A Regression Model to Predict Heat and Cooling Loads

Alberto Arath Figueroa Salomon

December 13, 2024

1 Introduction

Housing energy requirements are becoming increasingly demanding as we transition from mechanical interfaces to digital ones, and as new electric car power demands emerge. One crucial aspect of energy consumption is heat and cooling loads, which can account for up to 50 percent of total energy usage. This model aims to identify the features that make a house more energy-efficient, as cooling and heating loads are influenced by the thermal capacity of the homes.

2 Data

Data to be analized is based on public repository [Cho22].

2.1 Instance Composition

The dataset is composed of labeled data with 707 rows and 64 features. The training data has a synthetic origin, so there is no need for data preparation, allowing us to focus solely on the application of machine learning techniques.

2.2 Predictor Set

- x_1: **Relative Compactness (RC)**: Relative compactness is a measure of how compact the building is. It can be represented as:

$$RC = \frac{\text{Volume of the building}}{\text{Minimum envelope surface area for that volume}}$$

- x_2: **Surface Area (SA)**: The total surface area of the building:

$$SA = \text{Total Area of all external surfaces (walls, roof, floor, etc.)}$$

- x_3: **Wall Area (WA)**: The area covered by the building's walls:

$$WA = \text{Height of the wall} \times \text{Width of the wall}$$

- x_4: **Roof Area (RA)**: The area covered by the building's roof:

$$RA = \text{Length of the roof} \times \text{Width of the roof}$$

- x_5: **Overall Height (H)**: The total height of the building:

$$H = \text{Base to Top Height of the building}$$

- x_6: **Orientation (O)**: Orientation describes the direction the building faces:

$$O \in \{\text{North, South, East, West}\}$$

- x_7: **Glazing Area (GA)**: The total area of walls that are made of glass:

$$GA = \text{Sum of all glassed surfaces (windows, panels, etc.)}$$

- x_8: **Glazing Area Distribution (GAD)**: The distribution of glazing area across the building:

$$GAD = \frac{\text{Glazing Area in a particular wall}}{\text{Total Glazing Area}}$$

2.3 Categorical Variable Handling

Orientation (O) is a categorical variable. One-Hot Encoding (OHE) is used to treat this feature in the model.

2.4 Target Variables

- y_1: **Heating Load (HL)**: The total load required to heat the building, expressed as:

$$HL = f(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

- y_2: **Cooling Load (CL)**: The total load required to cool the building, expressed as:

$$CL = f(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

This encoding allows the application of machine learning algorithms such as softmax regression, decision trees, or neural networks to predict the likelihood of each class based on the input features.

3 Exploratory Data Analysis (EDA)

3.1 Correlation Matrix

The feature set shows strong correlations in clusters. A PCA approach might help, even though there aren't that many features.

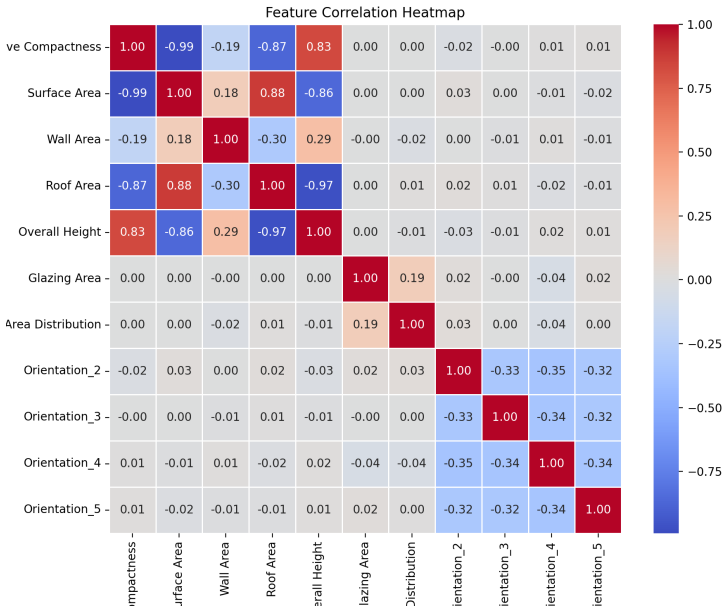

Figure 1: Feature Correlation Heatmap

3.2 Outlier data distribution exploration

No significant observations to take actions.

