Predictive Modeling of Thermal Efficiency in Residential Buildings: Analyzing Heat Absorption and Emission Based on Architectural Features

Le Nguyen Quoc Anh
SE192149 AI1907

***Abstract*—This study investigates predictive modeling for thermal efficiency in residential buildings by analyzing architectural features that affect heat absorption and emission. Utilizing a dataset with attributes such as compactness, surface area, and glazing properties, we assess how these factors influence heating and cooling loads. Regression models, including Multiple Linear Regression, Random Forest, and Support Vector Regression (SVR), provide insights for optimizing building designs to enhance energy efficiency and support sustainable architectural practices.**

I. Introduction

Efficient energy use in residential buildings is essential for sustainability and cost-effectiveness. Designing with energy efficiency in mind can optimize heating and cooling loads, reducing energy consumption. The primary goal of this study is to predict the heating load ($y1$) and the cooling load ($y2$) based on various architectural characteristics, as shown in Table 1.

TABLE 1
Data Description

Variable	Description
$X1$	Relative Compactness
$X2$	Surface Area
$X3$	Wall Area
$X4$	Roof Area
$X5$	Overall Height
$X6$	Orientation
$X7$	Glazing Area
$X8$	Glazing Area Distribution
$y1$	Heating Load
$y2$	Cooling Load

II. Analytical Approach

The analysis involved the following steps:

1) **Data Cleaning:** Addressed missing values and verified data integrity.
2) **Descriptive Statistics:** Calculated metrics such as mean, median, variance, and standard deviation to understand feature distributions.
3) **Correlation Analysis:** Assessed relationships between variables using the Pearson correlation coefficient:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}. \quad (1)$$

4) **Data Visualization:** Used heatmaps, scatter plots and others to illustrate relationships between variables.
5) **Model Development and Evaluation:** Developed and evaluated regression models (Multiple Linear Regression, Random Forest, and Support Vector Regression) based on Mean Squared Error (MSE) and the R^2 score.

III. Data and Descriptive Statistics

The dataset used in this analysis consists of 768 samples, each capturing attributes related to building characteristics, as well as their corresponding heating and cooling loads. Descriptive statistics for each variable are provided in Table 2, which summarizes key measures like mean, standard deviation, minimum, and maximum values for all attributes. This statistical summary provides insight into the data distribution and variability, setting the foundation for further analysis.

TABLE 2
Descriptive Statistics of Dataset Variables

Attribute	Mean	Std Dev	Min	Max
Heating Load	21.34	9.5	10.1	42.5
Cooling Load	15.3	7.8	5.6	31.6
Wall Area	85.2	15.3	40.5	105.6

	Relative_Compactness	Surface_Area	Wall_Area	Roof_Area	Overall_Height	Orientation	Glazing_Area	Glazing_Area_Distribution	Heating_Load	Cooling_Load
count	768.000000	768.000000	768.000000	768.000000	768.00000	768.000000	768.000000	768.00000	768.000000	768.000000
mean	0.764167	671.708333	318.500000	176.604167	5.25000	3.500000	0.234375	2.81250	22.307201	24.587760
std	0.105777	88.086116	43.626481	45.165950	1.75114	1.118763	0.133221	1.55096	10.090196	9.513306
min	0.620000	514.500000	245.000000	110.250000	3.50000	2.000000	0.000000	0.00000	6.010000	10.900000
25%	0.682500	606.375000	294.000000	140.875000	3.50000	2.750000	0.100000	1.75000	12.992500	15.620000
50%	0.750000	673.750000	318.500000	183.750000	5.25000	3.500000	0.250000	3.00000	18.950000	22.080000
75%	0.830000	741.125000	343.000000	220.500000	7.00000	4.250000	0.400000	4.00000	31.667500	33.132500
max	0.980000	808.500000	416.500000	220.500000	7.00000	5.000000	0.400000	5.00000	43.100000	48.030000

Fig. 1. Table of descriptive statistics.

In Figure 1, the table is shows descriptive statistics for various building features in a dataset by using describe() function.

IV. Visualization

The data will be visualized into several plots below in order to have a further approach to the data.

