DATA 1030: Final Project Report

Using Machine Learning to Estimate Energy Performance of Residential Buildings

Due on Wednesday, December 3, 2019

Doc. Andras Zsom

TTh 1pm-2:20pm

Cheng Zeng

School of Engineering, Brown University

December 2, 2019

github: https://github.com/cengc13/DATA1030_final_project

.

1 Introduction

1.1 Motivation

This final project aims to understand the energy efficiency of residential buildings. Given eight attributes of a given, there are two variables named **heating load** and **cooling load** characterizing the energy efficiency. Since the outputs are continuous variables. It is essentially a regression problem. It is important because it could help to design environmentally friendly buildings from the theoretical perspective. On-site experiments could be intractable, time and effort consuming.

1.2 Dataset

This dataset originates from the UCI energy efficiency. It has **768** simulation points. Each point has ten features, including 8 input features and 2 target features. No missing values exist for this dataset. In Table. 1, following the classical mathematical convention, X denotes input variables whereas Y denotes target variables. To make it more clear, the exact meaning of the two target variables are elaborated.

- heating load: the amount of heat energy that would need to be added to a space to maintain the temperature in an acceptable range.
- cooling load: the amount of heat energy that would need to be removed from a space (cooling) to maintain the temperature in an acceptable range.

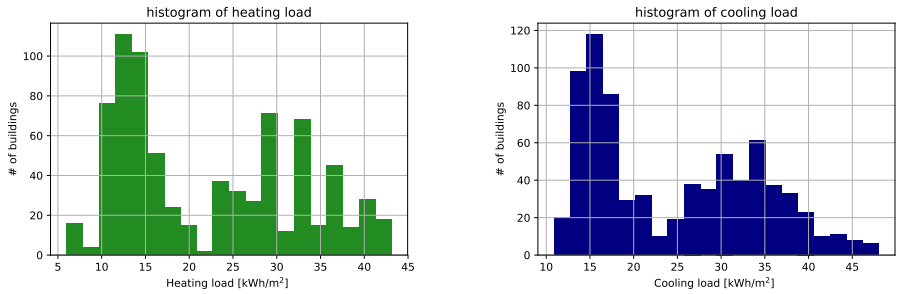
Table 1: Input and output variables.

Math symbols	Input/Output variable	possible values	units/description
$X1$	Relative compactness	12	absolute unit
$X2$	Surface area	12	m^2
$X3$	Wall area	7	m^2
$X4$	Roof area	4	m^2
$X5$	Overall height	2	m
$X6$	Orientation	4	2:North, 3:East, 4:South, 5:West
$X7$	Glazing area	4	0%, 10%, 25%, 40% (of floor area)
$X8$	Glazing area distribution	5	0:Unknow, 1:Uniform, 2:North, 3:East, 4:South, 5:West
$Y1$	Heating load	586	kWh/m^2
$Y2$	Cooling load	636	kWh/m^2

2 Exploratory Data Analysis

We first investigate the properties of two target variables. Fig. 1(a) and Fig. 1(b). As one can see, the two target variables do not follow a simple distribution such as normal distribution. Also, we can see that the two target variables have similar shape of distribution. It indicates that a building with high heating load may also display cooling load.

The Fig. 2(a) and Fig. 2(b) indicate that the relationship between the input variables and target variable is not trivial. It suggests that a simpler learner such as linear regression is not able to describe this complicated behavior, we need more sophisticated model such as random forest regressor to justify this relationship.



(a) Heating load (b) Cooling load

Figure 1: Properties of two target variables

The Fig. 3(a) and Fig. 3(b) indicate that the relationship between the input variables and target variable is not trivial. It suggests that a simpler learner such as linear regression is not able to describe this complicated behavior, we need more sophisticated model such as random forest regressor to justify this relationship.

We compute the correlation matrices for both target variables and rand with the absolute variables with respect to the target variable. In Fig. 4(a) and Fig. 4(b) implies that overall height (X5), roof area (X4), surface area (X2) and compactness (X1) show the highest linear correlation with both target variables.

