Estimation of Energy Performance of Buildings Using Machine Learning Tools

Rasoul Rashidifar, Frank F. Chen
Department of Mechanical Engineering, University of Texas at San Antonio, One UTSA Circle San Antonio, San Antonio, TX 78249, USA

Abstract

In this project, it is developed a machine learning framework to study the effect of eight input variables on two output variables, namely heating load (HL) and cooling load (CL), of residential buildings. The association strength of each input variable with each of the output variables using statistical analysis tools is investigated, in order to identify the most strongly related input variables. A linear regression model as a baseline model on data is fitted and then in this project is considered in terms of two different aspects, classification and regression. For classification, it is defined a distance between the amount of data as a category and fit logistic regression on data and also investigated the influence of the number of classes on the accuracy of the model. In other hands, for regression, random forest regression is used to fit a model on data and compare a classical linear regression approach against a powerful state of the nonlinear non-parametric method, random forests, to estimate HL and CL. Extensive simulations on 768 diverse residential buildings show that we can estimate HL and CL with mean square error and R-square. The results of this study support the feasibility of using machine learning tools to estimate building parameters as a convenient and accurate approach.

Keywords

Energy Performance of building, Machine Learning, Regression and Classification Method, Random Forest.

1. Introduction

There has been a considerable body of research on the topic of the energy performance of buildings (EPB) recently due to growing concerns about energy waste and its perennial adverse impact on the environment [1], [2]. Building energy simulation tools are currently widely used to analyze or forecast building energy consumption, in order to facilitate the design and operation of energy efficient buildings since practice has shown that the results of the simulations can often accurately reflect actual measurements [3]. Nowadays, there are a lot of research about using modern technology in residential building, Piezo-sensors are one of them that widely use in buildings [4]. Also, some automatic devices (robotics) are another tools that use in residential building to decrease consumption of energy [5]. Using advanced dedicated building energy simulation software may provide reliable solutions to estimate the impact of building design alternatives; however this process can be very time-consuming and requires user-expertise in a particular program. Hence, in practice many researchers rely on machine learning tools to study the effect of various building parameters (e.g. compactness) on some variables of interest (e.g. energy) because this is easier and faster if a database of the required ranges of variables is available [6]. Various machine learning techniques such as polynomial regression [7], support vector machines (SVM) [8] artificial neural networks (ANN) [9] and decision trees [2] have been explored to predict various quantities of interest in the context of EPB. Machine learning tools have also been explicitly used in predicting HL and CL [6]. In this project, the effect of eight input variables to determine the output variables (HL and CL) of the residential building is investigated. In this case, a classical linear regression as a baseline method and two different classification and regression models is considered. In this study, it is fitted two model logistic regression and random forest on data and investigated and discussed results which are obtained from them.

2. Data

In this study, data are included 8 variables and 2 outputs [10]. This data is generated in 12 building where each building form is composed of 18 elements. The simulation assumes that the buildings are in Athens, Greece, residential with seven persons. All the buildings have the same volume, which is 771.75 m3, but different surface areas and dimensions. The materials used for each of the 18 elements are the same for all building forms. In these buildings, it has been used three types of glazing areas, which are expressed as percentages of the floor area: 10%, 25%, and 40%. Furthermore, five different distribution scenarios for each glazing area were simulated: 1) uniform: with 25% glazing on each side, 2) north: 55% on the north side and 15% on each of the other sides, 3) east: 55% on the east side and 15% on each of the other sides, 4) south: 55% on the south side and 15% on each of the other sides, and 5) west: 55% on the west side and 15% on each of the other sides [6]. In addition, some samples has no glazing areas. Finally, all shapes were rotated to face the four cardinal points. Also, for each of the 768 buildings it has been recorded heating and cooling load [6]. The heating load (HL) is the amount of heat energy that would need to be added to a space to maintain the temperature in an acceptable range and the cooling load (CL) is the amount of heat energy that would need to be removed from a space (cooling) to maintain the temperature in an acceptable

range [11]. Table 1 summarizes the input variables and the output variables in this study, introduces the mathematical representation for each variable, and indicates the number of possible values [6].