A. Correlation Matrix

The correlation heatmap reveals several relationships between the features in the dataset. Here are some key insights:

1) **High Positive Correlation:**
   - Heating_Load and Cooling_Load have a strong positive correlation (0.98), suggesting that buildings with high heating loads also tend to have high cooling loads.
   - Overall_Height and Heating_Load (0.89) as well as Overall_Height and Cooling_Load (0.90) also

show high positive correlations, indicating that taller buildings may require more heating and cooling energy.

2) **High Negative Correlation:**
   - Relative Compactness has a strong negative correlation with Surface Area (-0.99), suggesting that as compactness increases, the surface area tends to decrease significantly.
   - Roof Area and Overall Height have a strong negative correlation (-0.97), meaning that as building height increases, roof area tends to decrease (possibly indicating taller structures with smaller roof areas).
   - Roof Area and Heating Load as well as Roof Area and Cooling Load both show strong negative correlations (-0.86), implying that buildings with larger roof areas may require less heating and cooling.
3) **Weak or No Correlation:**
   - Orientation has very low or near-zero correlation with most other features, indicating that building orientation has little influence on the other variables in this dataset.
   - Glazing_Area and Glazing_Area_Distribution also show very low correlations with other features, suggesting that window/glazing specifications may not significantly impact the heating or cooling loads.
4) **Interpretation for Energy Loads:**
   - Features like Overall_Height, Roof_Area, and Surface_Area appear to have substantial influence on Heating_Load and Cooling_Load. This insight could help in designing energy-efficient buildings by optimizing these features.

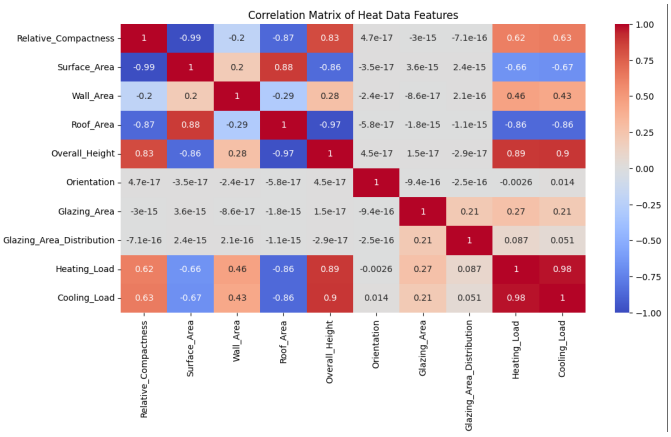

Fig. 2. Heatmap highlights the relationships among various features and suggests which features are most impactful on heating and cooling loads.

B. Factors affecting heat load

In this subsection, we will discuss the main features that affect the heating load and the cooling load, which have a high possible / negative correlation with the two above.

1) **Surface Area and Heating Load**: In Figure 3, we can observe a trend where lower Surface Area values are associated with higher Heating Load values, while higher Surface Area values are associated with lower Heating Load values. This suggests an inverse relationship, where increasing the surface area may decrease the heating load needed, possibly due to improved thermal efficiency or insulation

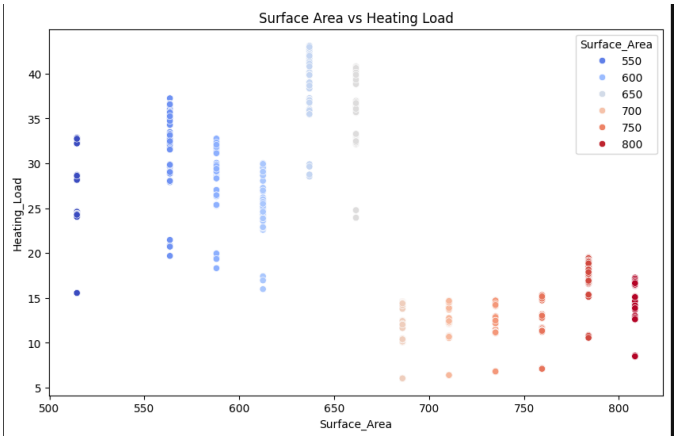

Fig. 3. scatter plot visualizes the relationship between Surface Area and Heating Load.

