EAI Endorsed Transactions
on Energy Web

Research Article EAI.EU

Extreme Gradient Boosting Algorithm for Energy Optimization in buildings pertaining to HVAC plants

Monika Goyal, Mrinal Pandey*

Department of Computer Science and Technology, ManavRachna University, Faridabad, India

Abstract

With the recent advancements in technology, energy is being consumed at a great pace in almost every region. Buildings are the biggest consumer of energy, almost 40% of total energy is being consumed by the buildings. The purpose of this research is to investigate Ensemble Learning based optimal solution for predicting energy consumption in Heating, Ventilation and Air Conditioning (HVAC) plants as the HVAC unit consumes a large percentage of energy in buildings. The study focuses on Cooling Tower data of HVAC plants as Cooling Tower carries a major responsibility for maintaining ambient within a building. In this paper,four Regression techniques namely Multiple Linear Regression, Random Forests, Gradient Boosting Machines and ExtremeGradient Boosting have been experimented. The findings reveal that Extreme Gradient Boosting Ensemble outperforms with higher accuracy and lower in overfitting.

Keywords: Machine Learning, Ensemble, Energy Optimization, HVAC,Extreme Gradient Boosting.

Received on 31 August 2019, accepted on 02 May 2020, published on 15 May 2020

doi: 10.4108/eai.13-7-2018.164562

*Corresponding author. Email:mrinal@mru.edu.in

1. Introduction

The ever-increasing dependency on technology in almost every sector in the present world has put forward the issue of quick consumption of energy. Energy is required for the smooth functioning of various types of machines like microwaves, refrigerators, water heaters, washing machines and other appliances in homes and hotels, Computers, printers in offices, heavy machines like those required for welding, packaging, robotics in industries, energy required in means of transport. Although, advancements in technology have made human life easier and comfortable, but they have taken a toll on natural non-renewable resources. This is so because energy is conventionally generated by the burning of fossil fuels like oil, coal and natural gas, which take millions and millions of years to form. If the consumption of precious energy continues at such a high rate, soon the non-renewable energy sources will deplete. Also, burning of fossil fuels releases carbon, More the fossil fuels burnt more carbon is emitted which consequently leads to air and water pollution [1]. Furthermore, to ensure the availability of energy for future generations, its misuse should be tackled urgently.

End users can contribute towards preventing the misuse of energy by following simple measures like i) switching off the electrical appliances when not in use ii) maintaining indoor temperatures by maintaining indoor temperatures in buildings by keeping windows, window shades and blinds open or close as required iii) changing normal light bulbs with CFLs and LED lights iv) changing air conditioner, furnace and heat pump filters at regular intervals v) installing high efficiency and solar heating systems in place of old conventional water heaters. Such measures may help saving energy to some extent.

To conserve energy on a larger level, identification of areas that utilize a huge amount of energy is necessary. Studies reveal that globally Buildings are the largest consumer of energy followed by Transportation and Industry. As per the reports, buildings are accountable for nearly 40% of energy usage throughout the world [2-5] and it is greater than the energy consumed by the other two areas namely industry and transport which as per the reports comes out to be 32% and 28% respectively. These reports motivated the researchers to target the biggest energy consumer i.e. buildings.

The energy consumption pattern analysis in buildings shows that the biggest chunk is consumed by HVAC (Heating, Ventilation and Air conditioning) system. An HVAC unit consists of distinct components like Chillers, Cooling towers, Primary pumps, and Secondary pumps. Each of these components performs its designated work and consumes the power to operate. HVAC consumes around 40% - 50% of the total energy consumed in a building [6-7]. This study covered the way to analyze the energy consumption profile of HVAC.

To handle the issue of energy consumption in buildings several researchers across the world have applied Machine Learning, as it has a vast variety of techniques that aid in the analysis of various applicative domains effectively according to the problem specification [8].Apart from applying the ML algorithms individually, ensemble models can be created by specifically combining different models, which improve the effectiveness of the model in terms of accuracy and performance.