Table1: Input and output variables - units and number of possible value

Data	Description (unit)	Number of possible value
X1	Relative Compactness - No units	12
X2	Surface Area - m^2	12
X3	Wall Area - m^2	7
X4	Roof Area - m^2	4
X5	Height - m	2
X6	Orientation - 2:North, 3:East, 4:South, 5:West - No units	4
X7	Glazing Area - 0%, 10%, 25%, 40% (of floor area) - No units	4
X8	Glazing Variations - 1:Uniform, 2:North, 3:East, 4:South, 5:West	6
Y1	Heating Load - kWh/m^2	586
Y2	Cooling Load - kWh/m^2	636

3. Methodology

In this section, I summarize the problem and describe the baseline and a proposed model to solve this problem. According to the points mentioned above, in this project the purpose is estimation of two targets including HL and CL by using 8 variables that are shown in table 1. This problem is considered in 2 aspects separately classification and regression model.

3.1. Overview of Model

By looking at the data and relationship between them it is clear that in this special project I have some classified data that I might use classification method, logistic regression, on them. In figure 1, the histogram plot from data is shown.

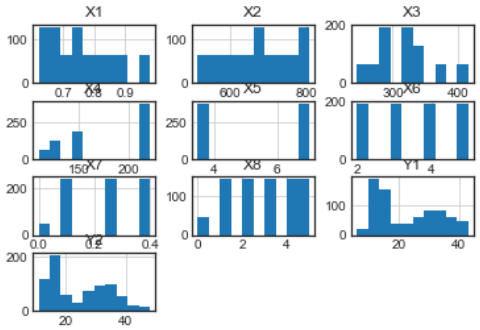


Figure 1: histogram plot from all variables and output

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Logistic regression transforms its output using the logistic sigmoid function (equation 1) to return a probability value. It is a widely used technique because it is very efficient, does not require too many computational resources, it's highly interpretable, it doesn't require input features to be scaled, it doesn't require any tuning, it's easy to regularize, and it outputs well-calibrated predicted probabilities. These might be some reasons that I use this model in my project.

$$f(x) = \frac{1}{1+e^{-z}}, \qquad z = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \tag{1}$$

On the other hand, in data which is about energy efficiency, there is a wide range of data. It is correct that data is classified but based on data that has a vast range, it seems that I also can investigate this data on the regression model. In this case, multi-linear regression is an appropriate baseline method. A linear regression model predicts the target as a weighted sum of the feature inputs (equation 2). The linearity of the learned relationship makes the interpretation easy. Linear regression models have long been used by statisticians, computer scientists and other people who tackle quantitative problems.

$$f(x) = y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \tag{2}$$

In the regression method, the decision tree method is a conceptually simple, yet powerful nonlinear method that often provides excellent results. A natural extension of DT is random forests (RF) which is an ensemble machine learning technique capable of performing both regression and classification tasks using multiple decision trees and a statistical technique called bagging. Bagging along with boosting are two of the most popular ensemble techniques which aim to tackle high variance and high bias. RF instead of just averaging the prediction of trees uses random sampling of training observations when building trees and random subsets of features for splitting nodes. In other words, the random forest builds multiple decision trees and merges their predictions together to get a more accurate and stable prediction rather than relying on

individual decision trees [12]. On many problems the performance of random forest is very similar to boosting, and they are simpler to train and tune. As a consequence, random forest are popular, and are implemented in variety of packages [13].

4. Experimentation

This section describes the data exploration and statistical concepts and the machine learning techniques which are used to analyze the data in python.

4.1. Data exploration

The first step in most data analysis applications is the exploration of the statistical properties of the variables. In this case, there is no missing data. In tables 2 and 3, some statistical information about data is shown. One tool for analyzing data has been used is the Spearman rank correlation coefficient.