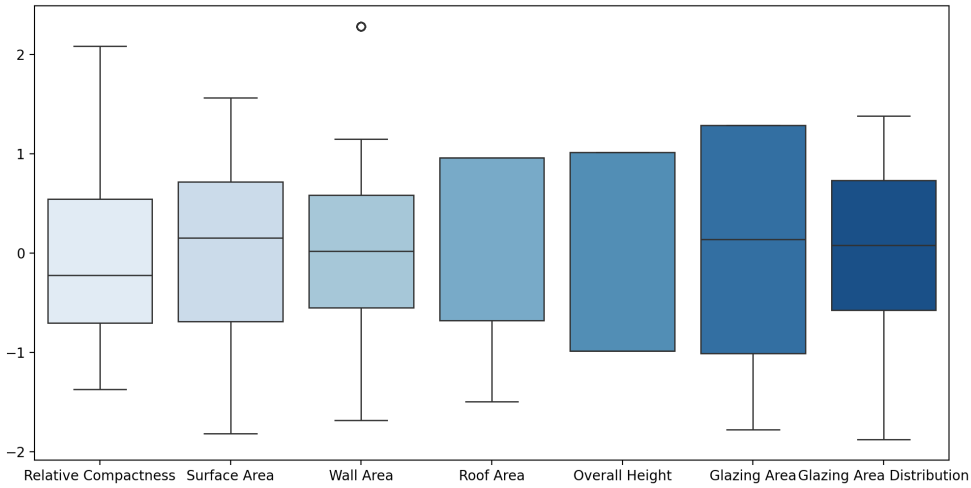

Figure 2: Feature Correlation Heatmap

3.3 Features Correlation Matrix

4 Methodology

In this study, we performed hyperparameter tuning using Grid Search, testing multiple machine learning algorithms to identify the best-performing model. The following algorithms were evaluated:

4.1 Metrics Used

Root Mean Squared Error (RMSE)

The RMSE measures the square root of the average squared differences between predicted and actual values. It is given by:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where:

- y_i: Actual value,
- $\hat{y}_i$: Predicted value,
- n: Number of samples.

R-Squared (R^2)

The R-Squared score measures the proportion of variance explained by the model. It is defined as:

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

Where:

- y_i: Actual value,
- $\hat{y}_i$: Predicted value,
- $\bar{y}$: Mean of actual values,
- n: Number of samples.

4.2 Pre-processing Algorithms

- **Standard Feature Regularization:** Standardizes features by centering them at zero and scaling to unit variance:

$$z = \frac{x - \mu}{\sigma}$$

  Where:

  - x: Original feature value.
  - μ: Mean of the feature.
  - σ: Standard deviation of the feature.

- **One-Hot Encoding (OHE):** Converts a categorical variable into binary features, one for each class:

$$\text{OHE}(x) = [x_1, x_2, \ldots, x_k], \quad x_i \in \{0, 1\}, \quad \sum_{i=1}^{k} x_i = 1$$

  Where:

  - k: Number of unique classes.
  - x_i: Binary variable indicating presence (1) or absence (0) of class i.

- **Principal Component Analysis (PCA):** PCA reduces dimensionality by transforming data into a new set of axes (principal components). The optimal number of components (axes) can be found iteratively using Grid Search with Cross-Validation (2-folds):

  - Each fold evaluates the cross-validation score.
  - Grid Search determines the number of principal components that maximize the model's performance.

4.3 Linear Regression Model

If there aren't regularization techniques and the instance set is not too large, the solver will use the normal equation to compute weights, unless specified otherwise.

4.4 Hyperparameter Tuning

Grid search was used to determine the best number of axes in PCA. This is mainly for the sake of study, as no significant improvement is expected.

