VRIJE UNIVERSITEIT AMSTERDAM

Master Thesis

Big Data to Small Footprints: Predicting Office Operating Carbon Emissions Using Machine Learning

by

Stan Brouwer

(2671939)

First supervisor:	Ronald Siebes
Daily supervisor:	Ronald Siebes
Second reader:	Second reader's name

July 15, 2025

Submitted in partial fulfillment of the requirements for the VU degree of Master of Science in Information Sciences

DECLARATION OF AUTHORSHIP

I, **Stan Brouwer**, declare that this thesis titled "Big Data to Small Footprint: Predicting Office Operating Carbon Emissions Using Machine Learning" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Stan Brouwer

Date: 11 July 2025

Big Data to Small Footprints: Predicting Office Operating Carbon Emissions Using Machine Learning

Stan Brouwer
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
s.j2.brouwer@student.vu.nl

ABSTRACT

Context. Building energy consumption is influenced by many complex factors, including building characteristics, usage patterns, and climate. Accurate prediction of energy use intensity (EUI) is essential to improve energy efficiency and support sustainable building management.

Goal. This study aims to evaluate and compare the performance of various statistical and machine learning models in predicting EUI for office buildings, highlighting the key predictors and the effectiveness of nonlinear approaches.

Method. We analyze a large data set of office buildings, applying linear regression, decision trees, and random forest models. We evaluated the accuracy of the model using multiple performance metrics and examined the influence of predictors such as operating hours, heating and cooling degree days, and floor area.

Results. Nonlinear models, especially random forest regression, significantly outperformed linear regression, achieving higher predictive accuracy ($R^2 = 0.65$) and lower error rates. Operating hours and climate variables were identified as the most influential predictors. Linear models showed limited explanatory power due to complex, non-linear relationships in the data.

Conclusions. Machine learning methods like random forests better capture the complexity of building energy use and improve prediction reliability. Future research should incorporate additional contextual factors and explore advanced modeling techniques to support more adaptive and precise energy management strategies.

1 INTRODUCTION

Greenhouse gas (GHG) emissions associated with human activities have already caused 1.1 [0.95-1.20] ° C of global warming above preindustrial levels [10], leading to irreversible changes in the climate system. In response, the Paris Agreement [19] has established goals to limit global warming to well below 2 ° C and preferably 1.5 ° C. However, current policy outcomes fall short: Under existing commitments, the world is on track to 2.8 ° C of warming by the end of the century [58, 72]. To avoid this outcome, great reductions in emissions are required. The UNEP notes that achieving the 1.5 ° C target requires a decrease in current emissions by 42% by 2030 and 57% by 2035 [58].

Corporations are under growing pressure to align their operations with climate goals. Governments are introducing stricter regulations, such as the EU Corporate Sustainability Reporting Directive [13], which requires larger companies to disclose climate-related risks, greenhouse gas emissions and sustainability. At the same time, social expectations are shifting. Public awareness and concern about climate change create reputational incentives for companies to demonstrate environmental responsibility [30, 80]. Investors also are increasingly valuing environmental, social, and governance (ESG) factors in their decision-making, favoring organizations with sustainability strategies [13, 29, 36, 67]. In addition, legal accountability is on the rise. Landmark rulings such as Milieudefensie v. Royal Dutch Shell underscores the growing role of the judiciary in enforcing corporate climate responsibility, further compelling companies to act.

The role of office buildings. Office buildings present a significant opportunity to reduce greenhouse gas emissions through improved energy management [58]. Buildings account for approximately 40% of total energy consumption and GHG emissions related to energy in both the US and the EU [28, 73]. Although the global proportion is lower (between 20% and 40%), it is expected to grow rapidly with the urbanization and development of new economies [62]. A large portion of this energy use is still derived from fossil-fuel heating, cooling, or electricity systems, making the sector critical for decarbonization.

The energy use in building operations alone is estimated to account for 19% of global GHG emissions. [70] Commercial buildings, including offices, are particularly energy intensive, often consuming up to three times more energy per square meter than residential buildings [9]. Despite increasing awareness, energy consumption in office buildings continues to increase, posing challenges to climate goals. This highlights the need for better evaluation and management of energy performance within the sector [58].

Managing and reducing energy consumption starts with quantifying energy use, which is typically normalized for floor area and period and expressed as energy use intensity (EUI; $\mathrm{kWh \cdot m^{-2} \cdot year^{-1}}$ [37, 50]. Generally, there are two methodological approaches to assess energy performance: physics-based simulations that use detailed structural and environmental data in combination with energy simulation tools (EnergyPlus, TRNSYS, ESP-r) to calculate consumption; or data-driven and statistical models that infer patterns from historical building and environmental data to predict energy use [65].

Although simulation-based models are widely used in energy optimization and certification schemes, they are computationally intensive and often impractical for large-scale or early-stage evaluations. Optimization algorithms relying on simulation tools suffer from high time overhead, making them costly to scale or apply across portfolios [6, 68]. In contrast, data-driven models offer a more scalable alternative, reducing costs and enabling rapid evaluation of energy performance across large datasets [22, 23]. They are also increasingly used to support design decisions, optimize HVAC systems, and benchmark energy performance [1, 46]. Importantly,

classification-based techniques can facilitate the analysis of complex variables such as occupant behavior or usage patterns, which are difficult to model by simulation [79].

Practical barriers. Although technological solutions such as smart meters, advanced HVAC systems, and retrofitting exist and are often cost-effective, their implementation remains limited, and thus the challenges are informational and economic rather than technical [81]. Reliable performance data are often missing or not transparent, preventing markets from operating efficiently [38]. This hinders investors and tenants from evaluating energy efficiency, reducing incentives to invest in sustainable buildings.

Transparency initiatives such as the EU Building Energy Performance Directive (EPBD), ENERGY STAR, and LEED have improved the availability of information. Studies show that certified buildings command rental and sales premiums [24, 32, 49]. However, these certifications are inconsistent between regions, not always publicly accessible, and often rely on complex, expert-driven simulations.

Practical constraints further hinder sustainability initiatives. Office buildings are often multi-tenant and rely on shared energy meters. Tenants are often charged based on floor area rather than actual usage. This structure limits the feedback tenants receive on their actual energy use and undermines incentives to reduce consumption [43]. At the portfolio level, building owners and real estate managers struggle to collect and compare energy performance data, making it difficult to identify underperforming assets, comply with ESG mandates, or justify sustainability investments. Providing performance data has wider economic benefits, reduces adverse selection, aligns stakeholder incentives, and allows more targeted, cost-effective policies [3, 35].

