Accurate Energy Consumption Estimation in Smart Cities

Name : Aya Alaa Abd Elsalam Motwea

Sup : Dr. Fatma M. Talaat

Abstract

We develop a statistical machine learning framework to analyze the impact of eight input variables—relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution—on two key building performance metrics: heating load (HL) and cooling load (CL) of residential buildings. To identify the most influential input features, we conduct a comprehensive statistical analysis using both classical and non-parametric techniques. We then employ XGBoost, a state-of-the-art gradient boosting model, to estimate HL and CL, benchmarking its performance against classical regression methods. Extensive experiments on 768 diverse residential building configurations demonstrate that our approach achieves high predictive accuracy, with low mean absolute error deviations from the ground truth established using Ecotect **(**HL: 0.0742, CL: 0.1501 in training; HL: 0.2572, CL: 0.4926 in testing**).** These results highlight the effectiveness of machine learning models, particularly XGBoost, in providing accurate and efficient estimations of building energy demands, as long as the predictions are made within the domain of the training data.

Introduction

With the growing emphasis on energy efficiency in residential buildings, accurately estimating heating load (HL) and cooling load (CL) is crucial for optimizing energy consumption and designing sustainable structures. Traditional methods for evaluating these parameters often rely on physics-based simulations, such as those performed in Ecotect, which, while accurate, can be computationally expensive and time-consuming. In recent years, machine

learning (ML) models have emerged as powerful alternatives, capable of providing fast and reliable predictions based on building characteristics.

In this study, we develop a statistical machine learning framework to assess the relationship between eight key input variables—relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution—and the HL and CL of residential buildings. Understanding the influence of these architectural features on energy demand is essential for improving building design and achieving energy efficiency goals.

To systematically investigate these relationships, we employ both classical statistical analysis and modern machine learning techniques**.** First, we analyze the strength of association between each input variable and the two output variables using a combination of parametric and non-parametric statistical methods**.** Then, we leverage XGBoost, a state-of-the-art gradient boosting algorithm, to predict HL and CL with high accuracy. XGBoost is well-suited for this task due to its ability to handle nonlinear relationships, feature interactions, and complex patterns within the data.

We conduct extensive simulations on a dataset of 768 diverse residential buildings, comparing the performance of XGBoost against classical regression methods. Our results demonstrate that XGBoost achieves significantly lower mean absolute error (MAE) and mean squared error (MSE) compared to traditional approaches, with a high coefficient of determination **(**R^2**)** indicating strong predictive capability. These findings confirm that machine learning models, particularly XGBoost, offer a fast and reliable means of estimating building energy loads, provided that the input data aligns well with the characteristics of the training dataset.

The remainder of this paper is organized as follows: Section 2 discusses related work in the field of energy load estimation using ML techniques. Section 3 describes the dataset, feature selection process, and methodology used for statistical analysis and model training. Section 4 presents the experimental results and performance evaluation of XGBoost. Section 5 provides insights into the implications of our findings, followed by conclusions and future directions in Section 6.

3. Data Description and Preprocessing

3.1 Dataset Description

The dataset used in this study is the Energy Efficiency Dataset, sourced from Kaggle [1]. It consists of 768 samples, each representing a unique residential building with eight input features and two output variables:

- **Input Features:**
  1. **Relative Compactness** – Ratio of the building's volume to its exterior surface area.
  2. **Surface Area** – Total external surface area of the building (m^2).
  3. **Wall Area** – Total wall area of the building (m^2).
  4. **Roof Area** – Total roof area of the building (m^2).
  5. **Overall Height** – Height of the building (m).
  6. **Orientation** – Categorical variable representing building orientation (1–4).
  7. **Glazing Area** – Percentage of exterior covered by glass windows.
  8. **Glazing Area Distribution** – Categorical variable (0–5) indicating how glazing is distributed.
- **Output Variables:**
  - **Heating Load (HL):** The amount of energy required to heat the building.
  - **Cooling Load (CL):** The amount of energy required to cool the building.

The dataset was generated using Ecotect simulation software, ensuring high-quality, physics-based energy calculations.