In this paper four Regression techniques were applied to predict the power consumption of HVAC plants. First, Multiple Linear Regression, Second the Random Forest as a bagging variant, Third Gradient Boosting Machines and fourth Extreme Gradient Boosting as a boosting variant have been experimented. Extreme Gradient Boosting (XGBoost) is one of the latest homogenous ensemble techniques of boosting variant ,proposed by Tianqi Chen and Carlos Guestrin [9] which has decision tree based underlying structure and uses gradient boosting frame work. The applications of XGboost have not been explored much in the area of energy optimization. Thus this paper presents a case study of XGBoost, particularly in the area of energy consumption due to cooling tower of HVAC plants. The results of the experiments also prove that XGBoost outperforms other ensemble algorithms by an appreciable margin.

The organization of this paper is as follows: Section 1 introduces the problem and gives a brief overview of the Machine Learning techniques. Related work is presented in Section 2. Section 3 describes the Machine Learning algorithms used for performing our experiments. The Methodology of the work done is described in Section 4. Section 5 concludes the paper followed by references.

2. Related Work

This section investigates the work of several authors related to the optimization of energy within buildings.

Authors in [10] proposed two frameworks for anomaly detection in HVAC power consumption. One was a pattern based anomaly classifier called CCAD-SW(Collective contextual anomaly detection using sliding window) which created overlapping sliding windows so that anomalies can be pointed out as soon as possible. This framework made use of bagging for improved accuracy. Another was a prediction based anomaly classifier called EAD (Ensemble Anomaly Detection) which used Support Vector Regression and Random Forests. Experiments were performed on HVAC power consumption data collected from a school in Canada and results show that EAD performed better than CCAD-SW in terms of sensitivity and reducing False Positive rate.

Decision Tree Analysis was performed in [11] to predict the cost estimations of HVAC while designing buildings. The HVAC subsystems are CP (Central Plant) system, WD (Water Side Distribution) system, and AC (Air Conditioning) system. Different combinations of these sub systems result in different costs of HVAC plants. The study was carried out in office buildings in Korea. The study showed that the AC component of HVAC has maximum impact on the cost followed by CP and then WD has minimum impact.

Authors in [12] applied six Regression techniques: Linear Regression, Lasso Regression, Support Vector Machine, Random Forest, Gradient Boosting and Artificial Neural Network on for estimating Energy Use Intensity in Office buildings and energy usage by HVAC, plug load and lighting based on CBECS 2012 microdata. Out of them, Random Forest and Support Vector Machine were found comparatively robust.

A sensor-based model was proposed by the authors of [13] for forecasting the energy consumed by a multi-family residential building in New York City. The model was built using Support vector regression. Authors analyzed the prediction performance through the perspective of time and space and found that the most optimal prediction was hourly prediction at by floor levels.

In another paper [14] the authors developed a framework in which they used clustering algorithm and semi-supervised learning techniques to identify electricity losses during transmission i.e. between source and destination. The technique also helps in optimizing the losses. Deep learning is used for semi-supervised machine learning because of its ability to learn both labeled and unlabeled data. The electricity consumption, heating, cooling and outside temperature data was obtained from a research university campus in Arizona.

The work done in [15] used various supervised classifiers- DT(Decision Trees), DA(Discriminant Analysis), SVM (Support Vector Machines) and KNN (K- Nearest Neighbours) to disaggregate the data of power consumption by multiple HVAC units into that consumed by individual HVAC, while the data was retrieved collectively from single meter to reduce cost and complexity. The Power consumption information of individual appliances is necessary for accurate energy consumption monitoring. The experiment was performed by collecting data from a commercial building in Alexandria. The results show that K- Nearest Neighbours was most efficient in power disaggregation.

A component-based Machine Learning Modelling approach was proposed [16] to counter the limitations of the Building Energy Model for energy demand prediction in buildings. Random Forest was selected andapplied on

the climate data collected from Amsterdam, Brussels, and Paris. MLMs excel over BEM as they generalize well under diverse design situations.

