University of
St Andrews

CS5014: Machine Learning
Practical 1: Predicting energy performance of residential buildings

Student ID: 180007800

University of St Andrews | FOUNDED 1413

Table of Contents

University of St Andrews | FOUNDED 1413

Introduction

This report describes the derivation process for a regression model which predicts the cooling load and heating load of residential buildings. The regression is performed on a data set consisting of the cooling load, heating load of the residential buildings and a combination of 8 different input parameters like relative compactness, surface area, wall area and roof area and so on. The dataset consists 8 numerical attributes and has a total of 768 samples. The samples were collected from 768 simulated buildings.

In order to allow the reader to understand all the steps of the model derivation the report is structured as follows. In the first section of the report the initial data cleaning is described. After the data cleaning the report looks at the given data by some general described statistics. Furthermore, the data is visualised to get an idea how the data is distributed and correlated. In the following Feature Selection section, the most important features are generated, and a regression model is fitted. In the evaluation section the model is evaluated. Finally, the report summarizes the results and critically discusses them.

Data Cleaning and Creation of Test and Training sets

The initial data set has kept a very high quality. There are completed values for all the parameters and all the parameters have the same number of samples. Therefore, it is not necessary to remove any samples. In addition, the format of all variables is numerical. However, the type of two variables among them were found to “float” by using .dtypes, which were different from others, so the type of them were transferred to “int”.

After the cleaning, the dataset was split up into a training and a test data set. The step plays an important role in evaluating the regression model later on with unknown data. I used sample() function of the panda package to seperate the training set and testing set. 80% of the dataset are set to training data. In addition, the random_state is set to 200 in order to reproduce the results. In the future, the training data is used for training the model and the test data is used for testing.

Data Analysis and Visualisation

The first step in the data analysis is providing some general descriptive statistics (Table 5 in Appendix) for all variables. The mean and ranges of the different variables could provide an idea about how much they vary. Next, it is more important to analyse the independent variables and dependent variables.

The histograms of probability density (Figure 1) present the statistical properties of the input variables. These distributions clearly demonstrated that none of the variables follow the normal distribution.

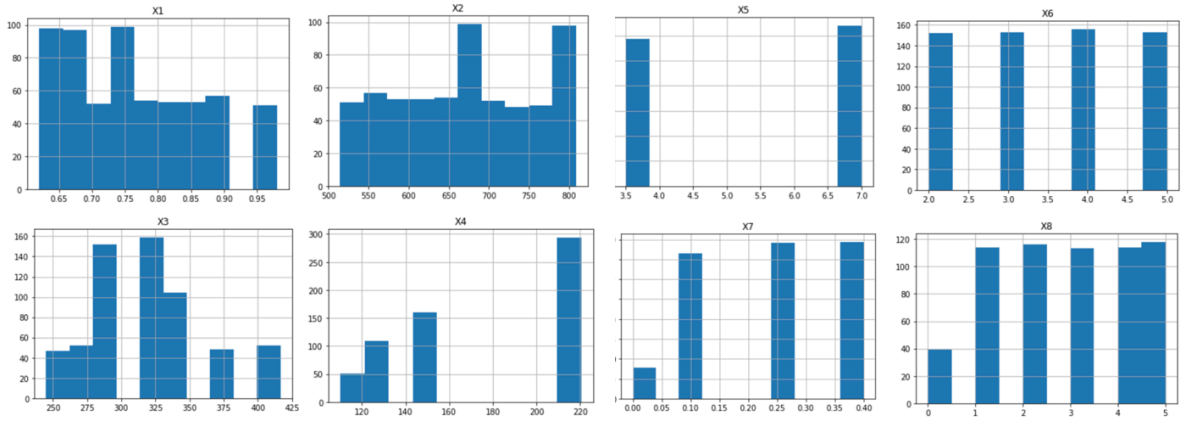

Fig. 1. Histograms demonstrating probability density of the eight input variables.

