Machine learning approaches for estimating building energy consumption

To cite this article: Liangyu Liu *et al* 2020 *IOP Conf. Ser.: Earth Environ. Sci.* **474** 052072

Machine learning approaches for estimating building energy consumption

Liangyu Liu[1], Ningyi Liu[2], Yilin Zhang[1], Yumeng Li[3], Xiaobo Rui[4] and Zi Yang[5*]

[1]Department of electrical and computer engineering, The Ohio State University, Columbus, OH, 43210 USA

[2]Department of statistics, The Ohio State University, Columbus, OH, 43210 USA

[3]Department of integrated systems engineering, The Ohio State University, Columbus, OH, 43210 USA

[4]State Key Laboratory of Precision Measuring Technology and Instruments, Tianjin University, Tianjin 300072, China

[5]Department of materials science and engineering, The Ohio State University, Columbus, OH, 43210 USA

[*]Corresponding author's email: 821365245@qq.com

Abstract. Using building data and corresponding weather conditions provided by ASHRAE, a statistical method that carefully measures features and applies both linear regression and gradient boosting machine models to predict and analyse building energy consumption was developed. Comparison of the predicted and actual energy usage indicates our model can predict energy consumption within an acceptable error range. Such statistical model has the potential to be widely used to monitor energy consumption and measure energy savings for various kinds of buildings in the future. If additional data is available, this method will be more widely applicable to other sectors such as industrial facilities.

1. Introduction

From houses and hotels to schools and skyscrapers, buildings in the United States consume almost 40% of the country's energy for cooling, heating, lighting and equipment operation. Any efficient climate protection strategy must take residential and commercial buildings, which are responsible for approximately 40% of U.S. carbon dioxide emissions, into account [1]. Since the energy consumption of buildings is significant, the prediction of energy use in buildings is important in order to improve their energy performance and reduce environmental impact. Furthermore, energy demand forecasting also helps design and construct energy-saving buildings.

Energy consumption of a building is affected by internal factors such as primary use, number of floors and area of each floor, as well as external factors such as local weather (temperature, precipitation, wind speed, etc.). Such a complex situation makes it difficult to implement the prediction of building energy usage. So far, several prediction methods such as engineering models and artificial intelligence methods have been developed. The engineering methods, which use physical theories to characterize thermal dynamics and energy behaviours of buildings, have been well developed over the past fifty years. Al-Homoud reviewed a simplified engineering method called 'degree day' in which only one index, daily temperature, is analysed. This method performs well for predicting the energy usage of small buildings [2]. Yik et al. calculated cooling load profiles of different buildings using simulation tools [3]. Artificial Neural Networks (ANNs) have been studied for more than twenty years and are the most widely used artificial intelligence methods to estimate the energy use of buildings. This approach is effective for such complex application and solving nonlinear problems. For instance, Kreider et al. studied a recurrent neural network on hourly energy consumption to predict energy needs for building heating and cooling using only the weather and timestamp datasets [4].

In this paper, we propose a statistical method that carefully measures features and applies both linear and gradient boosting machine (GBM) models to predict and analyse building energy consumption based on building meta data and weather conditions provided by ASHRAE. A comparison of the predicted and actual energy usage indicates our model can predict energy consumption within an acceptable error range. In the future, this model has the potential to be widely used to monitor energy consumption and measure energy savings for various kinds of buildings. If additional data is available, this method will be more widely applicable to other sectors such as industrial facilities.

2. Methods

Energy usage, characterized by meter reading, can be defined as a regression problem given observable variables such as weather, building information, and site id. For input data, there were 7 weather variables, 2 building info variables, 1 site id variable and 1 time variable. The output was the predicted meter reading. Two models were studied: linear regression model and gradient boosting model. For each model, in addition to using all variables as input, multivariate methods were also applied to reduce the dimension of weather features. For all developed models, evaluation metric used for meter reading prediction is the Root Mean Squared Logarithmic Error (RMSLE), which is defined as

$$\epsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(p_i + 1) - log(a_i + 1))^2}$$

where n is the total number of observations. p_i is the predicted meter reading. a_i is the ground truth meter reading [5]. Smaller error means better prediction. ASHRAE dataset has a dedicated testing set which was used for prediction accuracy analysis.