Table 2: Statistical information about inputs and outputs

	X1	X2	X3	X4	X5	X6	X7	X8	Y1	Y2
count	768	768	768	768	768	768	768	768	768	768
mean	0.764167	671.7083	318.5	176.6042	5.25	3.5	0.234375	2.8125	22.3072	24.58776
Std	0.105777	88.08612	43.62648	45.16595	1.75114	1.118763	0.133221	1.55096	10.0902	9.513306
min	0.62	514.5	245	110.25	3.5	2	0	0	6.01	10.9
max	0.98	808.5	416.5	220.5	7	5	0.4	5	43.1	48.03

The Spearman rank correlation coefficient can characterize general monotonic relationships and lies in the range between -1 and 1, where a negative sign indicates inversely a proportional and positive sign indicates a proportional relationship. For instance, as you can see, X6 has no correlation with other variables and it is almost zero.

Table 3: Spearman Rank Correlation

	X1	X2	X3	X4	X5	X6	X7	X8	Y1	Y2
X1	1	-9.92E-01	-2.04E-01	-8.69E-01	8.28E-01	0	1.28E-17	1.76E-17	0.622272	0.634339
X2	-9.92E-01	1	1.96E-01	8.81E-01	-8.58E-01	0	1.32E-16	-3.56E-16	-0.65812	-0.673
X3	-2.04E-01	1.96E-01	1	-2.92E-01	2.81E-01	0	-7.97E-19	0.00E+00	0.455671	0.427117
X4	-8.69E-01	8.81E-01	-2.92E-01	1	-9.73E-01	0	-1.38E-16	-1.08E-16	-0.86183	-0.86255
X5	8.28E-01	-8.58E-01	2.81E-01	-9.73E-01	1	0	1.86E-18	0.00E+00	0.88943	0.895785
X6	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	1	0.00E+00	0.00E+00	-0.00259	0.01429
X7	1.28E-17	1.32E-16	-7.97E-19	-1.38E-16	1.86E-18	0	1	2.13E-01	0.269842	0.207505
X8	1.76E-17	-3.56E-16	0.00E+00	-1.08E-16	0.00E+00	0	2.13E-01	1	0.087368	0.050525
Y1	6.22E-01	-6.58E-01	4.56E-01	-8.62E-01	8.89E-01	-0.00259	2.70E-01	8.74E-02	1	0.975862
Y2	6.34E-01	-6.73E-01	4.27E-01	-8.63E-01	8.96E-01	0.01429	2.08E-01	5.05E-02	0.975862	1

4.2. Evaluation method

Each model needs to have a criterion to evaluate it. In this project, I have used the machine learning tools that were linear regression, random forest, and logistic regression in order to train some data that is called train data and then use a testing subset, which is used to assess the learner's generalization performance. In this case, for the classification model that uses logistic regression, I use the accuracy of the model which is trained by train data by test data. Also, there are some evaluation methods for the regression model; we record the mean absolute error (MAE), the mean square error (MSE), the mean relative error (MRE) and R^2 for both training and testing subsets (equations 3-5).

$$MAE = \frac{1}{N}\sum_{n=1}^{i}|y_i - \hat{y}_i| \quad (3)$$

$$MSE = \frac{1}{N}\sum_{n=1}^{i}|y_i - \hat{y}_i|^2 \quad (4)$$

$$MRE = \frac{1}{N}\sum_{n=1}^{i}\frac{|y_i - \hat{y}_i|}{y_i} \quad (5)$$

4.3. Model

Generally, the configuration of the model in both classification and regression model is that after import data in python data are divided into 2 sections, train and test that usually 70% is train and 30% is test data. Then on train data, the model is fitted, in this project either classification or regression, and then with using test data, the model would be evaluated. In the following paragraphs, I aptly elaborate on the models that I set up in this project.