Motivation. With increased attention to decarbonization, there is a growing body of research on energy consumption in buildings. Much of it focuses on the residential sector, while office buildings remain relatively understudied. As office buildings have different characteristics that differentiate them from residential structures, they require separate analysis.

This research contributes to the scientific literature by applying ML/statistical techniques to large data sets of more than 7,000 US office buildings. We evaluated different machine learning and statistical methods that estimate the energy consumption of US office buildings. Novel is the combination with weather data based on the location and size of the considered data set.

The study is structured around the following research questions: *RQ*1: What are the key characteristics (e.g., building size, age, number of occupants, operating hours) most strongly determining energy use in US office buildings? *RQ*2: How do different modeling techniques compare in accuracy when used to predict yearly energy consumption?

By developing and evaluating the precision of predictive models, this study aims to offer a faster and more scalable approach to support effective energy management, targeted retrofits, and strategic decision-making in the commercial real estate sector.

2 BACKGROUND

The energy consumption of buildings, and in particular office buildings, contribute significantly to overall energy use and CO_2 emissions [70, 75]. Consequently, developing accurate models for the prediction of energy consumption has become crucial to optimize energy consumption, reduce costs, and promote sustainability in the office environment [48]. Scholars have recognized the importance of energy prediction in achieving energy efficiency and cost savings since the mid-1980s [12].

Drivers of office building energy usage. Energy consumption in buildings is influenced by many physical, operational, behavioral, and environmental variables. Physicists and engineers tend to create physical models that calculate heat dissipation and radiation, while the data-driven domains use big data approaches.

From an architectural perspective, studies have emphasized the importance of building envelope properties such as insulation, wall-to-window ratio, air tightness, HVAC system efficiency, and the impact of passive design features (orientation, shading, thermal mass) on energy demand [62, 71].

In behavioral sciences, researchers have identified occupant behavior and organizational culture as significant drivers of variations in energy use. Research indicates that user habits such as thermostat settings, window opening, and equipment usage can lead to large differences in energy performance, even among technically equivalent buildings. This "performance gap" is often attributed to behavioral unpredictability[34, 63]).

Empirical studies have attempted to quantify the relative impact of different energy consumption factors, often using regression models or machine learning. Most of the variables investigated are discussed below.

Floor area. Almost all analyses find that larger offices use more energy, roughly in proportion. A nationwide Korean study reports that floor area alone explains approximately 90% of the variation in office energy usage (R^2 = [0.89 - 0.91]) [40]. Often, regression models include the floor area. Based on a New York City office data set, Kontokosta reports that each additional m^2 of floor area raised the yearly energy use intensity by 0.20 kWh/m^2 ($p < 0.05$) [42]. Although energy usage scales roughly linearly with floor area, it is hypothesized that larger office buildings benefit from economies of scale and have lower energy consumption per floor area (energy intensity). Kim & Kim [40] found that the coefficient of determination between total floor area and energy consumption was higher with a quadratic model for large office buildings than with a linear model, indicating possible scale benefits. However, the authors also note that the linear regression model performed better for smaller offices and suggest the application of segmented regression. These interactions could be further complicated by the interactions between building age, energy rating, and building age and size as newer buildings tend to be better insulated and larger. In absolute terms, the pattern is clear: Larger buildings generally use more energy.

Number of occupants. Studies consistently find that the intensity of energy increases with the number of people inside a building. By including workers per area as a positive factor, Kontokosta found

that adding one worker per m^2 increased annual energy usage with about 3.07 KWh/m^2 (p > 0.01) [42].

Operating hours. The longer an office is used per week, the more energy it consumes. Regression studies often include operating hours as a predictor. For example, Sharp (1996) identified operating hours as a strong driver of energy usage. Sharp found that each additional open hour per week corresponded to an increase in energy use of 1.41 kWh/m^2 (p < 0.05), approximating the reported 1.94 kWh/m^2 (p <0.0001) from the ENERGY START RATING TEAM (2019) technical report [26, 66]

Building age. Many regressions find that newer offices use more energy per area. Kontokosta found a negative correlation of energy use with age: offices older than 80 years used approximately 30% less energy per area than the average office [42]. This trend is also observed in a study in the UK, reporting that the intensity of electricity is higher in recently built offices [**?**] The Building Energy Research Center of Tsinghua University (2023) attributes the observed trend of increased intensity of the energy of office buildings mainly to the increasing prevalence of air conditioning systems [9]

Weather influences. Weather significantly affects the use of energy in commercial buildings, which must be taken into account. The simplest approach is to consider degree days: summing the differences between outdoor temperature and a base temperature, loosely corresponding to the temperature gradient between indoor and outdoor temperatures [47]. More refined methods fit regression models of energy and weather variables. With the rise of machine learning, more advanced methods such as neural networks, random forests, etc. can model non-linear responses to temperature and other factors. Each approach to considering the influence of weather uses common meteorological features, such as temperature, humidity, solar radiation, and wind speed, to estimate what the energy use would have been under a reference climate. By removing weather effects, these models enable better year-over-year or between-building comparisons [44]

The most common normalization method for adjusting for the effects of weather is the heating and cooling degree-day (HDD, CDD) method [2], although the methodology for normalizing for other variables is similar. Degree days represent the total positive or negative differences between a set temperature and the average temperature for a given period of time [7], which has been specified as 18.3 ° C (65 ° F) in the US. Kissock et al. discussed the Variable Base Degree-Day (VBDD) method, in which the most optimal base temperature that provides the best statistical fit is calculated. In the basic degree-day method, a regression model correlates energy use with degree days, which is then evaluated on its R^2 value. For VBDD, multiple models are developed with different base temperatures, and the model with the highest R^2 value is selected [41].

3 METHODS

This study develops predictive models for the annual EUI of office buildings using building characteristics and climate data. The methodology is structured into four main sections: are organized into four main sections: Data Collection & Preprocessing, Feature Selection, Model Development, and Model Evaluation. An a priori significance level of $\alpha \leq 0.05$ was applied throughout.

3.1 Data collection & preprocessing

Data were compiled from 26 publicly available datasets that contain annual energy consumption and associated building characteristics [4, 5, 14, 15, 17, 18, 20, 25, 31, 33, 45, 52–57, 59–61, 64, 69, 76**?** –78].