One multivariate method was Principal Component Analysis (PCA). To reduce the dimension of weather variables, PCA was performed on the 7 variables. Covariance matrix and correlation matrix were built using all data from different sites and spectral decomposition of covariance matrix was done. Five largest components were selected to keep 80% of total variance. For PCA, the assumptions are no existed outliers and variables are linearly related. Outliers were checked and removed during exploratory data analysis (EDA) [6].

The other multivariate method was Linear Discriminant Analysis (LDA), in which weather features were studied to separate different time frames and site ids using discriminant analysis. Within-group variation and between-group variation matrices were calculated in order to solve the first 2 eigenvectors of the generalized eigen decomposition. There are two assumptions for LDA. The first assumption is homoscedasticity, meaning different groups should have different covariance structure. This assumption is hard to check because true labels of different energy consumption groups are unknown, but the covariance of building usage can be obtained to evaluate this assumption. The other assumption is that the discriminant that best separates various groups needs to be a linear combination of original variables [7].

GBM model is based on the fact that the fastest decrease of a cost function in a relatively small neighbourhood can be found [8]. It is a machine learning technique for both regression and classification problems. Like other boosting models, GBM is usually ensemble of weak prediction models generated by minimizing the loss function. In this study, lightGBM optimization was chosen with log function to build a loss function. Features of lightGBM include leaf-wise growth strategy, histogram split search algorithm and categorical feature support, which make LightGBM an efficient and accurate framework.

3. Results and discussions

3.1. PCA analysis of weather variables

PCA analysis was done on weather dataset in order to find the correlation between month, air temperature, cloud coverage, dew temperature, precipitation depth in one hour, sea level pressure, wind direction and wind speed. It is necessary to reduce the number of variables in weather feather and only use principal elements for further analysis. According to PCA results generated using covariance and correlation matrices shown in Figure 1, there is an obvious indication of scaling problem with the original data. Therefore, correlation matrix was chosen instead of covariance matrix in this study to compare data samples from various populations.

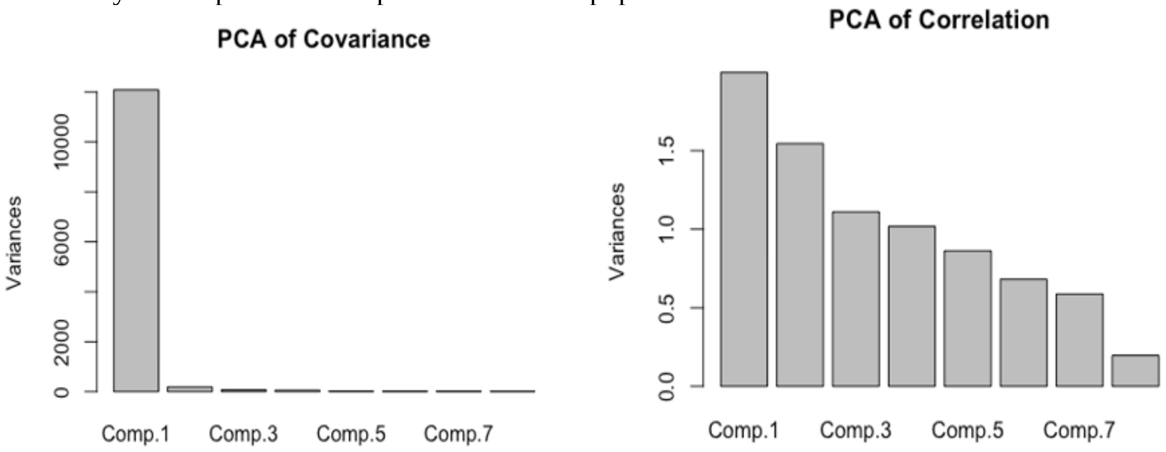

Figure 1. Scree plot of Principal Component Analysis.

According to PCA results listed in Table 1, 80% was chosen as cut-off point and the first 5 components were used as principal components. Combined with real-life meanings, the first 5 components are intercepted as follows: the first component is highly positively correlated to air temperature and dew temperature. It also correlates positively to month and negatively to sea-level pressure, wind direction and wind speed. Hence, it was defined as temperature component. The second component is highly negatively correlated to wind speed, sea-level pressure and wind direction. It was defined as atmosphere component. The third component has the highest correlation to precipitation depth in one hour and could coverage. It was obviously a precipitation component. The fourth component is highly correlated to month and wind direction. Since wind direction depends on the month of a year, the fourth component was defined as date component. The fifth component is a difference between cloud coverage and precipitation with sea level and wind speed positive correlated, so it was a thunder component.