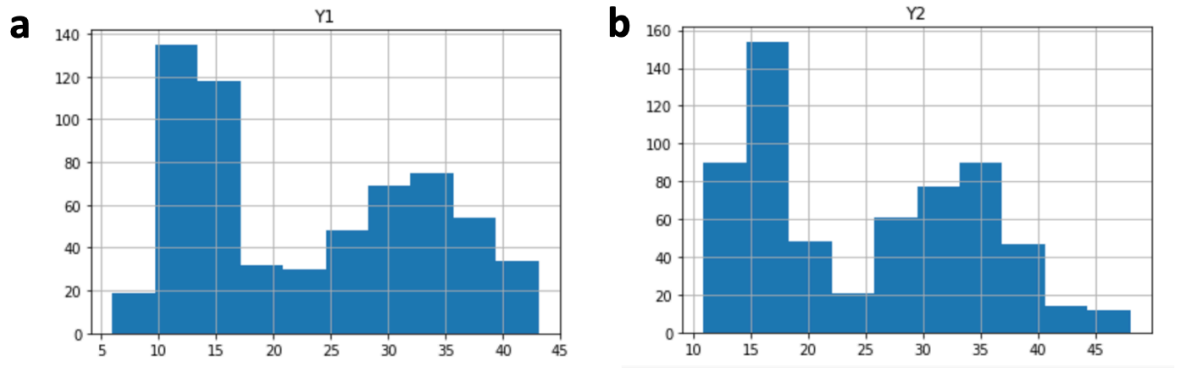

Fig. 2. Histograms demonstrating probability density of (a) heating load, or (b) cooling load.

The descriptive statistics (Table 6 in Appendix) gives an idea that Y1 has a minimum value of 6.01 and a maximum value of 43.1, the mean is 22.778 and the median is 22.795. Since the median is close heavily to mean, the data is nearly distributed evenly on both the right side and the left side of the mean. However, the histogram of Y1 in Figure 2a demonstrates that the number of samples (more than 200) that ranges from 10 to 15 is much more than others (less than 100). Thus, the samples which is between 12.5 to 15 might appear frequently than others. Another descriptive statistic (Table 7 in Appendix) gives an idea that Y2 has a minimum value of 10.9 and a maximum value of 48.03, the mean is 25.035 and the median is 25.135. The distribution is nearly even which is similar with Y1 seen from the descriptive data. However, the histogram of Y2 in Figure 2b presents that more than 200 samples distribute between 15 to 20, and the number of samples in other ranges is less than 100. In addition, the histograms of these

variables show that none of the variables follows Gaussian distribution. Thus, the problem is that whether classical linear regression technique would fail to predict the output variables.

In order to answer the question, the correlation between different variables is analysed to explore the distribution can be explained by either linear correlation or non-linear correlation.

Correlation Analysis

Due to the data is non-Gaussian, the Spearman rank correlation coefficient was used to get an idea how the variables are related based on the association strength of each input variable with each of the two outputs. The range of the Spearman rank correlation coefficient remains between –1 to 1, where 1 means the most perfect positive correlation, -1 means the most perfect negative correlation and 0 means no correlation. The higher correlation between the two variables means the more powerful influence of the two variables. But we could not confirm that a low correlation means a low influence between them, because the relation might be non-linear. Therefore, we could get a rough idea which input variables would have more powerful influence on the prediction of the output variables.

From the Table 1, we can see that the first five input variables appear strong relation with the output variables. In addition, X1 (relative compactness) and X2 (surface area) are shown to be inversely proportional in Table 2 and the reason is that the volume of the simulating buildings is assumed to be constant. Another interesting finding: there is a strong relation between X4 (roof area) with X1 and X2, and a strong relation between X5 (height) with X1 and X2, furthermore the magnitude of the rank correlation coefficient between X4 and X5 is revealed to be –0.972.

Input variable	Y1	Y2
X5	0.861	0.864
X4	-0.799	-0.798
X1	0.626	0.65
X2	-0.626	-0.65
X3	0.47	0.422
X7	0.316	0.281
X8	0.066	0.047
X6	-0.013	0.008

Table 1: Correlation estimated using the Spearman rank correlation coefficient of the eight input variables (X1. . . X8) with heating load (Y1) and cooling load (Y2).