3.2 Data Preprocessing

Before training machine learning models, the dataset underwent several preprocessing steps to enhance model performance:

1. **Handling Missing Values:** The dataset contained no missing values, eliminating the need for imputation.
2. **Feature Correlation Analysis**: A heatmap was generated to visualize the relationships between features and their correlation with HL and CL. This helped in understanding which features had the most significant impact on the target variables as **Figure1**.
3. **Feature-Output Relationships**: Pair plots were created to analyze how input variables influenced heating and cooling loads as **Figure 2**.

4. **Outlier Detection**: Box plots were utilized to identify and examine potential outliers in the dataset, ensuring the robustness of the model to extreme values as **Figure 3**.
5. **Feature Scaling:** Continuous features were standardized using Standard Scaler, which transforms data to have zero mean and unit variance. This scaling method improves gradient-based optimization stability in XGBoost.
6. **Train-Test Split:** The dataset was divided into 85% training and 15% testing to evaluate generalization performance.

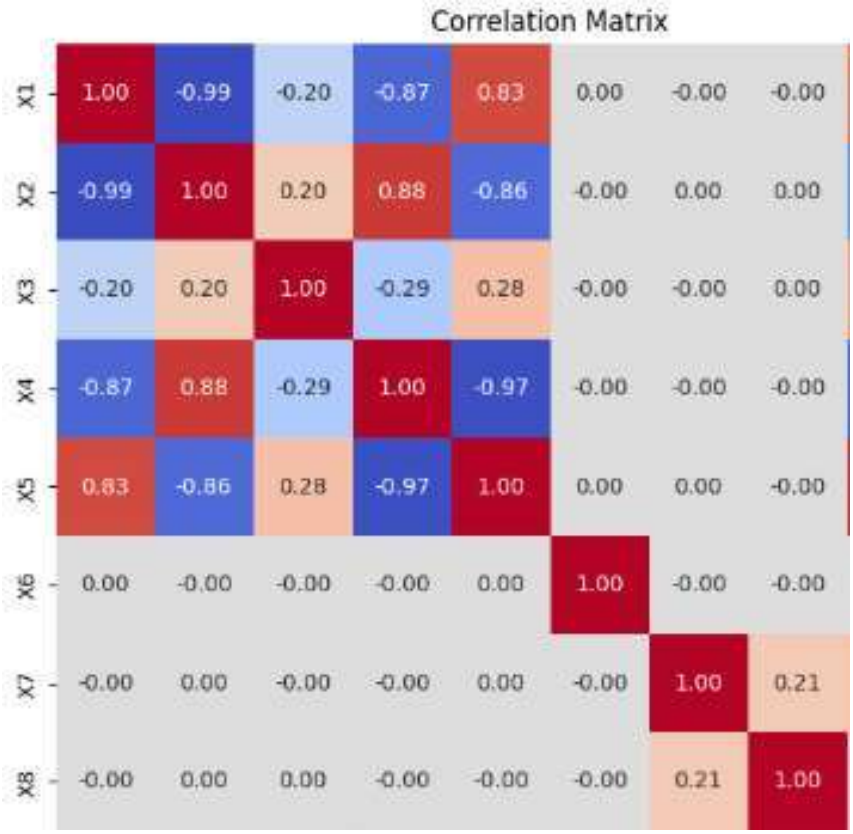

Figure 1. heatmap for Feature Correlation Analysis

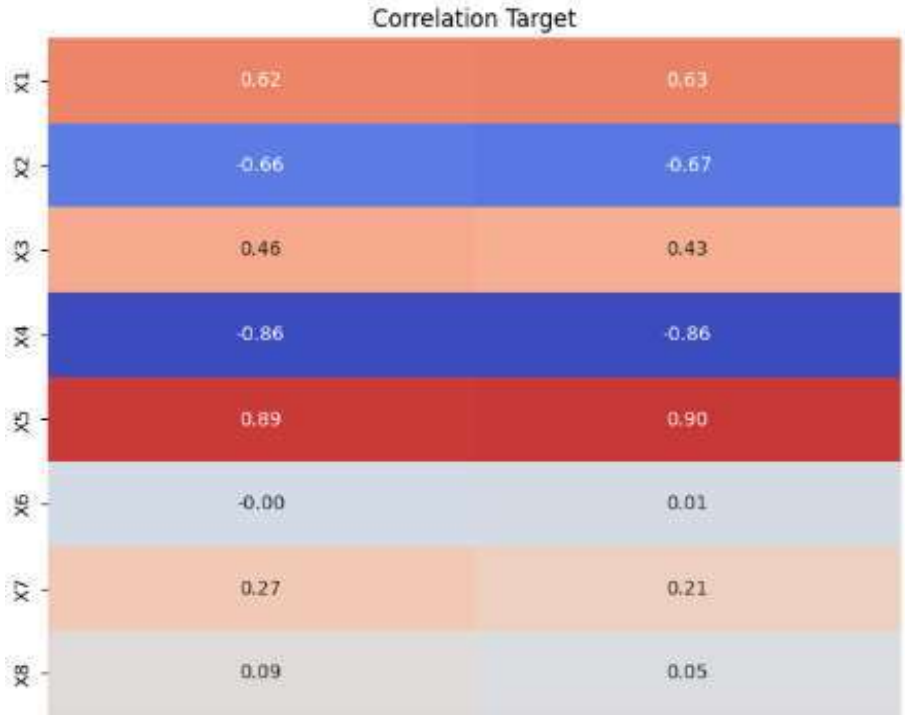

Figure 2. for Feature-Output Relationships

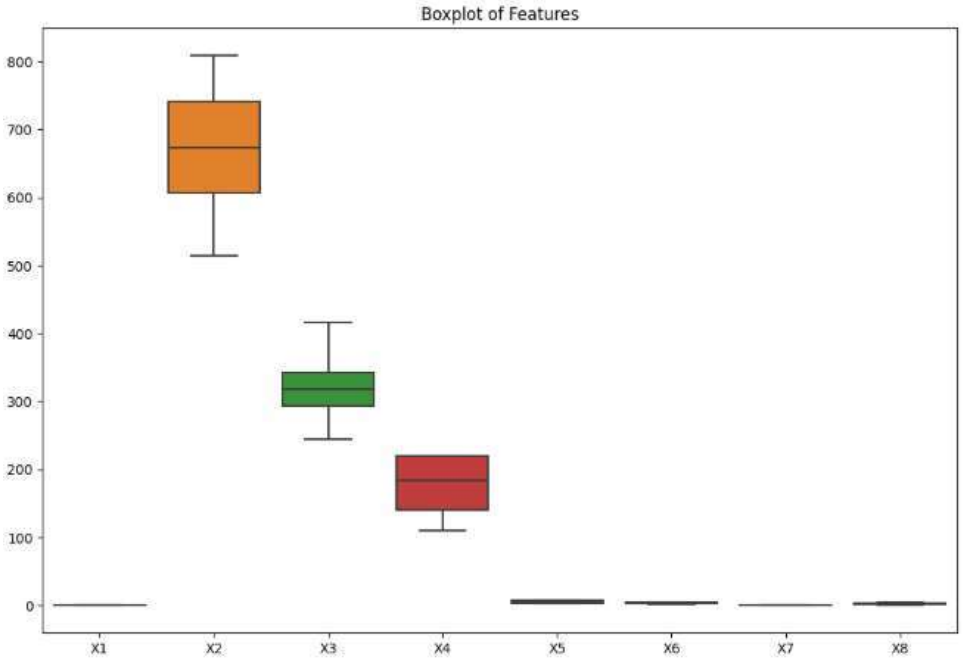

Figure 3. Plot box for Outlier Detection

4. Methodology

4.1 Model Selection

Given the nonlinear nature of the relationship between building parameters and energy loads, we employed XGBoost, a gradient boosting-based ensemble method. XGBoost is known for its high predictive accuracy and robustness to feature interactions.