Table 1. PCA loadings of weather variables.

	Comp.1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6	Comp. 7
month	0.284	0.142	0.141	0.748	-	0.541	0.137
air_temp	0.631	-0.134	-0.163	-	-	-0.264	-
cloud_coverage	-	-0.279	0.609	-0.169	0.696	-	0.181
dew_temp	0.650	-0.106	-	-	-	-0.214	-
precip_depth_1_hr	-	-0.149	0.718	-	-0.650	-0.132	-0.122
sea_level_pressure	-0.141	0.528	0.158	0.365	0.266	-0.539	-0.426

wind_dir	-0.221	-0.460	-0.168	0.488	-	-0.502	0.467
wind_speed	-0.165	-0.600	-0.119	0.195	0.110	0.175	-0.720

3.2. LDA analysis of weather variables

Weather features were used to cluster different sites. As shown in Table 2, the first discriminant component was a combination of dew temperature, cloud coverage and sea level pressure. The second discriminant component was the difference between air temperature and dew temperature combined with cloud coverage. These two components can discriminate among different sites.

Table 2. LDA loadings of weather features with respect to sites.

	air_temp	dew_temp	cloud_coverage	precip_depth_1_hr	sea_level_pressure	wind_dir	wind_speed
0	0.0056	0.1287	0.2298	0.0123	0.1050	0.0009	0.0590
1	-0.2331	0.2200	-0.7790	-0.0296	-0.0433	0.0017	0.1653

Weather features were also used to cluster different months. As shown in Table 3, the first discriminant component was a combination of air temperature, dew temperature and cloud coverage. The second discriminant component was mainly related to dew temperature.

Table 3. LDA loadings of weather features with respect to months.

	air_temp	dew_temp	cloud_coverage	precip_depth_1_hr	sea_level_pressure	wind_dir	wind_speed
0	-0.1516	-0.1059	0.1301	-0.0081	-0.0386	-0.0013	-0.0239
1	-0.0761	-0.1190	0.0725	-0.0066	-0.0002	-0.0013	0.0230

Figure 2 shows the discriminant analysis of weather features with respect to sites and months. By using discriminant components, summer and winter months were distinguished successfully.

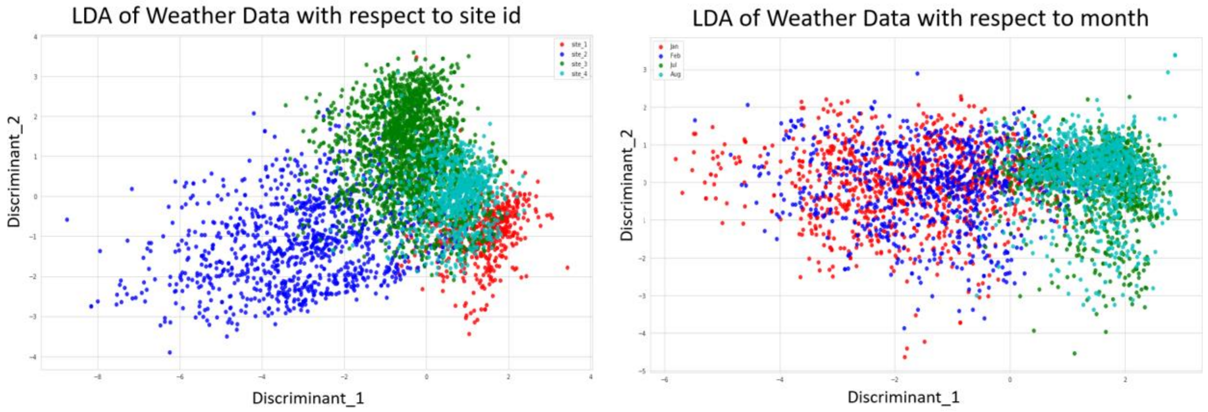

Figure 2. Discriminant analysis of weather features with respect to (left) sites and (right) months.