Only records explicitly labeled as office buildings were retained. Further filtering ensured inclusion of entries with positive and non-missing values for the following variables: site energy use (defined as the total annual site energy consumption in MWh), floor area, year built, energy source indicators and location (city or state).

All imperial units were converted to SI units except for time and energy use (reported in years and MWh for interpretability). The energy use intensity was calculated as the energy use at the site divided by the area of the floor.

The heating degree days and the cooling degree days were sourced from the National Weather Service Climate Prediction Center, which provides aggregated monthly degree day statistics for 359 major US cities [51]. The building locations were matched with city-level climate data; If city data were unavailable, the weighted average of the state population was applied. The final data set includes 32,686 yearly observations in 7,226 unique office buildings, spanning 2010 to 2023.

3.2 Feature selection

The predictor variables were initially assessed for compliance with the assumptions of normality, homoskedasticity, and independence. Distributions were visually inspected using histograms and Q-Q plots. The z-values of the skewness and kurtosis were calculated, with ±2 as the threshold for normality. Shapiro-Wilk's test was performed on random samples (n=2,000 per variable) to mitigate sensitivity to large sample sizes (Field, 2009). For non-normal variables, log-transformations were applied and reassessed. Heteroskedasticity was evaluated with Levene's test.

3.3 Correlation analysis

The association between candidate predictors and yearly EUI was examined using Pearson's correlation coefficient (PCC) or Spearman's rho in the case of nonnormality. Cohen's (1988) guidelines on the interpretation of correlation coefficients served as a reference, where correlation values of r = 0.1, 0.3 and 0.5 indicate small, moderate or high correlation. However, it should be repeated that the interpretation of the coefficients ultimately depends on their context and purpose, and given the relatively large sample size in our dataset (n > 30,000), even relatively minor correlation coefficients can achieve statistical significance and reflect real associations [16].

3.4 Multicollinearity assessment

An intervariable correlation matrix and Variance Inflation Factors (VIF) were calculated to detect multicollinearity. Variables with a VIF greater than 5 were considered for removal or combination to ensure model stability.

3.5 Development of predictive models

3.5.1 Multiple linear regression (OLS). A simple linear parametric model was developed to predict the energy use of buildings only with electricity. A separate multivariate linear regression model was developed to predict both the electric and fuel energy use for mixed-source buildings. The assumptions of normality, homoskedasticity and independence were evaluated by visual inspection of the histogram and QQ plots, and with Shapiro-Wilk and Levene's tests. The general form was as follows:

$$\text{EUI}_i = \beta_0 + \beta_1 \cdot \text{FloorArea}_i + \beta_2 \cdot \text{YearBuilt}_i + \beta_3 \cdot \text{OperatingHours}_i + \beta_4 \cdot \text{CDD}_i + \beta_5 \cdot \text{HDD}_i + \epsilon_i$$

With:

- EUI_i: Energy Use Intensity of building i.
- β_0: Intercept term.
- β_1: Coefficient for floor area.
- FloorArea_i: Floor area of building i.
- β_2: Coefficient for year built.
- YearBuilt_i: The year building i was constructed.
- β_3: Coefficient for operating hours.
- OperatingHours_i: Building operating hours i.
- β_4: Coefficient for cooling degree days.
- CDD_i: Cooling Degree Days at the building's location.
- β_5: Coefficient for heating degree days.
- HDD_i: Heating Degree Days at the building's location.
- ϵ_i: Error term.

3.5.2 Decision tree regression. A decision tree model was developed to capture potentially non-linear relationships between building characteristics and energy use. Unlike linear models, decision trees recursively divide the data into smaller groups based on 'decision rules' that maximize the difference in the outcome variable between branches. This allows the model to 'learn' threshold effects and interactions that might be difficult to represent parametrically. To tune the model and prevent overfitting, key hyperparameters, such as the maximum number of splits (tree depth), the minimum number of samples required to split a node, were optimized using 5-fold cross-validation. In this approach, the training data are split into five equal subsets (folds). The model is trained on four sets and validated on the remaining set, the process being repeated five times with rotation in order for each set to be used as a validation set exactly once. The average performance across the iterations provides a robust estimate of how well the model generalizes to new data.

Before hyperparameter tuning, we randomly split the data set into 80% training data and 20% testing data, which were not used for the 5-fold cross-validation. We tested each combination of hyperparameters with tree depths ranging from 2 to 20 and minimum samples per split ranging from 2 to 20. The accuracy of the model was evaluated using the mean absolute error (MAE), which measures the average absolute difference between predicted and actual annual energy values. The hyperparameter combination with the lowest average MAE across validation folds was selected. The final decision tree was then re-trained on the full training dataset using this optimal configuration and tested on the unseen holdout set.

3.5.3 Random forest regression. To improve prediction accuracy and mitigate overfitting observed in single decision trees, a random forest model was implemented [27]. This ensemble method builds multiple decision trees using bootstrapped samples and random feature selection, averaging their predictions to enhance generalization. Hyperparameter tuning followed the same 5-fold cross-validation and holdout approach. Feature importance metrics were extracted to interpret predictor influence.

3.6 Model evaluation

Linear regression models were evaluated using R^2, adjusted R^2, MSE, RMSE, and MAE on the complete dataset. Nonlinear models were assessed on the 20% holdout set to provide unbiased estimates of predictive performance.

Evaluation metrics include:

EVALUATION METRICS

Evaluation metrics include:

- **Root Mean Squared Error (RMSE)**: Square root of the average squared differences between predicted and actual values, penalizing larger errors [39].
- **Mean Absolute Error (MAE)**: Average absolute difference between predictions and observations, providing a unit-consistent error measure.
- **Mean Absolute Percentage Error (MAPE)**: Average percentage deviation between predictions and actual values, enabling scale-independent comparison [21].
- **Mean Bias Error (MBE)**: Average bias indicating systematic over- or underprediction.
- **Coefficient of Variation of RMSE (CV_RMSE)**: RMSE normalized by the mean of observed values, allowing comparability across datasets [11].
- **Coefficient of Determination (R^2)**: Proportion of variance in the dependent variable explained by the model.

Where:

- A_t = actual value
- P_t = predicted value
- n = number of observations
- $\bar{A}$ = mean of actual values

4 RESULTS

4.1 Statistics

The data set considered contained 32,686 valid yearly observations in 7,226 unique office buildings, with observation dates ranging from 2010 to 2023. The descriptive statistics are reported in Table 1.

