CPSC 6300 Applied Data Science Individual Project

Energy Efficiency Analysis of Residential Buildings using Machine Learning Algorithms

Victor Liu

Introduction

In this project, I built three algorithmic models of machine learning, including linear regression model, SVM model and random forests model, to explore the relationship between the energy performance of residential buildings and parameters of residential buildings with 12 different shapes simulated in Ecotect. The buildings vary in the glazing area, the glazing area distribution, and the orientation, amongst other parameters. Before train the dataset, I use backward stepwise selection method to choose the "better" predictors as the reduced model based on AIC. Comparing to the full model, we have:

$$Y1' = \beta_0+\beta_1X1+\beta_2X2+\beta_3X3+\beta_5X5+\beta_7X7+\varepsilon \text{ (Reduced)}$$

$$Y1 = \beta_0+\beta_1X1+\beta_2X2+\beta_3X3+\beta_4X4+\beta_5X5+\beta_6X6+\beta_7X7+\beta_8X8+\varepsilon \text{ (Full)}$$

Also, I use the methods of normalization and lasso to improve the models. Then I fit the models with three Machine Learning Algorithms, we could easily estimate the heating load and cooling load by these models with high accuracy (Actually I use only heating load as response because of the high linear correlation between heating load and cooling load). In addition, I also evaluate the different performances of these models with the same dataset by 10-fold cross validation.

Dataset

This dataset comes from a building energy simulation tools named Ecotect, which generates 768 observations totally. This dataset consists of 8 predictors: relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, glazing area distribution, and 2 responses: heating load and cooling load. The corresponding mathematical representation are: X1, X2, X3, X4, X5, X6, X7, X8, y1 and y2. Amongst the 8 predictors, orientation, glazing area and glazing area distribution are indicator variables. Except them, the rest of these predictors and the 2 responses are all numbers. Besides, I split the total dataset into train set and test set randomly at a ratio of 7:3 to better assess their results.

Exploratory data analysis

According to the correlation matrix and scatter plot matrix, we find that there exists multicollinearity in the dataset, such as X1, X2 and X4. What's more, y1 and y2 have a significant linear relationship. So in the following analysis, I would focus on the effect of the input variables on y1(heating load) only. And I would drop some predictors to compare their performances with the original one as I mentioned in the Introduction.

Link to online resource: http://archive.ics.uci.edu/ml/datasets/Energy+efficiency
Citation: 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 (the paper can be accessed from [Web Link])

	X1	X2	X3	X4	X5	Y1	Y2
X1	1.0000000	-0.9919015	-0.2037817	-0.8688234	0.8277473	0.6222722	0.6343391
X2	-0.9919015	1.0000000	0.1955016	0.8807195	-0.8581477	-0.6581202	-0.6729989
X3	-0.2037817	0.1955016	1.0000000	-0.2923165	0.2809757	0.4556712	0.4271170
X4	-0.8688234	0.8807195	-0.2923165	1.0000000	-0.9725122	-0.8618283	-0.8625466
X5	0.8277473	-0.8581477	0.2809757	-0.9725122	1.0000000	0.8894307	0.8957852
Y1	0.6222722	-0.6581202	0.4556712	-0.8618283	0.8894307	1.0000000	0.9758618
Y2	0.6343391	-0.6729989	0.4271170	-0.8625466	0.8957852	0.9758618	1.0000000

