 2018. < https://www.imt.org/coming-to-nyc-building-energy-grades/ > [accessed 2018-5-26].

[60] New York Post. Managing property in NYC is about to get a lot harder; 2018a. < https://nypost.com/2018/01/02/managing-property-in-nyc-is-about-to-get-a-lot-harder/ [accessed 2018-5-26].

[61] New York Post. New Yorks dumb way to go green; 2018b. < https://nypost.com/2018/01/08/new-yorks-dumb-way-to-go-green/ > [accessed 2018-5-26].

[62] Crain's New York Business, A failing grade for the city's new energy-efficiency scoring system; 2018. < http://www.crainsnewyork.com/article/20180116/OPINION/180119939/editorial-a-failing-grade-for-the-citys-new-energy-efficiency-scoring-system > [accessed 2018-5-26].