The results of the normality tests are presented in Table 2. In combination with the visual inspection of the histograms and Q-Q plots it was assumed that none of the variables demonstrated normality.

4.2 Feature selection

The correlation analysis (Table 3) revealed a strong correlation between the number of people and floor area ($r = 0.88$), both showing VIF values greater than five (5.74 and 5.34). Due to this multicollinearity and the theoretical relevance of floor area to EUI

(e.g., surface-to-volume ratio effects), the number of people was excluded from linear models. After removal, all VIFs dropped below 5 (Table 4).

4.3 Linear regression model

The linear regression results are presented in Table 5. Although all predictor variables in the model reach statistical significance, the predictive power of linear regression is low (R^2 = 0.09). Operating hours is by far the strongest predictor. As expected from energy models, HDD and CDD significantly influence the intensity of energy use. In line with [42], newer buildings are slightly more energy intensive.

4.4 Model comparison

The Random Forest model clearly outperformed both the linear regression and the decision tree models across all metrics (Table 6). It achieved the highest coefficient of determination (R^2 = 0.65), indicating that it explains approximately 65% of the variance in the building energy use intensity, a substantial improvement over the Linear Regression (R^2 = 0.09) and Decision Tree (R^2 = 0.25). The Random Forest also had the lowest errors: its *RMSE* (0.08) and *MAE* (0.05) were significantly smaller than those of linear regression (*RMSE* = 0.02; *MAE* = 0.09) and the decision tree (*RMSE* = 0.12; *MAE* = 0.08), suggesting better predictive accuracy and precision. The MAPE of 23.63% indicates that, on average, the predictions deviated from the actual values by about 24% for the random forest model, which is considerably better than the Linear Regression (44.70%) and the Decision Tree (39.60%).

The *MBE* was near zero for all models, with the Random Forest showing a slight negative bias (-0.005), implying a small tendency to underpredict but overall negligible.

Finally, the Coefficient of Variation of RMSE (*CVRMSE*), which normalizes error by the mean of the response variable, was lowest for the Random Forest (0.31), further confirming its superior relative performance.

5 DISCUSSION

This study aimed to evaluate the performance of predictive models for building energy use intensity (EUI) by evaluating multiple statistical and machine learning approaches on a large dataset of office buildings. The predictor variables deviated from normality and thus linear regression oversimplifies the intricate relations between predictor and outcome variables. Although statistically significant among predictors, the explanatory power of the linear regression was low ($adjusted\ R^2$ = 0.09). This aligns with the known complexity of the factors that influence energy use, which linear models may not fully capture. Operating hours emerged as the most influential predictor, underscoring the importance of building utilization patterns in energy demand. Additionally, HDD and CDD significantly affected EUI, highlighting the role of local climate in energy consumption patterns.

Operating hours were the most influential predictor in the linear model, with a $1.35 KWh$ increase per additional operating hour per week. This is closely aligned with the findings reported by Sharp and ENERGY STAR, who identified operating hours as major driver of energy use [26, 66]. Additionally, HDD and CDD significantly affected EUI, repeating the influence of local climate conditions. The observed positive coefficient for year built suggests that newer buildings tend to consume more energy per area, a trend also observed by others, likely caused by the growing prevalence of energy intensive HVAC systems in modern offices [9, 42]

While floor area is frequently reported as the single most powerful predictor of total energy consumption, its coefficient in this analysis ($0.15 KWh/m^2$) is somewhat lower than those reported in other work (e.g., Kontokosta's 0.20 kWh per m²) [42]. This may reflect differences in modeling energy use per area rather than total consumption, as well as interactions with other variables such as occupancy and building age. In particular, the coefficient for the density (0.07 $\mathrm{kWh} \cdot \mathrm{person}^{-1} \cdot \mathrm{m}^{-2}$ positive but smaller in magnitude compared to earlier research (3.07 $\mathrm{kWh} \cdot \mathrm{person}^{-1} \cdot \mathrm{m}^{-2}$ [42], possibly due to differences in scaling, normalization, or building operational characteristics in this dataset.

Nonlinear models, particularly the random forest regression, outperformed linear regression and the decision tree model. The Random Forest's superior R^2 (0.65) and lower error metrics (RMSE, MAE, MAPE) demonstrate how it captures threshold effects that the other models cannot. We speculate that the random forest performs better than the decision tree due to its ensemble nature which reduces overfitting risks inherent to single decision trees [8]

The predictive power of our linear regression model (R^2 = 0.09)is lower than that reported by Kontokosta (R^2 = 0.20) [42]. This performance difference may be due to the narrower geographic scope of Kontokosta's study, which focused solely on New York City buildings, or the inclusion of a larger number of predictor variables in their model.

Finally, while floor area is the most influential predictor of a building's total energy use, models generally achieve much higher predictive performance when estimating total consumption rather than EUI. For example, Sharp reports predictive performance ranging between R^2 = 0.74 and R^2 = 0.88 for models predicting total energy use [66]. In contrast, we deliberately chose to predict energy use per unit area to facilitate comparisons between buildings of different sizes, recognizing that this approach may inherently limit the overall fit of the model but provides more actionable insights into relative energy performance.

6 LIMITATIONS (OR THREATS TO VALIDITY)

This study has several limitations that may affect the interpretation and generalizability of the findings. Following the classification by Wohlin *et al.* [74], we organize these threats to validity as follows.

6.1 Internal Validity

Internal validity refers to whether the observed effects can confidently be attributed to the predictors rather than other factors. One threat is the potential for unobserved confounding variables, such as building envelope characteristics, HVAC system efficiency, or occupant behavior beyond the reported hours and density. While the models included climate factors (HDD and CDD) and utilization patterns (operating hours, occupant density), other relevant variables were not available in the dataset. This limitation may bias the estimated relationships.

Another internal threat is measurement error. Some variables, such as floor area and operating hours, may have been reported inconsistently across data sources. We conducted basic outlier screening and removed clearly implausible values, but residual inaccuracies may remain.

6.2 External Validity

External validity concerns the generalizability of our results to other contexts. Our dataset includes office buildings only from within the US, and thus does not represent office buildings globally, where regulations, building codes, construction practices, cultural practices and climate differs. Accordingly, results should be interpretated with caution and future work should include greater topological diversity.