3.3. Comparison of Accuracy of Different Models

Firstly, a linear regression model was built as prediction baseline. Weather variables were input as the first 5 principal components. Based on the regression result, P-values for all variables were close to zero, indicating all variables were significant in the linear model. R-squared of 0.988 means that 98.8% of the variance in response variable (average energy usage) was predictable from independent explanatory variables. All features showed the fitness of the linear regression model was good. After natural logarithm transformation, every one unit increase of site id, primary use, square feet, month, air temperature, could coverage, dew temperature, precipitation in one hour, sea level pressure, wind direction and wind speed will cause 0.0148, -0.0098, 0.8225, 0.0492, 0.0114, 0.0295, -0.0130, 0.0018, -0.0053, -0.0002, -0.0204 unit increase of the natural logarithm of meter reading, respectively.

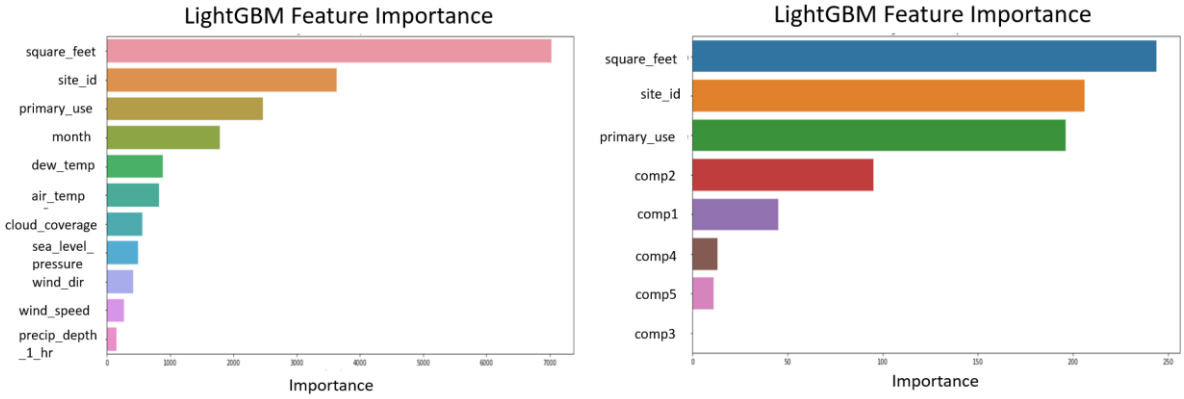

Figure 3. Importance of features (left) without PCA components and (right) with PCA components.

Two LightGBM models were developed in this study. One was trained with same variables used in linear regression model, and the other was trained with the first 5 PCA components that explained 80% of weather parameters. Category data like primary_use was encoded to 0-15. LightGBM models were trained with 4-folder validation to avoid overfitting, meaning 25% of data was used as validation data. K-folder validation strategy uses part of the data as validation dataset, and if the training score keeps improving while validation score starts to slip back, training is then stopped. The importance of training features with and without PCA components are shown in Figure 3. The training was stopped when the score of the validation set started to decrease. After natural logarithm transformation, the best score for both training set and validation set was RMSLE. Figure 4 shows the residual plots for all three models.

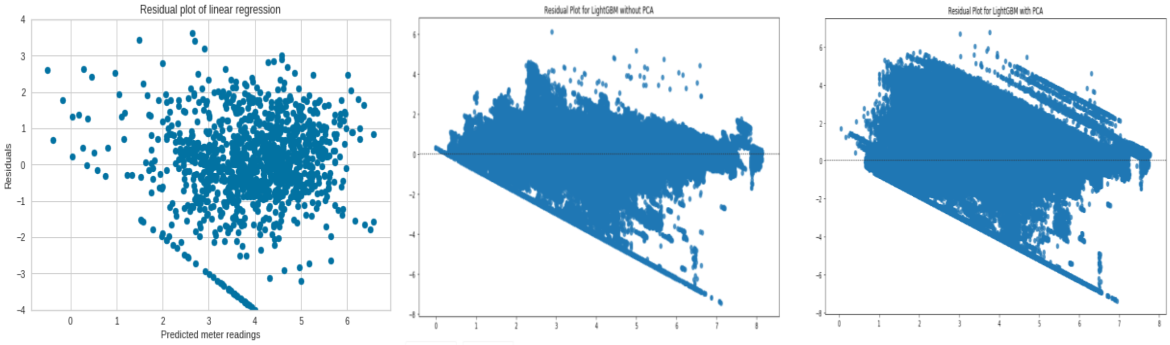

Figure 4. Residual plots of (left) linear prediction (middle) LightGBM without PCA and (right) LightGBM with PCA.