Table 1 Correlation Matrix

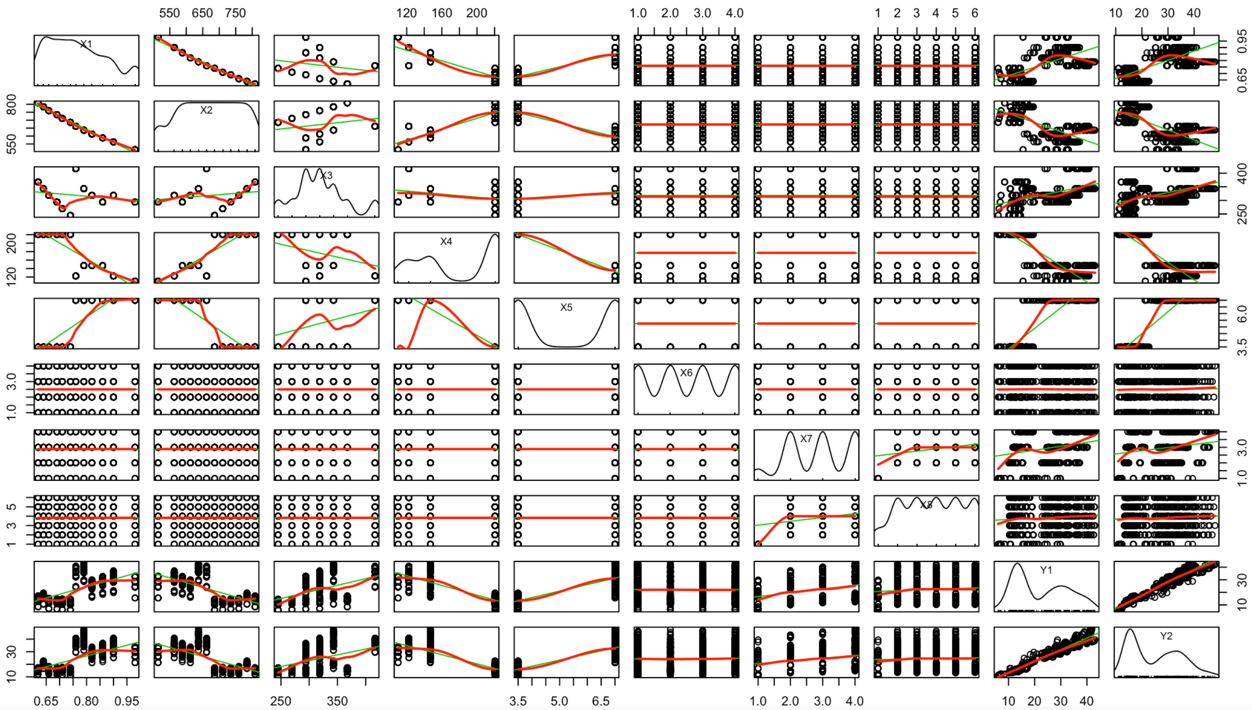

Figure 1 Scatter Plot Matrix

Model selection and validation

I try many methods, such as normalization and lasso, to improve the performance of these models (mainly based on the values of MAE and MSE), though, they do not perform much better, neither does the reduced model generated by backward stepwise selection comparing to the full model.

```
Coefficients: (2 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  73.252226  22.304002   3.284  0.00109 **
X1          -60.845247  12.030196  -5.058 5.87e-07 ***
X2           -0.079146   0.020096  -3.938 9.31e-05 ***
X3            0.056657   0.007924   7.150 2.92e-12 ***
X4                  NA         NA      NA       NA
X5            4.394712   0.401570  10.944  < 2e-16 ***
X63           0.106166   0.345764   0.307  0.75893
X64           0.089745   0.345888   0.259  0.79538
X65           0.131105   0.343797   0.381  0.70310
X70.1         5.465823   0.599671   9.115  < 2e-16 ***
X70.25        7.678098   0.598460  12.830  < 2e-16 ***
X70.4        10.183107   0.596513  17.071  < 2e-16 ***
X81           0.575018   0.404086   1.423  0.15533
X82           0.581134   0.407015   1.428  0.15395
X83           0.206011   0.392430   0.525  0.59983
X84           0.429866   0.405782   1.059  0.28993
X85                 NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.825 on 525 degrees of freedom
Multiple R-squared:  0.9237,    Adjusted R-squared:  0.9217
F-statistic: 454.1 on 14 and 525 DF,  p-value: < 2.2e-16
```