The work done in [17] witnessed the collection of energy consumption data of a house in Belgium and outside weather data and application of four Machine Learning algorithms namely Multiple Linear Regression, Support Vector Machine with Radial Kernel, Random Forest and Gradient Boosting Machines to predict the energy consumption and to rank the parameters according to their importance in prediction. They proposed that GBM was best at prediction.

The author in [18] proposed an ensemble technique which is a linear combiner of five different predictor models: ARIMA, RBFNN, MLP, SVM and FLANN. The combiner model was applied on stock exchange data for predicting the closing price of stock markets and it proved to be better in terms of accuracy as compared to individual models.

In yet another paper [19] the authors applied several Supervised Machine Learning techniques including Classification, Regression and Ensemble techniques to estimate the air quality of Faridabad by predicting the Air Quality Index. The algorithms applied include Decision Tree, SVM, Naïve Bayes, Random Forests, Voting Ensemble and Stacking Ensemble. They concluded that Decision Tree, SVR and Stacking Ensemble outperform other methods in their respective categories.

A framework was proposed by the authors of [20] in which they selected 8 different characteristics of a residential building as input parameters and depicted their effect on the 2 output parameters- Heating load and Cooling load. Linear Regression and Random Forests were applied and results showed that Random Forests was better at predicting Heating and Cooling load in terms of accuracy.

3. Machine Learning

Machine Learning is the concept in which a machine learns and behaves in a certain manner when a particular type of data is fed as input. Machine Learning can be classified as Unsupervised Learning and Supervised Learning. Unsupervised Learning is usually descriptive in nature and the results are obtained in the form of patterns or groups depending upon a certain similarity metric. A common technique in unsupervised learning is clustering in which the given dataset is grouped into a given number of clusters depending upon the distance metric. Supervised Learning is predictive in nature, in which the input data is mapped to the desired output using a set of training data. Two common Supervised Learning Techniques are Classification and Regression.

3.1 Regression

Regression can be viewed as a statistical methodology generally used for numeric prediction. Regression can be classified as a) Linear Regression which involves finding the best line to fit two variables, such that one variable is independent called Predictor and can be used to predict the other variable which is dependent called Response, and b) Non – Linear Regression which involves more complex calculations and finds the best curve instead of best line. A common example is Polynomial Regression.

Multiple Linear Regression

Multiple Linear Regression [21-22]can be viewed as an extension to Linear Regression. MLRis used to model the dependence of a single response variable Y on multiple predictor variables X_1, X_2,……X_p. MLR can be graphically represented as shown in Figure 1.

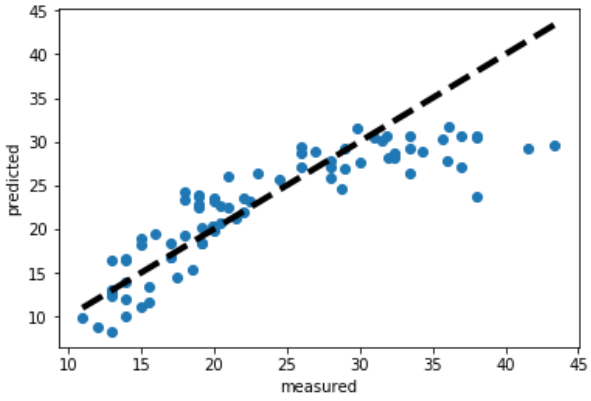

Figure 1. Multiple Linear Regression

The relationship of predictor and response variables can be expressed in the form of conditional expectation as shown in equation (1)

$$E(Y|X) = \beta_0 + \beta_i X_i \quad (1)$$

The slope β_i depicts the change in response variable Y when the predictor variable j is varied by one unit and other predictors are kept constant.