To reach the maximum prediction accuracy and avoid overfitting, time series feature was added to improve the model. Human activities vary at different time of a day, so various time features were built for accurate prediction, as well as lag and window features. The improved model achieved a RMSLE score of 0.66 on validation dataset, making it the model with the smallest error.

The prediction results of all models are summarized in Table 4. For linear regression model training with all variables, RMSLE was 1.38. When using principal components instead of weather variables, the RMSLE of linear regression model was 1.41. Training the LightGBM model with all variables resulted in a relatively low RMSLE of 0.84 on validation set. Using principal components instead of weather variables for LightGBM model could speed up the training process, but RMSLE increased to 1.35, indicating rough training result. Building a feature vector with generalized lag and window features could deliver more accurate results, but huge computational effort was consumed for such small improvement.

Table 4. Comparison of different methods in terms of prediction results.

Method	Linear regression	Linear regression + PCA	LightGBM	LightGBM + PCA	LightGBM + Additional features
RMSLE	1.38	1.41	0.84	1.35	0.66

4. Conclusion

By comparing the predicted energy usage with the real meter reading, questions like whether the building energy consumption is normal, or should the electrical system be replaced can be answered. After analyzing correlations between meter readings and environmental parameters, we are able to get insights on how to distribute energy more efficiently in the future. Furthermore, electrical companies can make budget plans more accurate with existing data by using statistical prediction models. In general, obtaining prediction result with high accuracy from a statistical model is time consuming and requires clean and not overlapped data with carefully tweaked training parameters. In this study, linear regression and lightGBM models with PCA analysis can only deliver a relatively rough result. Data manipulation method and energy prediction model were carefully chosen to balance the training consumption and result accuracy. If only the relationship between meter readings and weather factors is considered, a linear model or a correlation plot may be enough, while more accurate models could support better market incentives and enable lower cost financing. One caveat of this method is that the time series was not considered. Variables in different time slots were treated independently although there could be correlations between them along time. Overall, the lightGBM model with PCA analysis and additional feathers developed in this study is quite accurate and a small RMSLE value (less than 1) can be obtained.

References

[1] Zhao, H.X., Magoulès, F. (2012) A review on the prediction of building energy consumption. Renew. Sust. Energ. Rev., 16: 3586-3592.

[2] Al-Homoud M.S., (2001) Computer-aided building energy analysis techniques. Building and Environment, 36: 421-433.

[3] Yik, F.W.H., Burnett, J., Prescott I., (2001) Predicting air-conditioning energy consumption of a group of buildings using different heat rejection methods. Energy and Buildings, 33: 151-166.

[4] Kreider, J.F., Claridge, D.E., Curtiss, P., Dodier, R., Haberl J.S., Krarti, M., (1995) Building Energy Use Prediction and System Identification Using Recurrent Neural Networks. J. Sol. Energy Eng., 117: 161-166.

[5] ASHRAE-Great Energy Predictor III (2019). https://www.kaggle.com/c/ashrae-energy-prediction.

[6] Saeys, Y., Degroeve, S., Aeyels, D., Rouzé, P., Van de Peer, Y. (2004) Feature selection for splice site prediction: A new method using EDA-based feature ranking. BMC Bioinformatics, 5: 64.

[7] Huang, R., Liu, Q., Lu H., Ma, S. Solving the small sample size problem of LDA. Object recognition supported by user interaction for service robots, Quebec City, Quebec, Canada, 2002, pp. 29-32 vol.3.

[8] Cao, Y., Gui, L. Multi-Step wind power forecasting model Using LSTM networks, Similar Time Series and LightGBM. 2018 5th International Conference on Systems and Informatics (ICSAI), Nanjing, Jiangsu, China, 2018, pp. 192-197.