Table 1: Hyperparameter Combinations and Corresponding Test Scores with 2 K-folds

Hyperparameter Combination	**Mean Test Score**	**Rank Test Score**
Preprocessing num pipeline PCA components = 5	0.8945	1
Preprocessing num pipeline PCA components = 3	0.8384	2
Preprocessing num pipeline PCA components = 1	0.6438	3

5 Results and Discussion

The results are promising. The model performs reasonably well with a 95 percent variance acceptance.

5.1 Model Benchmark

Table 2: Training and Test Set Performance Metrics

Metric	Training Set	Test Set
RMSE	3.0672	3.1580
R^2	0.9481	0.9506

5.2 Feature Relevance to Determine Heating Efficiency

The most important PCA component is: PC2. The most important feature is: Wall Area. The worst feature is: Relative Compactness.

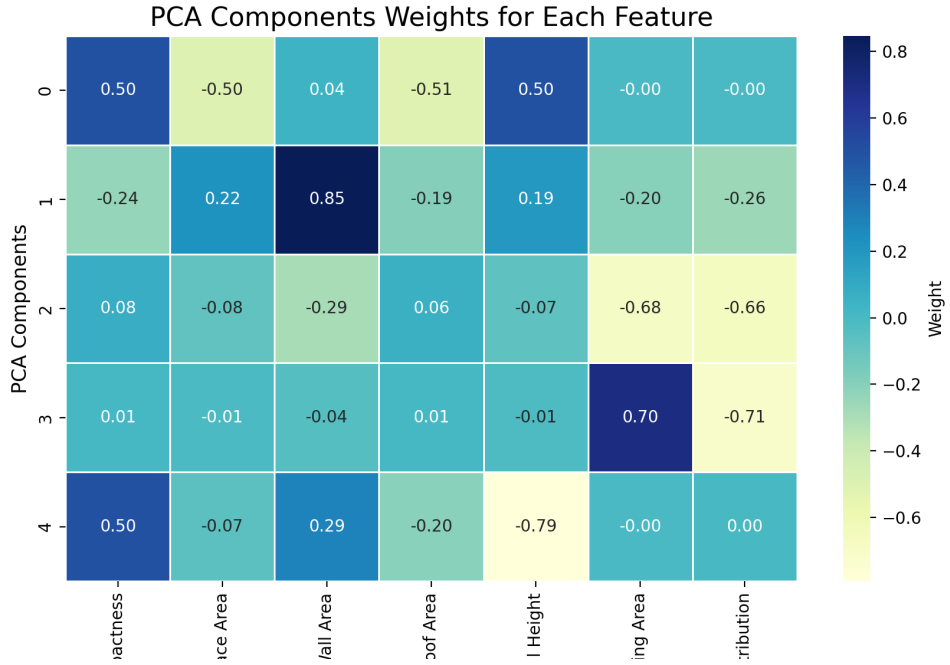

Figure 3: Features Correlation Heatmap

The wall area is the strongest feature, which is something that it does make sense which official data for heating load, as surface increases so the heat exchange surface. It is on the eyes of the study a reasonable conclusion.
However, the principal components themselves don't provide a clear interpretation of the properties along each axis. If additional features were included in the analysis, I would expect the new axes to become more closely correlated with the wall area feature.

6 Conclusion and Further Analysis

The regression model effectively explains both cooling and heating efficiency, demonstrating a strong correlation with the non-synthetic dataset. Although the model is relatively simple, it shows promising results. However, several improvements can be made:
- We analyzed an averaged result between cooling load and heating load; however, it might be valuable to perform separate analyses for each target variable individually. - There is a wealth of data available

on building construction parameters. By scaling the model, we could incorporate additional features such as geographic location, building materials, and city-specific heat signatures.

A Jupyter notebook

Code used for this analysis [Sal24].

References

[Cho22] Ujjwal Chowdhury. Energy efficiency dataset discussion, 2022. Accessed: June 7, 2024.

[Sal24] Alberto Arath Figueroa Salomon. Machine learning regression project. `https://github.com/AlbertoArath/IA/tree/main/MachineLearning/Projects/Regression_Report`, 2024. Accessed: 2024-12-13.