4.3.1. Classification using logistic regression

Using a logistic regression model on data is the first step in this project, I prepare and normalize data and then convert them from regression data to classification. In this case, I choose a bean size which is the distance between output values and then I classify them in several classes. It is clear that the bigger the bean size the smaller the number of classes. Another thing that

I consider in this project is choosing data. As mentioned before, data are divided into two parts train and test. For a training model that by using shuffle in python, I select this data randomly. It means in each time that I run code I have different data in train and test data. At the end of the code, I calculate the accuracy of the model by using test data. Bean size is a significant parameter in this method. So, for investigating its impact on the result, I decided to test four bean size and then calculate the accuracy of the model for each one. In table 4, it is shown that each bean size creates how many class and what is its accuracy model for each one.

Table 4: Accuracy of logistic regression with different classification

	Distance between beans	No. class	Accuracy (Y1)	Accuracy (Y2)
1	5	8	0.60	0.61
2	10	4	0.78	0.79
3	15	3	0.92	0.92
4	20	2	0.90	0.90

Table 4 shows that the accuracy of the model is better when there is a bigger bean size (less class). When we look at this result more precise, having less class means we lost some influence of variables on the model. In other words, for instance, when we have bean size equal 20 we have just 2 classes for all outputs however, according to table 1 for Y1 we have 586 possible value. Therefore, in result, this accuracy that we obtain when bean size is bigger is not real and it seems that this model is not reliable because, when we have small bean size accuracy is not acceptable and is almost 60% in the other hand when we have bigger bean size accuracy is better and even is near to 90% but as I mentioned above it is not reliable.

4.3.2. Regression using linear regression and random forest

The second approach in this project is a regression. In this case, I use linear regression as the baseline method and random forest regression as the final method. I divide data like what I did in logistic regression (split data and normalize them) and then I fit a linear regression on data. This model calculates coefficient and intercept a linear equation that would pass from data and also as I mentioned in the evaluation method section it would be obtained. Table 5 shows the coefficients that are obtained from this model.

Table 5: Coefficient and intercept for outputs in linear regression

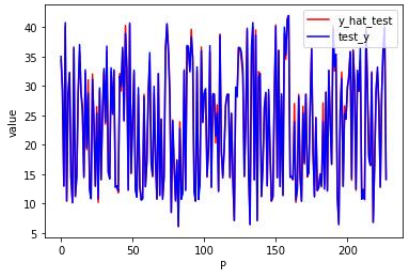
coefficients	B0	B1	B2	B3	B4	B5	B6	B7	B8
For Y1	81.42	-64.8	-2.04E+12	2.04E+12	4.09E+12	4.25	-2.42E-02	2.38E+01	2.43E-01
For Y2	94.43	-70.99	-1.31E+12	1.31E+12	2.62E+12	4.45E+00	9.29E-02	1.73E+01	5.12E-02

In the final stage of my project, I fit a random forest regression on data in python and investigate its results and compare it with the baseline model. Similar to previous models, after preparing data and normalizing that I set up a part to select train data randomly. It means that instead of selecting some certain data as train and others as a test data, python chooses data randomly it helps us to have an appropriate model with more precise results. In the model, the importance of each variables for each one of outputs is obtained. These results are illustrated in table 7.

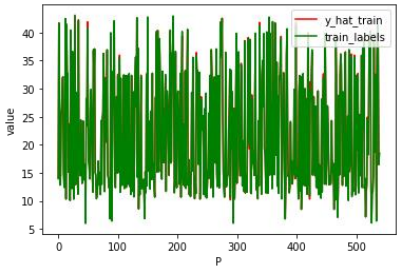
Table 6: Importance of Variables

coefficients	X1	X2	X3	X4	X5	X6	X7	X8
For Y1	0.234	0.23	0.033	0.24	0.175	0.0006	0.071	0.012
For Y2	0.259	0.158	0.042	0.06	0.413	0.011	0.041	0.014

In figure 2, there are diagrams of test and train data for both outputs (y1 and y2) versus predicted outputs that are obtained from the model. It is clear that the model is almost an appropriate model and it seems that errors are very low.