6.3 Construct Validity

Construct validity relates to whether the variables adequately capture the intended concepts. As energy demands are well defined and quantifiable, most concerns go out to the climate variables. HDD and CDD aggregate weather variation but do not capture the full effect of the weather, such as influences caused by humidity, sunshine or rain. Although these variables are commonly used proxies in building energy studies, they may imperfectly represent the underlying constructs.

6.4 Conclusion Validity

Conclusion validity refers to whether the statistical conclusions are credible. One issue is the non-normality of predictor distributions, which limits the suitability of linear regression. We addressed this by applying Random Forest regression, which does not assume linearity or normality and demonstrated substantially higher explanatory power (R^2 = 0.65). Nonetheless, model performance metrics may be optimistic due to the absence of external validation on unseen datasets. To reduce overfitting risk, we used out-of-bag estimation and cross-validation, but future work should validate models on truly independent samples.

7 CONCLUSION

This study compared different models to predict building energy use and showed that nonlinear methods like random forest clearly outperform simple linear regression. Using a large dataset, we found that factors like operating hours and climate have a great influence, and that linear models miss some of the complexity in energy patterns. Our work helps improve prediction accuracy and highlights the value of capturing nonlinear effects and local climate influences.

Future research could explore more advanced machine learning approaches and include more contextual data to build models that adapt better to different building types and locations, ultimately supporting smarter energy management.

REFERENCES

[1] A. S. Ahmad, M. Y. Hassan, M. P. Abdullah, H. A. Rahman, F. Hussin, H. Abdullah, and R. Saidur. 2014. A review on applications of ANN and SVM for building electrical energy consumption forecasting. *Renewable and Sustainable Energy Reviews* 33 (2014), 102–109. https://doi.org/10.1016/j.rser.2014.01.069

[2] Jan Akander, S. Alvarez, and G. Jóhannesson. 2004. Energy normalization techniques. In *Energy normalization techniques*. School of Architecture and the Built Environment (ABE, 57–70. http://www.diva-portal.org/smash/record.jsf?pid=diva2:795809

[3] George A. Akerlof. 1970. The market for "Lemons": quality uncertainty and the market mechanism. *The Quarterly Journal of Economics* 84, 3 (1970), 488. https://doi.org/10.2307/1879431

[4] Northwest Energy Efficiency Alliance. [n.d.]. Commercial Building Stock Assessment (CBSA) Dataset. Available at: https://bpd.lbl.gov/assets/public-datasets/Northwest(Accessed: 2025-07-15).

[5] Los Angeles. [n.d.]. Los Angeles Existing Buildings Energy and Water Efficiency (EBEWE) Program Dataset. Available at: https://bpd.lbl.gov/assets/public-datasets/Los(Accessed: 2025-07-15).

[6] F. Ascione, N. Bianco, C. De Stasio, G. M. Mauro, and G. P. Vanoli. 2014. A new methodology for cost-optimal analysis by means of the multi-objective optimization of building energy performance. *Energy and Buildings* 88 (2014), 78–90. https://doi.org/10.1016/j.enbuild.2014.11.058

[7] ASHRAE. 2009. *2009 ASHRAE Handbook - Fundamentals*. American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc., Atlanta, GA.

[8] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[9] Building Energy Research Centre of Tsinghua University. 2023. China's Building Energy Use and GHG Emissions. In *Decarbonize Urban Heating System*. Springer, 1–26. https://doi.org/10.1007/978-981-99-7875-5_1

[10] K. Calvin, D. Dasgupta, G. Krinner, A. Mukherji, P. W. Thorne, C. Trisos, J. Romero, P. Aldunce, K. Barret, et al. 2023. *Climate Change 2023: Synthesis Report, Summary for Policymakers*. Technical Report. IPCC. 1–34 pages. https://doi.org/10.59327/ipcc/ar6-9789291691647.001

[11] Young Tae Chae, Raya Horesh, Youngdeok Hwang, and Young M. Lee. 2016. Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings. *Energy and Buildings* 111 (2016), 184–194. https://doi.org/10.1016/j.enbuild.2015.11.045

[12] Michel Chammas, Abdallah Makhoul, and Jacques Demerjian. 2019. An efficient data model for energy prediction using wireless sensors. *Computers Electrical Engineering* 76 (2019), 249–257. https://doi.org/10.1016/j.compeleceng.2019.04.002

[13] B. Cheng, I. Ioannou, and G. Serafeim. 2013. Corporate social responsibility and access to finance. *Strategic Management Journal* 35, 1 (2013), 1–23. https://doi.org/10.1002/smj.2131

[14] City and County of Denver. [n.d.]. Denver Benchmarking Ordinance Dataset. Available at: https://bpd.lbl.gov/assets/public-datasets/Denver(Accessed: 2025-07-15).

[15] Kansas City. [n.d.]. Kansas City Energy Benchmarking Dataset. Available at: https://bpd.lbl.gov/assets/public-datasets/Kansas(Accessed: 2025-07-15).

[16] Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge, New York. https://www.utstat.toronto.edu/brunner/oldclass/378f16/readings/CohenPower.pdf Accessed: 2025-07-15.

[17] California Energy Commission. [n.d.]. California AB 802 Building Energy Benchmarking Program Dataset.

[18] California Energy Commission. [n.d.]. California Proposition 39 K–12 Program Dataset.

[19] United Nations Climate Change Conference (COP21). [n.d.]. Paris Agreement to the United Nations Framework Convention on Climate Change. Treaty.

[20] Montgomery County. [n.d.]. Montgomery County Benchmarking Ordinance Dataset. Available at: https://bpd.lbl.gov/assets/public-datasets/Montgomery(Accessed: 2025-07-15).

[21] Arnaud de Myttenaere, Boris Golden, Bénédicte Le Grand, and Fabrice Rossi. 2016. Mean Absolute Percentage Error for regression models. *Neurocomputing* 192 (2016), 38–48. https://doi.org/10.1016/j.neucom.2015.12.114 Advances in artificial neural networks, machine learning and computational intelligence.

[22] C. Deb, L. S. Eang, J. Yang, and M. Santamouris. 2016. Forecasting diurnal cooling energy load for institutional buildings using Artificial Neural Networks. *Energy and Buildings* 121 (2016), 284–297. https://doi.org/10.1016/j.enbuild.2015.12.050

[23] A. I. Dounis and C. Caraiscos. 2009. Advanced control systems engineering for energy and comfort management in a building environment: A review. *Renewable and Sustainable Energy Reviews* 13, 6 (2009), 1246–1261. https://doi.org/10.1016/j.rser.2008.09.015

[24] Piet Eichholtz, Nils Kok, and John M. Quigley. 2010. Doing Well by Doing Good? Green Office Buildings. *The American Economic Review* 100, 5 (2010), 2492–2509. http://www.jstor.org/stable/41038771

[25] Austin Energy. [n.d.]. Austin Energy Conservation Audit and Disclosure (ECAD) Ordinance Dataset.