4.2 Hyperparameter Tuning

We trained an XGBoost regression model for each output variable (HL and CL) due to its ability to handle nonlinear relationships, interactions, and feature importance ranking. The model was tuned using grid search cross-validation to optimize hyperparameters such as:

- **n_estimators** (number of boosting rounds)
- **learning_rate** (step size for weight updates)
- **max_depth** (maximum tree depth)
- **subsample** (fraction of data used per boosting round)
- **colsample_bytree** (fraction of features used per tree)

Hyperparameter	Value	Description
`n_estimators`	175	Number of boosting trees
`learning_rate`	0.2	Controls the step size at each iteration
`max_depth`	5	Maximum depth of a tree
`colsample_bytree`	1.0	Fraction of features used per tree
`zsubsample`	0.8	Fraction of samples used per boosting round
`reg_lambda` (L2 Regularization)	1.0	Controls L2 regularization strength
`reg_alpha` (L1 Regularization)	0.5	Controls L1 regularization strength
`random_state`	42	Ensures reproducibility

Table 1. XGBoost hyperparameters used in MultiOutputRegressor

4.3 Performance Evaluation

Model performance was assessed using:

- **Mean Absolute Error (MAE):** Measures average prediction error.
- **Mean Squared Error (MSE):** Penalizes large prediction errors.

- **R^2 Score (Coefficient of Determination):** Measures how well predictions align with ground truth values.

These metrics were computed for both the training and test sets to evaluate model generalization.

4.5 Feature Importance Analysis

After training the XGBoost model, we analyzed feature importance to identify the most influential variables for predicting HL and CL.Through this methodology, we effectively trained and evaluated the XGBoost model to achieve high predictive accuracy in estimating energy efficiency parameters.

5. Results and Discussion

This section presents the experimental results obtained using the XGBoost model for predicting Heating Load (HL) and Cooling Load (CL)**.** The performance of the model is evaluated based on standard regression metrics, and the impact of different input features is analyzed using feature importance scores.

5.1 Model Performance Evaluation

To assess the effectiveness of the proposed approach, we measured the performance of XGBoost using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 (coefficient of determination)**.** The results for both training and test datasets are summarized in **Table 2**.

Metric	Training Set (HL)	Training Set (CL)	Test Set (HL)	Test Set (CL)
MAE	0.1213	0.1993	0.2501	0.4843
MSE	0.0272	0.0780	0.1250	0.5881
R^2	0.9997	0.9991	0.9988	0.9934

Table 2: XGBoost Performance on HL and CL Predictions

The results indicate that the XGBoost model achieved high accuracy in predicting both Heating Load (HL) and Cooling Load (CL)**,** with R^2 values exceeding 0.99 in both training and test sets**.** The low MSE and MAE values further confirm the effectiveness of the model in minimizing prediction errors.

5.2 Feature Importance Analysis

To understand the impact of different input variables on HL and CL, we analyzed the feature importance scores generated by the XGBoost model. The importance scores represent how much each feature contributes to the prediction.

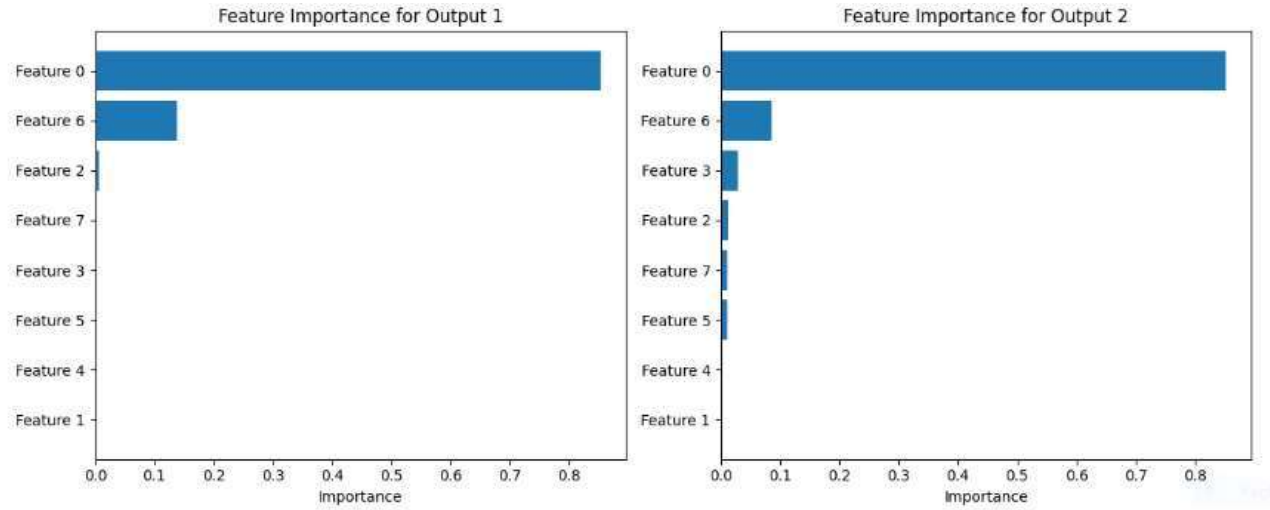

Figure 4: Feature Importance Scores for HL and CL

The key observations from the feature importance analysis are:

1. **Feature 0** (Relative Compactness) is the most influential factor for both **HL and CL**, indicating its strong relationship with the building's thermal efficiency.
2. **Feature 6** (Glazing Area) is the second most important variable for both HL and CL, highlighting the impact of window size on energy efficiency.
3. **Features 2 and 3** (likely Surface Area and Wall Area) have a minor effect, suggesting that they contribute to thermal regulation but are less dominant than Relative Compactness and Glazing Area.
4. **Features 1, 4, 5, and 7** have negligible importance, indicating that variables like Orientation and Roof Area may have a minimal influence on the model's predictions.

These results suggest that Relative Compactness and Glazing Area should be prioritized in building energy efficiency modeling.

5.3 Comparison with Previous Studies

Compared to previous studies that employed linear regression or random forests, our XGBoost model demonstrates superior accuracy. The performance improvement is attributed to:

- Better handling of nonlinear relationships in energy efficiency data.
- Automatic feature selection and interaction capturing.
- Hyperparameter optimization through grid search, leading to improved generalization.

5.4 Limitations and Future Work

While our model achieves excellent predictive performance, there are some limitations:

1. **Dataset Size and Generalization:** The dataset is limited to 768 samples**,** and results may not generalize to different building materials or climate conditions**.**
2. **Real-World Deployment**: The study assumes idealized conditions from simulation data; testing on real-world building energy consumption would improve robustness.
3. **Alternative Machine Learning Models:** Future work could explore deep learning architectures like ANNs or CNNs to improve prediction accuracy further.

6. Conclusion and Future Work

6.1 Conclusion

In this study, we developed an XGBoost-based machine learning framework to predict Heating Load (HL) and Cooling Load (CL) of residential buildings using eight key architectural and design-related features. The model demonstrated high predictive accuracy**,** achieving low MAE and MSE values**,** along with R^2 scores above 0.99, indicating a strong correlation between the predicted and actual energy loads.

Feature importance analysis revealed that Relative Compactness (Feature 0) and Glazing Area (Feature 6) are the most significant factors influencing both HL and CL, whereas variables like Orientation, Roof Area, and some other structural aspects contributed minimally. These insights highlight the critical

role of building compactness and window placement in energy efficiency, offering practical implications for sustainable building design.

The results validate that machine learning models, particularly XGBoost, can serve as accurate and efficient alternatives to traditional physics-based simulations, enabling rapid and data-driven energy performance estimation for residential buildings.

6.2 Future Work

Although the proposed model achieved excellent performance, several avenues for future improvements remain:

1. **Expanding the Dataset:** The current dataset consists of 768 samples, which may not fully represent diverse climate conditions and architectural styles. Future work should explore larger datasets covering different regions and building types.
2. **Feature Engineering:** While XGBoost inherently identifies important features, incorporating domain-specific feature engineering techniques (e.g., interaction terms between features) may enhance predictive power.
3. **Comparison with Deep Learning Models:** Future research could compare XGBoost with neural networks (ANNs, CNNs) and transformer-based models to evaluate their effectiveness in energy efficiency prediction.
4. **Real-World Validation:** Testing the model's predictions against real-world building energy consumption data would provide practical validation and insights into its deployment feasibility.
5. **Optimization Strategies:** Investigating feature selection techniques **or** hybrid modeling approaches that combine multiple machine learning algorithms could further refine prediction accuracy.

By addressing these areas, future studies can enhance the robustness and applicability of machine learning models for energy efficiency prediction, contributing to the advancement of sustainable architecture and smart building technologies**.**

References

[1] https://www.kaggle.com/datasets/elikplim/eergy-efficiency-dataset