3 Methods

3.1 Preprocessing

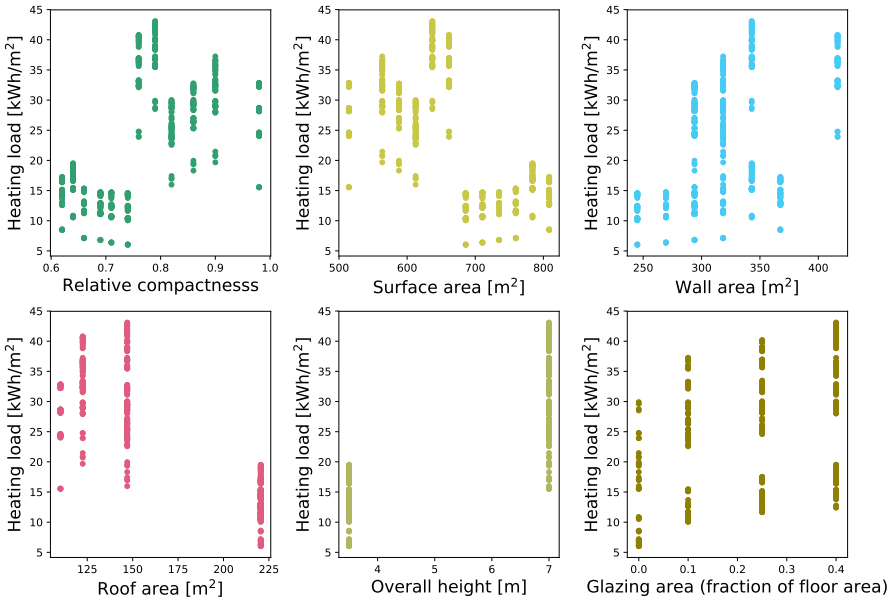
We elaborate on the preprocessing for various types of features

- Six continuous features (X1, X2, X3, X4, X5 and X7): since all the features are well bounded. *MinMaxScaler* is applied.
- Two categorical features (X6 and X8): one-hot encoding is used because the categories cannot be ordered.
- target variables (Y1 and Y2): in terms of the regression problem, it is kept as it is.

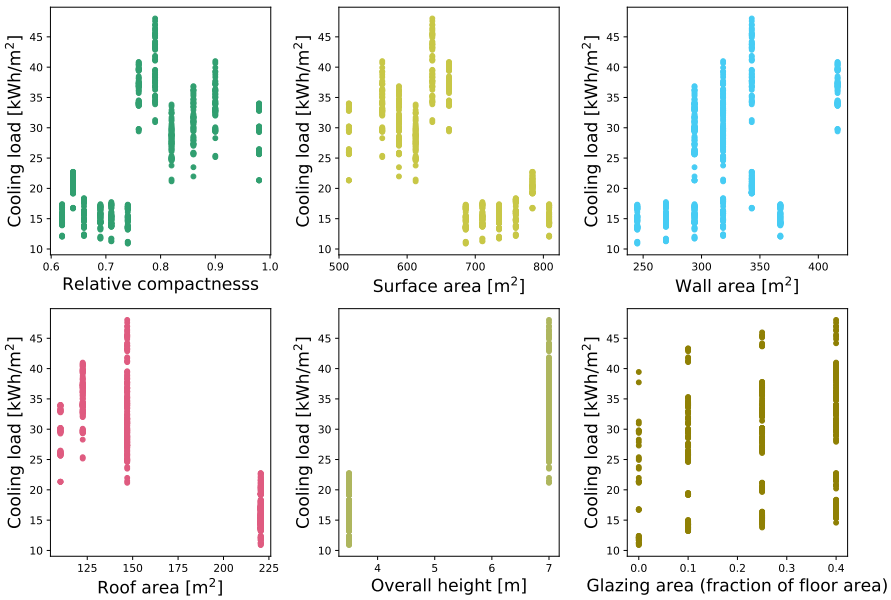
After preprocessing, there are in total 16 input features and 2 output features.

3.2 Machine learning pipeline

The following discussion applies to pipelines of models for both target variables. I used the ordinary k fold cross-validation (CV) because the dataset is iid. Five folds are used. The data splits to three parts (64% train, 16% CV, 20% test). The dataset if shuffled before splitting with given random state to make sure that this process is reproducible. Ten different random states are used to estimate the uncertainties due to data splitting. For non-deterministic methods, the uncertainty for the model itself is evaluated by fixing the random state for data splitting while varying the random state for model for 5 times. I tried four machine learning methods and tuned the hyperparameters using a grid search.