[26] ENERGY STAR. 2019. *ENERGY STAR score for offices in the United States*. Technical Report. ENERGY STAR. https://www.energystar.gov/sites/default/files/tools/Office_August_2019_508.pdf Technical Reference.

[27] Eliot Bytyçi Erblin Halabaku. 2024. Overfitting in Machine Learning: A Comparative Analysis of Decision Trees and Random Forests. *Intelligent Automation & Soft Computing* 39, 6 (2024), 987–1006. http://www.techscience.com/iasc/v39n6/59139

[28] European Commission. 2021. Energy performance of buildings directive. https://ec.europa.eu/. European Commission Press Release.

[29] Samuel Famiyeh and Agyemang Kwarteng. 2018. Implementation of environmental management practices in the Ghanaian mining and manufacturing supply chains. *International Journal of Productivity and Performance Management* 67, 7 (2018), 1091–1112. https://doi.org/10.1108/ijppm-04-2017-0095

[30] Caroline Flammer. 2012. Corporate social responsibility and shareholder reaction: the environmental awareness of investors. *Academy of Management Journal* 56, 3 (2012), 758–781. https://doi.org/10.5465/amj.2011.0744

[31] San Francisco. [n.d.]. San Francisco Existing Buildings Energy Performance Ordinance Dataset. Available at: https://bpd.lbl.gov/assets/public-datasets/San(Accessed: 2025-07-15).

[32] Franz Fuerst and Patrick McAllister. 2011. The impact of Energy Performance Certificates on the rental and capital values of commercial property assets. *Energy Policy* 39, 10 (2011), 6608–6614. https://doi.org/10.1016/j.enpol.2011.08.005 Sustainability of biofuels.

[33] Gainesville. [n.d.]. Gainesville Green Home Energy Tracking Dataset. Available at: https://bpd.lbl.gov/assets/public-datasets/Gainesville(Accessed: 2025-07-15).

[34] Kirsten Gram-Hanssen. 2010. Residential heat comfort practices: understanding users. *Building Research & Information* 38, 2 (2010), 175–186. https://doi.org/10.1080/09613210903541527

[35] Hannah Granade, Jon Creyts, Anton Derkach, Philip Farese, Scott Nyquist, and Ken Ostrowski. 2009. *Unlocking Energy Efficiency in the U.S. Economy*. Technical Report. McKinsey & Company. https://www.mckinsey.com/~/media/mckinsey/dotcom/client_service/epng/pdfs/unlocking%20energy%20efficiency/us_energy_efficiency_exc_summary.ashx

[36] Mohamad Harasheh, Ahmed Bouteska, and Rania Manita. 2023. Investors' preferences for sustainable investments: Evidence from the U.S. using an experimental approach. *Economics Letters* 234 (2023), 111428. https://doi.org/10.1016/j.econlet.2023.111428

[37] Tianzhen Hong, Le Yang, David Hill, and Wei Feng. 2014. Data and analytics to inform energy retrofit of high performance buildings. *Applied Energy* 126 (2014), 90–106. https://doi.org/10.1016/j.apenergy.2014.03.052

[38] Po-Hsuan Hsu, Xuan Tian, and Yan Xu. 2014. Financial development and innovation: Cross-country evidence. *Journal of Financial Economics* 112, 1 (2014), 116–135. https://doi.org/10.1016/j.jfineco.2013.12.002

[39] Rob J. Hyndman and Anne B. Koehler. 2006. Another look at measures of forecast accuracy. *International Journal of Forecasting* 22, 4 (2006), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

[40] H.G. Kim and S.S. Kim. 2020. Development of energy benchmarks for office buildings using the National Energy Consumption Database. *Energies* 13, 4 (2020), 950. https://doi.org/10.3390/en13040950

[41] J. Kissock, J. Haberl, and D. Claridge. 2003. *Inverse Modeling Toolkit: Numerical Algorithms for Best-Fit Variable-Base Degree Day and Change Point Models*. Technical Report. ASHRAE. https://oaktrust.library.tamu.edu/bitstream/1969.1/153708/1/ESL-PA-03-07-03.pdf

[42] Constantine E. Kontokosta. 2012. Predicting Building Energy Efficiency Using New York City Benchmarking Data. In *ACEEE Summer Study on Energy Efficiency in Buildings*. https://aceee.org/files/proceedings/2012/data/papers/0193-000114.pdf

[43] Lennart Kühn, Nikolas Fuchs, Lukas Braun, Leon Maier, and Dirk Müller. 2024. Landlord–Tenant Dilemma: How Does the Conflict Affect the Design of Building Energy Systems? *Energies* 17, 3 (2024), 686. https://doi.org/10.3390/en17030686

[44] H. Liu, J. Liang, Y. Liu, and H. Wu. 2023. A review of Data-Driven Building Energy Prediction. *Buildings* 13, 2 (2023), 532. https://doi.org/10.3390/buildings13020532

[45] Fannie Mae. [n.d.]. Fannie Mae Multifamily Energy Survey Dataset. Available at: https://bpd.lbl.gov/assets/public-datasets/Fannie(Accessed: 2025-07-15).

[46] F. Magoulès, H. X. Zhao, and D. Elizondo. 2013. Development of an RDP neural network for building energy consumption fault detection and diagnosis. *Energy and Buildings* 62 (2013), 133–138. https://doi.org/10.1016/j.enbuild.2013.02.050

[47] Ali Makhmalbaf, Varun Srivastava, and Ning Wang. 2013. Simulation-based weather normalization approach to study the impact of weather on energy use of buildings in the U.S.. In *Proceedings of BS2013: 13th Conference of International Building Performance Simulation Association*. https://buildingenergyscore.energy.gov/publications/weather_normalization.pdf [Conference proceeding].