In MLR the amount of variation determined by the regression model is known as the coefficient of determination and is expressed by equation (2)

$$R^2 = \frac{SSR}{SSR+SSE} \quad (2)$$

In the above equation, SSE is residual sum of squares and is given by

$$\sum(y_i - \acute{y}_i) \quad (3)$$

Where $\acute{y}_i = b_0 + \sum b_j x_{ij}$ (4)

and SSR is regression sum of squares and is given by

$$\sum(\acute{y}_i - Y^{-j})^2 \quad (5)$$

Theoretically, MLR seems to be similar to LR but the interpretation of results of MLR is comparatively complex, mainly due to correlation among predictor variables. When the correlation between predictors changes, it greatly affects the estimate of slopes and intercept if only one of those predictors is fitted. Ignoring of predictors which are of importance affects R^2,where R^2 is used to measure the predictive power of regression, and also to interpret regression coefficients. So for better predictions, the predictor variables should be chosen very carefully such that relevant ones are not missed and irrelevant ones are not used. The various indices which can be considered during MLR interpretation are regression weights, zero-order correlation coefficients, structure coefficients, relative weights,product measures, all possible subsets regression, dominance weights, and commonality coefficients.

3.2 Ensemble technique

In Ensemble techniques, regression is performed by integrating the results of several individual models with the objective of improving the accuracy and robustness of prediction in learning problems having a numerical response variable. The two most popular homogenous ensemble methods are Bagging and Boosting [23-24].

Random Forests

Random Forests [25-27] were developed as an extension to the popular ensemble technique called Bagging. It is a tree-based ensemble technique where each tree depends on a set of random variables. Random Forests can be used for Classification as well as Regression. The appeal of Random Forests lies in several features like their speed of training and prediction, built-in estimate of the generalization error, applicability for high dimensional problems, handling of missing values and outliers in predictor variables etc. Figure 2 shows the schematic diagram of Random Forests.

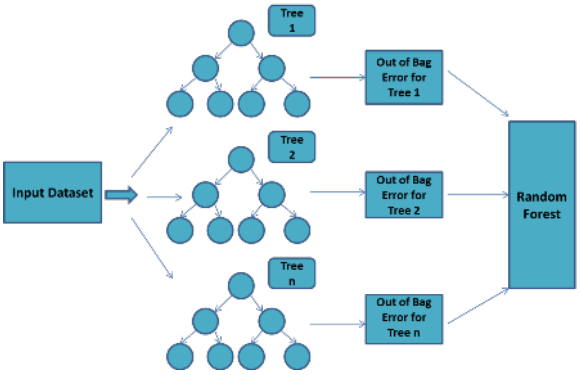

Figure 2. Random Forests

For predicting a continuous variable using Random Forests, the trees are grown depending on Θ,a random vector, in such a manner that h(x, Θ) which is the tree predictor takes on numeric values. The values of the response variable are numeric and it is assumed that the training sample is drawn independently from the distribution X of random vector Y. Equation (6) shows the mean square generalization error for a numeric predictor h(x)

$$E_{X,Y}(Y - h(X))^2 \quad (6)$$

The Random Forest predictor is constructed by taking the mean over k of the trees $\{h(x, \Theta_k)\}$

Equation (7) is a theorem which states the case when there are infinite numbers of trees in the forest

$$E_{X,Y}(Y - av_k h(X, \Theta_k))^2 \rightarrow E_{X,Y}(Y - E_\Theta h(X, \Theta))^2 \quad (7)$$

The above equation explains that by adding more number of trees the Random Forests do not overfit, but produces a limiting value of the generalization error.

Random Forests tend to be accurate and effective in prediction due to the right kind of randomness.

Gradient Boosting Machines

Boosting is a technique by which weak learners are converted into strong learners. In this method, each newly grown tree is a fit on an updated version of the original dataset. In boosting simple rules are combined to build an ensemble in a manner that results in improved performance of every ensemble member, that is, each member forming the ensemble is boosted. Where h_1, h_2,h_3…$h_T$as a set of hypotheses, and the composite hypothesis of ensemble be expressed as:

$$f(x) = \sum_{t=1}^{T} \alpha_t \, h_t(x) \quad (8)$$

Where t: ranges from 1 to T [28].

α_t: represents the coefficient which is used to combine the ensemble member h_t.

Gradient Boosting [29] builds additive regression models by iteratively fitting a simple base learner to currently updated pseudo-residuals by applying least squares at every subsequent iteration. The method behind Gradient Boosting Machines is shown in Figure 3.