(a) Heating load versus continuous features

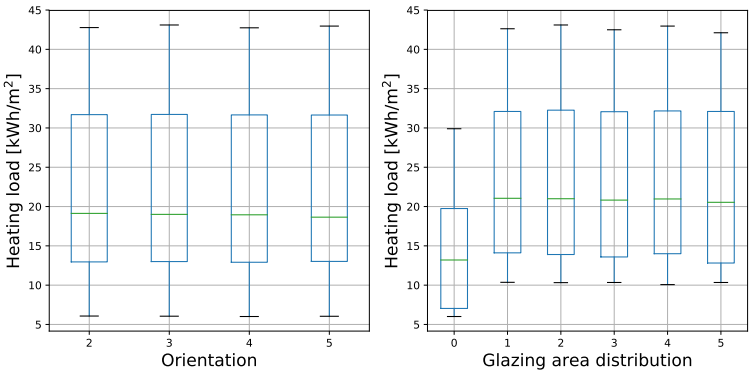


(b) Cooling load versus continuous features

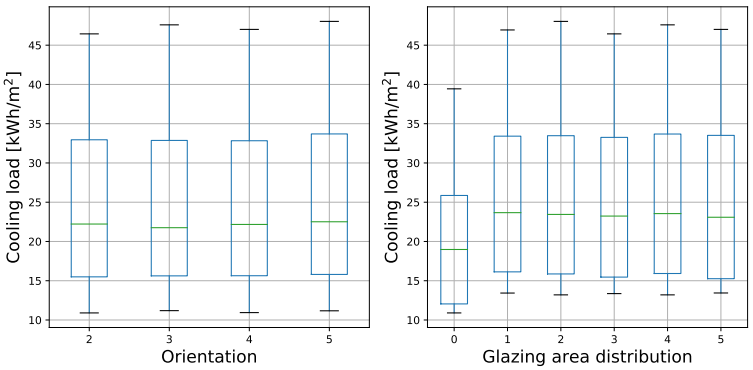
Figure 2: Interaction of two target variables with continuous features

For linear regression model, I used *Lasso* linear model and I tuned the regularization strength (alpha values) in the range between 0.0001 and 100. For random forest, the maximum depth and minimum samples for a split are tuned. I tried the range of 2 to 20 for maximum depth and the range of 2 to 12 for minimum samples for a split. For Support vector regression, I tuned the C and gamma, trying the range of 1 to 1000, and 0.0001 to 100, respectively. For multilayer perceptron regression, I tuned the regularization coefficients (alpha) in the range of 0.0001 to 0.01, and the number of nodes in hidden layers (I tied 10 and 15 nodes).

For the metrics, I used both mean squared error (MSE) and R2 score to evaluate the model performance. MSE is the typical metric used for evaluating a regression model. R2 score is useful when it comes to evaluating

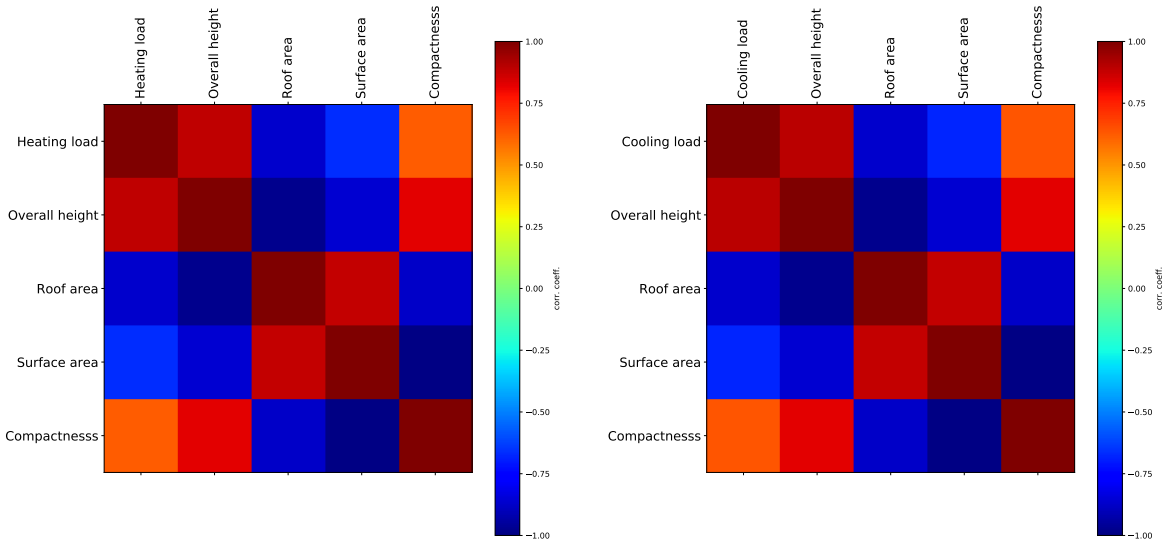


(a) Heating load versus categorical features



(b) Cooling load versus categorical features

Figure 3: Interaction of two target variables with categorical features



(a) Correlation matrix for heating load

(b) Correlation matrix for cooling load

Figure 4: Correlation matrices

the feature importance, especially in comparison of the baseline model whose R2 score is supposed to be 0.