[48] D. Mariano-Hernández, L. Hernández-Callejo, A. Zorita-Lamadrid, O. Duque-Pérez, and F. Santos García. 2021. A review of strategies for building energy management system: Model predictive control, demand side management, optimization, and fault detect diagnosis. *Journal of Building Engineering* 33 (2021), 101692. https://doi.org/10.1016/j.jobe.2020.101692

[49] Norman Miller, James Spivey, and Andrew Florance. 2008. Does Green Pay Off? *Journal of Real Estate Portfolio Management* 14, 4 (2008), 385–400. https://doi.org/10.1080/10835547.2008.12089822

[50] T. Nikolaou, D. Kolokotsa, G. Stavrakakis, A. Apostolou, and C. Munteanu. 2015. Review and State of the Art on Methodologies of Buildings' Energy-Efficiency Classification. In *Managing Indoor Environments and Energy in Buildings with Integrated Intelligent Systems*. Springer International Publishing, 13–31. https://doi.org/10.2174/9781608052851120101

[51] NOAA Climate Prediction Center. [n.d.]. Degree Days - U.S. Climate Data. https://www.cpc.ncep.noaa.gov/products/analysis_monitoring/cdus/degree_days/. Accessed: 2025-07-15.

[52] City of Berkeley. [n.d.]. Berkeley Building Emissions Saving Ordinance (BESO) Dataset.

[53] City of Boston. [n.d.]. Boston Building Emissions Reduction and Disclosure Ordinance (BERDO) Dataset.

[54] City of Boulder. [n.d.]. Boulder Building Performance Program Dataset.

[55] City of Cambridge. [n.d.]. Cambridge Building Energy Use Disclosure Ordinance (BEUDO) Dataset.

[56] City of Chicago. [n.d.]. Chicago Energy Benchmarking Ordinance Dataset. Available at: https://bpd.lbl.gov/assets/public-datasets/Chicago(Accessed: 2025-07-15).

[57] City of Evanston. [n.d.]. Evanston Benchmarking Ordinance Dataset. Available at: https://bpd.lbl.gov/assets/public-datasets/Evanston(Accessed: 2025-07-15).

[58] A. Olhoff, C. Bataille, J. Christensen, M. van den Elzen, T. Fransen, N. Grant, K. Blok, J. Kejun, E. Soubeyran, W. Lamb, K. Levin, J. Portugal-Pereira, M. Pathak, T. Kuramochi, C. Strinati, S. Roe, and J. Rogelj. 2024. Emissions Gap Report 2024: No more hot air . . . please! With a massive gap between rhetoric and reality, countries draft new climate commitments. https://doi.org/10.59117/20.500.11822/46404. United Nations Environment Programme.

[59] Orlando. [n.d.]. Orlando Building Energy Benchmarking Policy Dataset. Available at: https://bpd.lbl.gov/assets/public-datasets/Orlando(Accessed: 2025-07-15).

[60] Philadelphia. [n.d.]. Philadelphia Large Building Energy Benchmarking Ordinance Dataset. Available at: https://bpd.lbl.gov/assets/public-datasets/Philadelphia(Accessed: 2025-07-15).

[61] Portland. [n.d.]. Portland Building Energy Performance Reporting Dataset. Available at: https://bpd.lbl.gov/assets/public-datasets/Portland(Accessed: 2025-07-15).

[62] José Ramón Pérez and Fernando J. Rey. 2008. Climate change impacts on heating and cooling energy demand in buildings in the EU-25. *Energy and Buildings* 40, 12 (2008), 2227–2236. https://doi.org/10.1016/j.enbuild.2008.06.005

[63] Olivia G. Santin, Laure Itard, and Henk Visscher. 2009. The effect of occupancy and building characteristics on energy use for space and water heating in Dutch residential stock. *Energy and Buildings* 41, 11 (2009), 1223–1232. https://doi.org/10.1016/j.enbuild.2009.07.002

[64] Seattle. [n.d.]. Seattle Energy Benchmarking Law Dataset. Available at: https://bpd.lbl.gov/assets/public-datasets/Seattle(Accessed: 2025-07-15).

[65] S. Seyedzadeh, F. Rahimian, I. Glesk, et al. 2018. Machine learning for estimation of building energy consumption and performance: a review. *Visual Computing for Industry, Biomedicine, and Art* 6 (2018), 5. https://doi.org/10.1186/s40327-018-0064-7

[66] T. Sharp and Oak Ridge National Laboratory. n.d.. Energy benchmarking in commercial office buildings. Retrieved from https://www.aceee.org/files/proceedings/1996/data/papers/SS96_Panel4_Paper33.pdf.

[67] A. C. Shukla, S. Deshmukh, and A. Kanda. 2009. Environmentally responsive supply chains. *Journal of Advances in Management Research* 6, 2 (2009), 154–171. https://doi.org/10.1108/09727980911007181

[68] F. Smarra, A. Jain, T. de Rubeis, D. Ambrosini, A. D'Innocenzo, and R. Mangharam. 2018. Data-driven model predictive control using random forests for building energy optimization and climate control. *Applied Energy* (2018). https://doi.org/10.1016/J.APENERGY.2018.02.126

[69] Syracuse. [n.d.]. Syracuse Building Energy Benchmarking Ordinance Dataset. Available at: https://bpd.lbl.gov/assets/public-datasets/Syracuse(Accessed: 2025-07-15).

[70] J. Teng and H. Yin. 2023. Forecasting the carbon footprint of civil buildings under different floor area growth trends and varying energy supply methods. *Scientific Reports* 13, 1 (2023). https://doi.org/10.1038/s41598-023-49270-3

[71] U.Y.A. Tettey, A. Dodoo, and L. Gustavsson. 2018. Design strategies and measures to minimise operation energy use for passive houses under different climate scenarios. *Energy Efficiency* 12, 1 (2018), 299–313. https://doi.org/10.1007/s12053-018-9719-4

[72] United Nations Environment Programme. 2022. *Emissions Gap Report 2022*. Technical Report. United Nations Environment Programme. https://doi.org/10.18356/9789210023993

[73] U.S. Energy Information Administration. 2012. *Annual Energy Review 2011*. Technical Report. Washington, DC. https://www.eia.gov/totalenergy/data/annual/pdf/aer.pdf

[74] Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Experimentation in Software Engineering*. Springer Science & Business Media.

[75] Hongjie Wu, Zengwei Yuan, Lei Zhang, and Jun Bi. 2011. Life cycle energy consumption and CO2 emission of an office building in China. *The International Journal of Life Cycle Assessment* 17, 2 (2011), 105–118. https://doi.org/10.1007/s11367-011-0342-2

[76] New York. [n.d.]. New York Local Law 84 Benchmarking Dataset. Available at: https://bpd.lbl.gov/assets/public-datasets/New(Accessed: 2025-07-15).