Figure 2 Summary of Linear Regression

```
Coefficients: (2 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.06970    4.98612   4.827 1.82e-06 ***
X1          -21.90429    4.33087  -5.058 5.87e-07 ***
X2          -23.26904    5.90831  -3.938 9.31e-05 ***
X3            9.71660    1.35894   7.150 2.92e-12 ***
X4                 NA         NA      NA       NA
X5           15.38149    1.40549  10.944  < 2e-16 ***
X63           0.10617    0.34576   0.307    0.759
X64           0.08975    0.34589   0.259    0.795
X65           0.13110    0.34380   0.381    0.703
X70.1         5.46582    0.59967   9.115  < 2e-16 ***
X70.25        7.67810    0.59846  12.830  < 2e-16 ***
X70.4        10.18311    0.59651  17.071  < 2e-16 ***
X81           0.57502    0.40409   1.423    0.155
X82           0.58113    0.40702   1.428    0.154
X83           0.20601    0.39243   0.525    0.600
X84           0.42987    0.40578   1.059    0.290
X85                NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.825 on 525 degrees of freedom
Multiple R-squared:  0.9237,    Adjusted R-squared:  0.9217
F-statistic: 454.1 on 14 and 525 DF,  p-value: < 2.2e-16
```

Figure 3 Summary of Linear Regression - Normalization

Start: AIC=1136.5 **Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8**
Step: AIC=1130.67 **Y1 ~ X1 + X2 + X3 + X5 + X7 + X8**
Step: AIC=1125.87 **Y1 ~ X1 + X2 + X3 + X5 + X7**

Table 2 Backward Stepwise Selection

As for the three Machine Learning algorithms, the SVM algorithm does not perform well in this dataset. Except for that, simple linear regression and random forests perform well in the dataset. They both have a high R^2 (>0.9), and as their results of evaluation in train set are similar to the ones in test set, we will not take into account overfitting.

Random forests model has extremely better performance by contrast to simple linear regression model (Their values in MSE and MAE are much lower).

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  72.217037  22.134816   3.263  0.00117 **
X1          -60.247361  11.933177  -5.049 6.12e-07 ***
X2           -0.078232   0.019951  -3.921 9.96e-05 ***
X3            0.056613   0.007892   7.173 2.47e-12 ***
X5            4.404977   0.399503  11.026  < 2e-16 ***
X70.1         5.832414   0.536069  10.880  < 2e-16 ***
X70.25        8.038657   0.537796  14.947  < 2e-16 ***
X70.4        10.542730   0.536970  19.634  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.815 on 532 degrees of freedom
Multiple R-squared:  0.9232,    Adjusted R-squared:  0.9222
F-statistic: 914.2 on 7 and 532 DF,  p-value: < 2.2e-16
```

Figure 4 Summary of Linear Regression – Reduced

```
Aggregating results
Selecting tuning parameters
Fitting mtry = 16 on full training set
'data.frame':   3 obs. of  7 variables:
 $ mtry      : num  2 9 16
 $ RMSE      : num  2.271 0.777 0.594
 $ Rsquared  : num  0.956 0.994 0.996
 $ MAE       : num  1.591 0.448 0.371
 $ RMSESD    : num  0.272 0.201 0.125
 $ RsquaredSD: num  0.01518 0.0033 0.00159
 $ MAESD     : num  0.1799 0.0563 0.035
```

Figure 5 Results of Random Forests

	MSE	**MAE**
Linear Regression	7.83	1.97
Linear Regression(Reduced)	7.73	1.94
Linear Regression (Normalization)	7.83	1.97
Linear Regression (lasso)	7.91	1.99
SVM	52.46	6.47
Random Forests	0.30	0.37

Table 3 MSE&MAE from different models

Conclusion

Reduced model formed by dropping some predictors does not perform much better than full model, so we keep all the predictors. From the summary of linear regression, we find that except for X1(relative compactness) and X2(surface area) that have negative relationship with y1(heating load), the other predictors keep a positive relationship with y1. In addition, random forests model performs much better than the other two model. And SVM model is not appropriate for this dataset.