3.3 Results

The R2 score given by a baseline model that predicts the mean of y is 0. In table. 2 and table. 3, we show the MSE and R2 score of prediction on the testing set for heating load and cooling load respectively. Values following $\pm$ indicate the uncertainties due to data splitting. Values in the bracket are the uncertainties due to non-deterministic models, for which only results using MSE metric are evaluated.

Table 2: The test MSE and R2 score for heating load (Y1)

Methods	Test MSE	R2 score
Lasso Linear Model	7.938 ± 0.658	0.923 ± 0.005
Random Forest	$0.283 \pm 0.037(0.009)$	0.997 ± 0.0003
Support Vector Machine	6.650 ± 0.693	0.936 ± 0.004
Multilayer Perceptron Regression	$0.344 \pm 0.071(3.541)$	0.997 ± 0.001

Table 3: The test MSE and R2 score for cooling load (Y2)

Methods	Test MSE	R2 score
Lasso Linear Model	10.044 ± 1.183	0.890 ± 0.007
Random Forest	$3.317 \pm 0.426(0.233)$	0.964 ± 0.003
Support Vector Machine	10.082 ± 1.625	0.890 ± 0.012
Multilayer Perceptron Regression	$2.386 \pm 0.625(2.836)$	0.997 ± 0.001

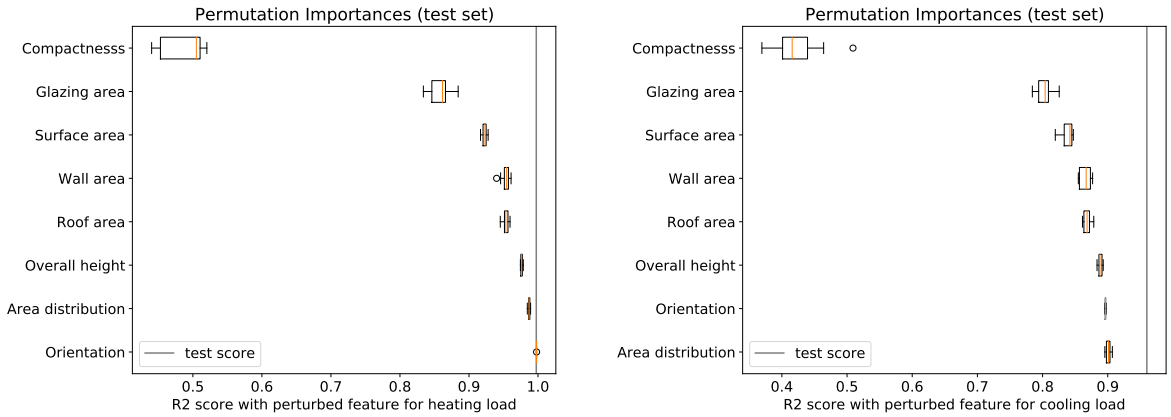
Among all the models studied, Random Forest is the best one for target variable of heating load (Y1), with MSE of R2 score as 0.283 and 0.997, respectively. For cooling load, multilayer perceptron regression (neural network) is the best model, with MSE and R2 score as 2.386 and 0.997, respectively. It is consistent with the finding from Fig. 2(b) that a complicated learner is needed to justify the curvy correlation between input variables and output variables. Another interesting observation is that heating load can be estimated much more accurately that cooling load, which might suggest that some input variables are more strongly associated the heating load. In terms of the uncertainties, in brief the more accurate the model is, the lower the uncertainty due to data splitting will be. However, for the neural network regressor, the uncertainty due to the model itself is very large compared to the other non-deterministic model, Random Forest.

3.4 Global feature importance

We note that there is a large model uncertainty for neural network regression. For the feature importance analysis, we will focus on the random forest model, which gives more stable and consistent performance. We will use R2 score as the metric for feature importance evaluation.