Test data(y1) Vs. y1 prediction



Train data(y1) Vs. y1 perdition

Test data(y2) Vs. y2 perdition Train data(y2) Vs. y2 prediction

Figure 2: Diagrams of test and train (y1 and y2) data versus y (y1 and y2) prediction

Another parameter that I consider in this model is the number of trees. However the number of tree helps the model to gain better results, it needs to have more computations and it is clear that when we have any numbers of data it would take more time. Therefore, in this case, users are looking for an optimized parameter that has both precise results and a suitable time. Therefore, in this case, users are looking for an optimized parameter that has both precise each other. There are results in table 8.

Table 7: evaluation criteria in random forest based on number of trees

	R-square		**MAE**		**MSE**		**RMSE**	
number of tree	Y1	Y2	Y1	Y2	Y1	Y2	Y1	Y2
10	0.99	0.96	0.36	1.04	0.23	2.97	0.48	1.72
50	0.99	0.97	0.36	0.93	0.29	2.6	0.54	1.61
100	0.99	0.97	0.34	1	0.26	2.68	0.51	1.63
200	0.99	0.96	0.35	1.04	0.25	2.98	0.5	1.72

In table 8 some errors as evaluation methods are compared each other for 4 different the number of trees and it clear that there is no difference between the value of errors and R-square when the number of trees are high or low. So, it seems that this parameter is not significant and 10 trees are chosen. Also, in this method we have 2 options for choosing criteria MSE and MAE, in this project, MSE is a parameter that is used in coding. According to table 7, it is clear that X6 has least importance and like linear regression, in the model, X6 is dropped and then model would be rerun. In table 9 the results that are obtained from new model is shown. This model is done with 10 trees for random forest.

Table 8: Results of RF after dropping X6

Drop X6	R-square		MAE		MSE		RMSE	
number of tree	Y1	Y2	Y1	Y2	Y1	Y2	Y1	Y2
10	0.99	0.95	0.32	1.28	0.21	4.7	0.46	2.16

In one of the research mentioned above [6], they obtain MAE with RF method on the same data and in this project the results are compared with their results. In this paper, they obtained MAE for the RF model as well, table 10 shows my results that are compared with them.

Table 9: MAE in Random Forest

	MAE from [6]	MAE
Y1(HL)	0.51	0.36
Y2(CL)	1.42	1.24

5. Discussion

I have developed a comprehensive framework to study HL and CL using a range of diverse input variables which included compactness, orientation and glazing properties. We demonstrated that we can accurately estimation HL with just 0.36 point error and CL with 1.05 point error these results are almost near to results that I gained in python. These results are shown in Table 7. I explored the statistical relationship between eight variables and two outputs, results illustrate that X5(Height), X1(relative compactness) and X3(wall area) have mostly associated with HL and CL and X4(Roof area), X2(Surface area) inversely associated with them.In this study, we use both classification and regression models. In the classification model, using logistic regression is not an appropriate tool for data and results illustrate that when we have larger bean size accuracy of model increase however it means that we have the smaller class for all data and it seems that this suitable accuracy is not real and reliable. As we know logistic regression is a simple and easy tool to use and interpret data and its result so in this case, it is better that we use a more precise and complex method because according to the data exploration section we have non-linear models and data. There is some suggestion in this case that it seems that SVM is a great model [8]. In the regression model, I fit a multilinear regression as a baseline model however R-square for outputs is roughly 0.85, histogram plots (figure 1) and the correlation between variables give ample evidence that linear techniques are not appropriate for available data in