	X1	X2	X3	X4	X5	X6	X7	X8
X1	1.0	-1.0	-0.248	-0.872	0.868	-0.012	0.008	-0.001
X2	-1.0	1.0	0.248	0.872	-0.868	0.012	-0.008	0.001
X3	-0.248	0.248	1.0	-0.193	0.222	-0.012	-0.006	-0.006
X4	-0.872	0.872	-0.193	1.0	-0.93	0.013	-0.004	0.005
X5	0.868	-0.868	0.222	-0.93	1.0	-0.016	-0.001	-0.004
X6	-0.012	0.012	-0.012	0.013	-0.016	1.0	0.021	0.039
X7	0.008	-0.008	-0.006	-0.004	-0.001	0.021	1.0	0.182
X8	-0.001	0.001	-0.006	0.005	-0.004	0.039	0.182	1.0

Table 2: Correlation evaluated using Spearman rank correlations between the eight input variables.

According to the distribution of variables in Figure 1, 2 and the scatter plots provided in Figure 3 in the appendix, we could get an idea that it might be impossible to find a suitable linear model to match the input variables and the output variables. But some correlation values between the input variables and output variables mean probability to predict output variables based on linear regression model. Therefore, I decided to try linear regression firstly, then try random forest in order to find an accurate mapping of the input variables to the output variables.

Feature Selection & Regression Model

Linear Regression

The feature selection is achieved by applying backward stepwise selection approach. The approach starts with a full set of features and deletes predictor which hurts least until the prediction of the model becomes significantly worse. The purpose of the approach is selecting features that provide the best performance. I used a linear regression model to select the features based on the mean squared error as a loss function. Only one of the variables was eliminated each time. Finally, the feature selection approach returned a linear regression model with 5 independent variables (X1, X2, X3, X5, X7). The eliminated variables are predicted to have the lowest linear correlation with the dependent variables so eliminating these variables benefits a more powerful linear regression model.

Table 3 compares the performance on the test data and the training data between the model with selected features (X1, X2, X3, X5, X7) and the models with all features. The experimental results from the five independent variables are the best so far. The results show a slight difference between them shown in Table 3. Therefore, the advantage of the feature selection approach is reducing the model complexity and improving the explanation.

Table 3 presents the MSE, MAE and R^2 of the training set and test set of predicting the output variables. The results demonstrate an accurate predicting result is available directly from the linear regression model.

a

	Training			Testing		
	MSE	MAE	R^2	MSE	MAE	R^2
Linear Model with feature selection	8.919	2.097	0.913	7.505	1.878	0.998
Linear Model with all features	9.044	2.147	0.912	7.536	1.955	0.920

b

	Training			Testing		
	MSE	MAE	R^2	MSE	MAE	R^2
Linear Model with feature selection	10.320	2.70	0.886	9.608	2.134	0.888
Linear Model with all features	10.383	2.365	0.885	9.812	2.270	0.885

Table 3: Training and testing evaluation metrics on (a) heating load, or (b) cooling load for the linear regression models

Random Forest

The non-linear model – random forest regressor was also used to predict the heating load and the cooling load. This model uses the training data to train a series of decision trees and the mean of the results of the individual decision trees is calculated and returned. Not unexpected, random forest regressor presents better predicting results shown in Table 4. In addition, I combined random forest regressor with a feature selection approach, the MSE of test data is as low as 0.290, and r^2 could be as high as 0.998.

a

	Training			Testing		
	MSE	MAE	R^2	MSE	MAE	R^2
Random Forrest Regressor with feature selection	0.223	0.340	0.998	0.174	0.310	0.998
Random Forrest Regressor with all features	0.219	0.335	0.998	0.172	0.304	0.998

b

	Training			Testing		
	MSE	MAE	R^2	MSE	MAE	R^2
Random Forrest Regressor with feature selection	2.648	1.060	0.971	2.250	0.971	0.974
Random Forrest Regressor with all features	2.664	1.062	0.971	2.300	0.923	0.973

Table4: Training and testing evaluation metrics on (a) heating load, or (b) cooling load for the Random Forrest Regressor

Summary & Discussion

At the beginning of the report, the initial dataset was analysed and cleaned, then the cleaned dataset was separated to training data and testing data. Then the variables were analysed by described statistics and histograms of probability density. The histograms showed that none of the variables followed Gaussian distribution and the number of the data within a certain range was much more than others within other ranges. This brought up this problem that the prediction could be achieved by either linear regression or non-linear regression.

Based on the correlation analysis on the Spearman rank correlation coefficient, I learnt that the first five input variables appeared strong correlation with the output variables. And the correlation value gave me an idea that linear regression model might be suitable to predict the output variables, but the predicting performance might not be as good as non-linear regression model. In addition, some of the input variables were strongly correlated with each other which demonstrated that they were not independent. Therefore, I decided to use a feature selection algorithm combined with linear regression and try to use random forest technique to compare performance with linear regression technique.