We determine the global feature importance by a permutation test. We found that relative compactness is the most important feature for both output variables (see Fig. 5(a) and Fig. 5(b)). It makes sense because it defines the surface-to-volume of the building, which is an important properties for heat transfer. Interesting, the second most important feature is the glazing area (X7), that does not show up in the simple linear correlation analysis. From engineering perspective, it is intuitively understood that glazing area is of great importance because the amount of glazing is associated with the heat adsorbed from the sun. It implies that statistical learning has potential in mining obvious pattern behind the data. The results fit into a human/academic context because it agrees with our common sense, and it is well-acknowledged in scientific community that relative compactness is an important factor related to the energy efficiency of buildings. I also studies the global feature importance using *Shapley*, the results are shown in Fig. 6(a) and Fig. 6(b),

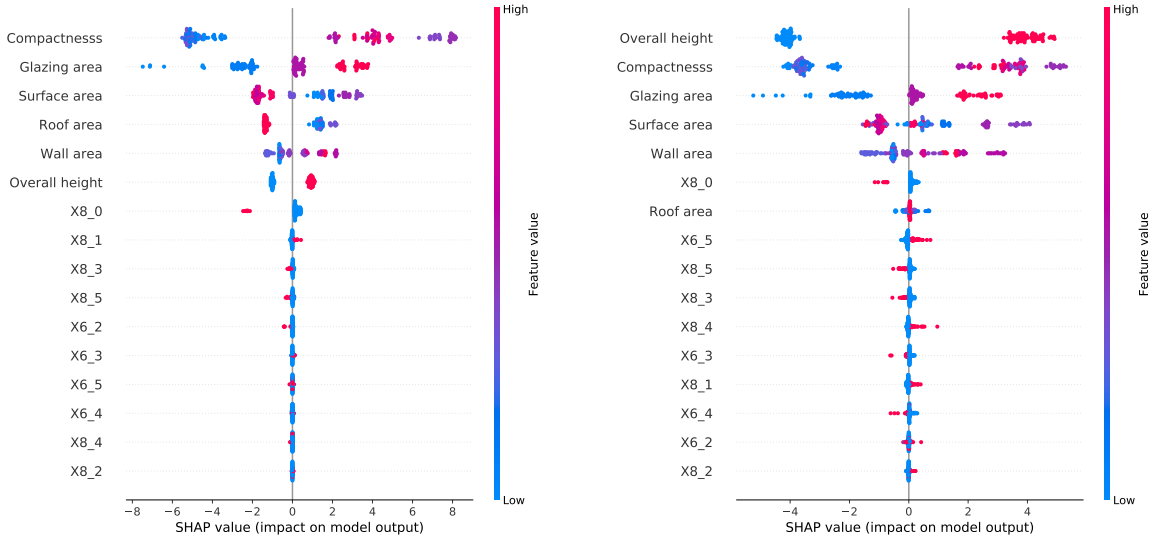
and it is consistent with permutation test although the order of importance of features for heating load is now slightly different than that for cooling load.



(a) Permutation test for heating load

(b) Permutation test for cooling load

Figure 5: Global feature importance: permutation test



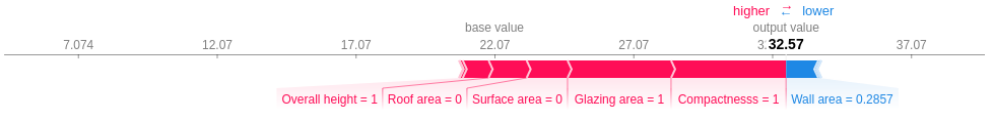
(a) Shap global feature importance for heating load

(b) Shap global feature importance for cooling load

Figure 6: Global feature importance: SHAP method

3.5 Local feature importance

With the help of *SHapley*, I studied the the local feature importance for an example point whose corresponding heating load and cooling load are 32.57 and 33.71, way above the base value of 22.07. It shows that high compactness and glazing area contributes the most to a high heating load whereas high overall height and compactness push the cooling load to the right.



(a) Shap local feature importance for heating load

(b) Shap local feature importance for cooling load

Figure 7: Local feature importance using shap

4 Outlook

We could try more other models to improve the performance, such as gradient boosting. We might have missed some important features for the prediction of cooling load, which might be the reason for the poor model performance on it. More formal tests need to be carried out to provide more insights. Also, we can try to tune more parameters, such as activation functions and learning rates for a neural network model.

5 References

1. The dataset is from UCI Machine Learning Repository
2. A. Tsanas, A. Xifara: 'Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools', Energy and Buildings, Vol. 49, pp. 560-567, 2012.