this application. For this reason, a non-linear model is more suitable for them. I fit the random forest model and also I calculate the importance of variables in this model. Interestingly, the most important variable (glazing area) is not the most correlated with either output variable. From an engineering perspective, it can be intuitively understood that the glazing area is of paramount significance to determine EPB [6]. This is because the amount of glazing determines the heat absorbed in a building due to the sun, and similarly, glazing is a source of heat leakage from the building to the environment [6]. In this project, for fitting a random forest model on data I face 2 parameters, number of trees and criteria. I develop four different number of trees and MSE as criteria. According to results which are illustrated in table 7, since there is no strong difference between values of errors in the various numbers of trees, I choose just 10 for it that would have fewer computations. Based on results obtained in this study, the methodology presented here is very general and strong that could be extended to encompass additional input variables (for instance climate and occupancy) and likely additional outputs.

Reference

[1] European Commission, "Directive 2002/91/EC of the European Parliament and of the Council of 16th December 2002 on the energy performance of buildings," *Off. J. Eur. Communities*, vol. L1/65 – L1/71, Apr. 2003.

[2] Z. Yu, F. Haghighat, B. C. M. Fung, and H. Yoshino, "A decision tree method for building energy demand modeling," *Energy Build.*, vol. 42, no. 10, pp. 1637–1646, Oct. 2010, doi: 10.1016/j.enbuild.2010.04.006.

[3] A. Yezioro, B. Dong, and F. Leite, "An applied artificial intelligence approach towards assessing building performance simulation tools," *Energy Build.*, vol. 40, no. 4, pp. 612–620, Jan. 2008, doi: 10.1016/j.enbuild.2007.04.014.

[4] H. Hoshyarmanesh, N. Ebrahimi, A. Jafari, P. Hoshyarmanesh, M. Kim, and H.-H. Park, "PZT/PZT and PZT/BiT Composite Piezo-Sensors in Aerospace SHM Applications: Photochemical Metal Organic + Infiltration Deposition and Characterization," *Sensors*, vol. 19, no. 1, p. 13, Jan. 2019, doi: 10.3390/s19010013.

[5] N. Ebrahimi, "Modeling, Simulation and Control of a Robotic Arm," engrXiv, preprint, Aug. 2019. doi: 10.31224/osf.io/t8fsr.

[6] A. Tsanas and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools," *Energy Build.*, vol. 49, pp. 560–567, Jun. 2012, doi: 10.1016/j.enbuild.2012.03.003.

[7] T. Catalina, J. Virgone, and E. Blanco, "Development and validation of regression models to predict monthly heating demand for residential buildings," *Energy Build. - ENERG BLDG*, vol. 40, pp. 1825–1832, Dec. 2008, doi: 10.1016/j.enbuild.2008.04.001.

[8] B. Dong, C. Cao, and S. E. Lee, "Applying support vector machines to predict building energy consumption in tropical region," *Scopus*, May 2005, Accessed: Nov. 30, 2019. [Online]. Available: https://scholarbank.nus.edu.sg/handle/10635/113973.

[9] J. Zhang and F. Haghighat, "Development of Artificial Neural Network based heat convection algorithm for thermal simulation of large rectangular cross-sectional area Earth-to-Air Heat Exchangers," *Energy Build.*, vol. 42, pp. 435–440, Apr. 2010, doi: 10.1016/j.enbuild.2009.10.011.

[10] "UCI Machine Learning Repository: Energy efficiency Data Set." http://archive.ics.uci.edu/ml/datasets/Energy+efficiency (accessed Nov. 30, 2019).

[11] William Rudoy and Joseph F.Cuba, "Cooling and Heating Load Calculation Manual'." American society of heating refrigerating and air-conditioning engineers, Inc.

[12] "Random Forest Regression model explained in depth," *GDCoder*, Jun. 04, 2019. https://gdcoder.com/random-forest-regressor-explained-in-depth/ (accessed Nov. 30, 2019).

[13] S. Tibshirani and H. Friedman, "Valerie and Patrick Hastie," p. 764.