[77] New York. [n.d.]. New York Local Law 87 Energy Audits Dataset. Available at: https://bpd.lbl.gov/assets/public-datasets/New(Accessed: 2025-07-15).

[78] New York. [n.d.]. New York Residential Statewide Baseline Study Dataset. Available at: https://bpd.lbl.gov/assets/public-datasets/New(Accessed: 2025-07-15).

[79] Zhun Yu, B. C. M. Fung, Fariborz Haghighat, Hiroshi Yoshino, and Edward Morofsky. 2011. A systematic procedure to study the influence of occupant behavior on building energy consumption. *Energy and Buildings* 43, 6 (2011), 1409–1417. https://doi.org/10.1016/j.enbuild.2011.02.002

[80] Hongxia Zhao and Frédéric Magoulès. 2012. A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews* 16, 6 (2012), 3586–3592. https://doi.org/10.1016/j.rser.2012.02.049

[81] Björn Ástmarsson, Per Jensen, and Esmir Maslesa. 2013. Sustainable renovation of residential buildings and the landlord/tenant dilemma. *Energy Policy* 63 (12 2013), 355–362. https://doi.org/10.1016/j.enpol.2013.08.046

Table 1: ***Descriptive statistics of building dataset variables***

Variable	*Unit*	*N*	*Mean*	*Median*	*Mode*	*SD*	*IQR*	*Min*	*Max*
id	categorical	32,686	–	–	PM1030273	–	–	–	–
year	year	32,686	2018.03	2019	2020	3.06	4	2010	2023
floor_area	m^2	32,686	24,121.99	10,866	5,574	36,099.34	20,437.75	80	427,412
year_built	year	32,256	1959.99	1969	1984	36.40	62	1,726	2022
number_of_people	number	11,847	812.93	290	100	1,390.71	677	1	14,673
occupant_density	occupants/m^2	11,847	230.88	213.44	213.68	166.44	121.70	0.554	5,515.72
operating_hours	hours/week	11,942	67.90	65	65	26.16	16	0.61	168
energy_star_rating	scale	29,216	68.21	75	83	24.10	30	1	100
electric_eui	MWh	32,686	3,873.73	1,554.67	117,771.78	7,362.82	3,237.91	0	158,361.14
fuel_eui	MWh	32,686	2,167.72	719.96	0	5,577.84	1,399.74	0	122,015.35
site_eui	MWh	32,550	6,054.90	2,450.57	177,818.08	11,798.23	4,545.35	13.52	251,713.44
cooling_dd	CDD	32,686	1,183.59	1,156	1,156	444.93	625	9	3,017
heating_dd	HDD	32,686	4,033.47	4,195	2,374	1,413.06	2,656	684	8,650
energy_intensity	MWh/m^2	32,550	0.245	0.211	0.657	0.168	0.128	0.00375	3.143

Table 2: ***Descriptive statistics of building dataset skewness, kurtosis, and normality tests***

Variable	*Z_Skewness*	*Abs_Skewness*	*Z_Kurtosis*	*Abs_Kurtosis*	*Shapiro_p*	*Levene_p*
year	-59.32	0.80	-4.07	0.11	1.31E-31	2.09E-132
floor_area	289.93	3.93	785.64	21.29	3.09E-56	9.88E-33
year_built	-27.08	0.37	-27.45	0.75	9.08E-28	0.0063
number_of_people	161.62	3.64	388.31	17.48	2.96E-57	2.39E-107
occupant_density	481.48	10.84	6448.35	290.24	3.79E-43	5.42E-06
operating_hours	100.98	2.26	135.88	6.09	1.80E-48	6.81E-07
energy_star_rating	-74.06	1.06	16.09	0.46	6.39E-35	0
electric_eui	493.09	6.68	2725.00	73.84	1.69E-60	1.61E-181
fuel_eui	631.92	8.56	4115.02	111.51	1.14E-65	3.46E-299
site_eui	478.76	6.50	2407.65	65.38	3.64E-61	9.48E-269
cooling_dd	-1.31	0.02	12.45	0.34	1.02E-16	4.64E-42
heating_dd	-28.21	0.38	-24.05	0.65	8.49E-23	8.77E-223
energy_intensity	371.90	5.05	1917.26	52.06	1.38E-47	0

Table 3: ***Correlation matrix of selected building variables***

Variable	*operating_hours*	*occupant_density*	*heating_dd*	*cooling_dd*	*year_built*	*floor_area*
operating_hours	1	0.0861	-0.0361	0.0321	0.0845	0.1463
occupant_density	0.0861	1	-0.0537	0.0879	0.0107	0.0454
heating_dd	-0.0361	-0.0537	1	-0.0925	-0.2929	0.0858
cooling_dd	0.0321	0.0879	-0.0925	1	-0.1396	0.1034
year_built	0.0845	0.0107	-0.2929	-0.1396	1	0.0867
floor_area	0.1463	0.0454	0.0858	0.1034	0.0867	1

Table 4: ***Variance Inflation Factors (VIF) for building variables***

Variable	*VIF*
operating_hours	1.032
occupant_density	1.018
heating_dd	3.986
cooling_dd	4.074
year_built	1.185
floor_area	1.179

Table 5: ***Regression coefficients and statistics***

Term	*Estimate*	*Std. Error*	*Statistic*	*p-value*
(Intercept)	-0.9795	0.1089	-8.997	2.78E-19
operating_hours	0.00135	6.12E-05	22.00	1.12E-104
occupant_density	6.97E-05	1.05E-05	6.66	2.96E-11
heating_dd	3.52E-05	5.29E-06	6.66	2.86E-11
cooling_dd	7.13E-05	1.01E-05	7.09	1.40E-12
year_built	0.00045	5.28E-05	8.59	9.90E-18
floor_area	1.53E-07	4.00E-08	3.83	1.31E-04

Table 6: ***Performance metrics of different models***

Model	*R_squared*	*MSE*	*RMSE*	*MAE*	*MAPE*	*MBE*	*CV_RMSE*
Linear_Regression	0.0894	0.0170	0.1305	0.0862	44.70	-0.0060	0.4946
Decision_Tree	0.2495	0.0141	0.1189	0.0783	39.57	-0.0031	0.4507
Random_Forest	0.6532	0.0065	0.0806	0.0477	23.63	-0.0048	0.3055