2) **Heating and Cooling Load by Overall Height**: In Figure 4 taller structures require more heating and cooling energy, possibly due to increased surface area exposed to external temperature variations.

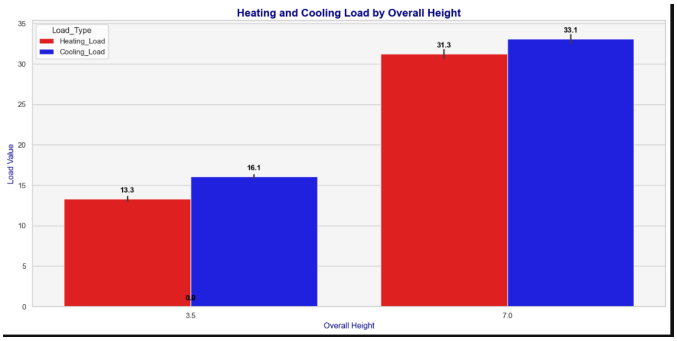

Fig. 4. scatter plot visualizes the relationship between Surface Area and Heating Load.

V. Performance Metrics

The primary performance metrics used to evaluate the predictive accuracy of each model include:

- **Mean Squared Error (MSE):** MSE measures the mean squared difference between actual and predicted values, calculated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

- **R-Squared (R^2 Score):** The R^2 score measures the proportion of the variance in the dependent variable that is predictable from the independent variables, defined as:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{3}$$

VI. Applying Models

A. Multiple Linear Regression

The regression model equation is:

$$y1 = \beta_0 + \beta_1 \cdot X1 + \beta_2 \cdot X2 + \ldots + \beta_8 \cdot X8 + \epsilon \tag{4}$$

c

B. Random Forest Regression

Random Forest Regression was applied to predict heating and cooling loads based on a set of building features. This model uses multiple decision trees and averages their predictions to capture complex relationships.The prediction is calculated as follows:

$$\hat{y} = \frac{1}{n} \sum_{i=1}^{n} T_i(x) \tag{5}$$

Where:

- $\hat{y}$ is the predicted value for input x.
- $T_i(x)$ is the prediction made by the i^{th} tree.
- n is the total number of trees in the ensemble.

References

C. Support Vector Regression (SVR)

Support Vector Regression (SVR) was applied using a radial basis function (RBF) kernel.

VII. Results and Model Comparison

The table below shows the results of MSE and R^2 Score for each model:

Model	MSE (Heating Load)	R^2 (Heating Load)	MSE (Cooling Load)	R^2 (Cooling Load)
Multiple Linear Regression	9.153	0.912	9.893	0.893
Random Forest	0.241	0.997	2.93	0.968
Support Vector Regression (SVR)	0.566	0.994	1.781	0.981

Fig. 5. Comparison of Model Performance

Based on the model performance, we observed that:

- Random Forest outperformed the other models in both heating and cooling load predictions, achieving the lowest MSEs and the highest R^2 scores. This indicates that Random Forest has superior accuracy and predictive power for this dataset.
- Support Vector Regressor (SVR) also performed very well, with slightly higher MSE values but still strong R^2 scores, especially in cooling load prediction. It is a strong alternative to Random Forest.
- Linear Regression had the lowest performance, with higher MSEs and lower R^2 scores for both heating and cooling loads, suggesting that it may not be as suitable for this dataset.

Summary:

- Best Model for Heating Load: Random Forest
- Best Model for Cooling Load: Support Vector Regressor (slightly better MSE than Random Forest)
- Best Model for predicting: Random Forest

VIII. Conclusion

This study confirms that key architectural factors can serve as reliable predictors of thermal efficiency in residential buildings. Random Forest achieved the best performance, followed by SVR and Multiple Linear Regression.

References

[1] Pearson, K. Notes on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

[2] Hastie, T., Tibshirani, R., & Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Springer*, 2009.

[3] Deisenroth, M. P., Faisal, A. A., & Ong, C. S. Mathematics for Machine Learning. *Cambridge University Press*, 2020, p. 291.

[4] Iyoboyi, S. A., & Jaji, M. S. The coefficient of determination R^2 is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Journal of Applied Statistics*, 2021.

[5] Breiman, L. Random Forests. *Machine Learning*, 45(1), 5–32, 2001.