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 switch from mechanical interfaces to digital ones, new electric car power demands. Once crucial aspect of energy consumption are heat and cooling loads which can take up 50 percent of the power intake.
This model looks to find the features that make a house more energy efficency as cooling and heat loads depend on the thermal capacity of the homes

2 Data

2.1 Instance composition

Data set is composed by labeled data, 707 rows and 64 features. Training data has a synthetic origin so there is really no need for data preparation which is good as we can focus solely in 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 categorical variable, One Hot Encoder column insetion is being used to treat this feature into 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 To 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

Feature set shows a strong correlation in clusters a PCA aproach might work help even though there aren't that many features.

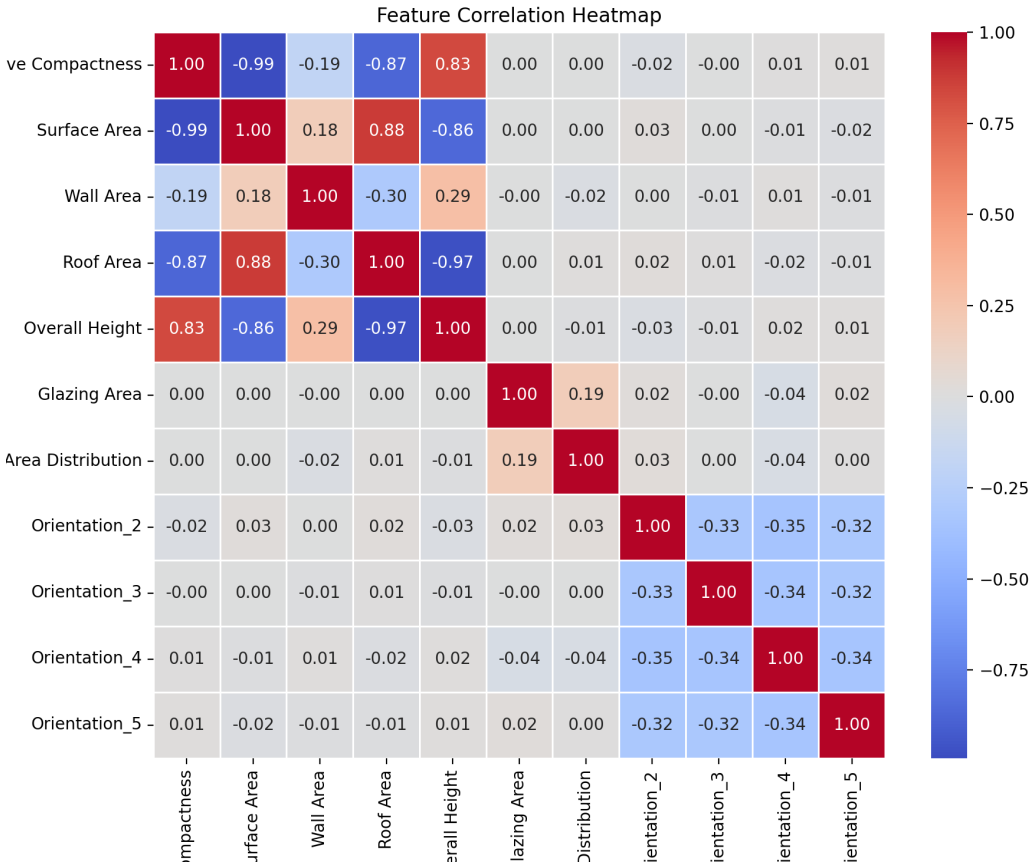

Figure 1: Features Correlation Heat marginparwidth

3.2 Features correaltion 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 Metric 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 ($\mathbf{R^2}$)

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 instance set is not too large, solver will use will use normal equation if not specified. Normal equation is used to compute weights.

4.4 Hyperparameter Tuning

Grid search was used to determine best number of axis in PCA, this is more for the sake of study as no much 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

Results are not that exciting model performs reasonable well for a 95 percent variance acceptance

Table 2: Training and Test Set Performance Metrics

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

Scikit-learn has been widely used for machine learning tasks [**?**]. Vector-borne diseases are a significant challenge[**?**].

A Additional tables