During the analysis, the predicting results of linear regression suggested that linear regression with all features could be used to predict the output variables and the feature selection algorithm is able to optimise the performance. As expected, random forest achieves superior performance. The reason might be that random forest is more suitable to the complicated relationship between these variables.

University of St Andrews | FOUNDED 1413

Appendix

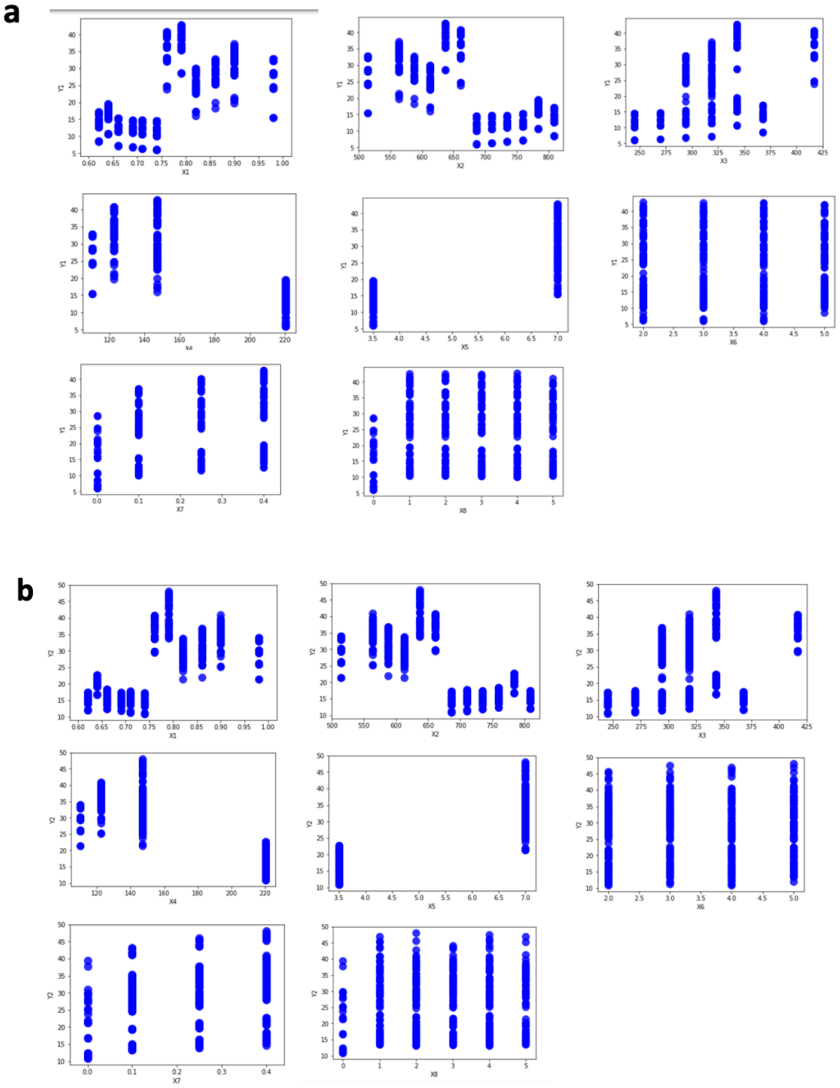

Fig. 3. Scatter plot illustrating visually the relationship between each normalized input variable and the normalized outputs a) the heating load, or b) the cooling load.

	X1	X2	X3	X4	X5	X6	X7	X8	Y1	Y2
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

Table 5: Descriptive statistics for the whole data set

University of St Andrews | FOUNDED 1413

```
count    614.000000
mean      22.778029
std       10.135196
min        6.010000
25%       13.305000
50%       22.795000
75%       32.195000
max       43.100000
Name: Y1, dtype: float64
```

Table 6: Descriptive statistics for heating load

```
count    614.000000
mean      25.034821
std        9.528055
min       10.900000
25%       15.800000
50%       25.135000
75%       33.635000
max       48.030000
Name: Y2, dtype: float64
```

Table 7: Descriptive statistics for cooling load