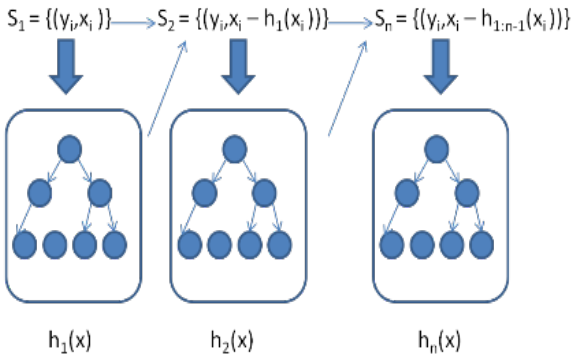

Figure 3. Gradient Boosting Machines

The objective of gradient boosting is to generate a function F*(x) which maps x to y, so that when the joint distribution of all values (y,x) is taken, the expected value of Ψ(y, F(x)) which is some specified loss function is minimized [30]. This relation is depicted in equation (9).

Where y: is the random output or dependent variable and x={x_1, x2, ...x_n} is a set of random input variables.

$$F^*(x) = \arg\min E_{y,x}\Psi(y, F(x)) \quad (9)$$

Gradient Boosting machines are highly flexible. There are several tuning parameters that increase their flexibility including the Number of trees, Depth of trees, Learning rate and Subsampling. To improve the performance of GBM, randomness was included in the original algorithm. During training, each iteration a random sub-sample of the training data is drawn without replacement from the whole training data set. Instead of using the whole training dataset, this sub-sample is used to fit the base learner and compute the update in the model for the current iteration[30]. Simulation studies show that the performance of the model is dependent upon the average absolute error of the derived approximation F′(x) while predicting each target F*(x) as shown in equation (10)

$$A(\acute{F}) = E_x|F^*(x) - \acute{F}(x)| \quad (10)$$

Extreme Gradient Boosting

Extreme Gradient Boosting [31] can be viewed as a scalable tree boosting algorithm which has key features of execution speed and model performance. Figure 4 describes the algorithm.

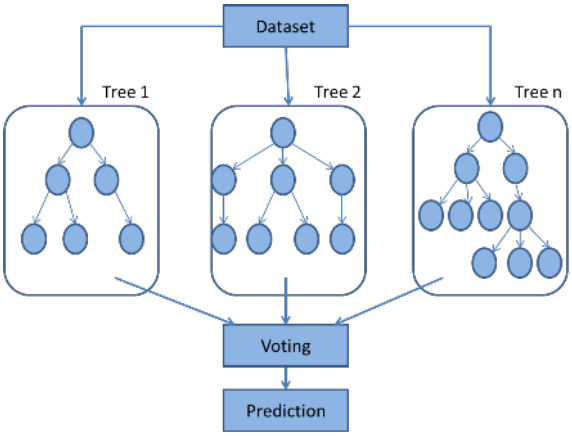

Figure 4. Extreme Gradient Boosting

To ensure scalability of the algorithm several optimizations included are: a novel tree learning algorithm to handle sparse data, parallel and distributed computing for speed up learning, out-of-core computation.
Equation (11) defines the model:

$$\pounds(t) = \sum_{i=1}^{n} l\left(Y_i, \hat{Y}_i(t-1) + f_t(x_i)\right) + \Omega(f_t) \quad (11)$$

Where l :The differentiable convex loss function which measures the difference between predicted and target values
Ŷi(t-1): is the prediction of i-th instance at the t-th iteration

4. Methodology

This section of the paper outlines the workflowbeginning with data collection, then various steps of data pre-processing followed by application of ML algorithms and results. Figure 5represents the methodology adopted in the paper.

4.1 Data collection and Description

The data of HVAC plants was collected from a hotel building in New Delhi, India for this research. It consists of HVAC data from sensor recordings at every 5-minute intervals for one year from Oct 2017 to September 2018. The data were categorized into two categories as Humidity (May-Nov) and Non-Humidity (Dec–April) depending upon two weather conditions.

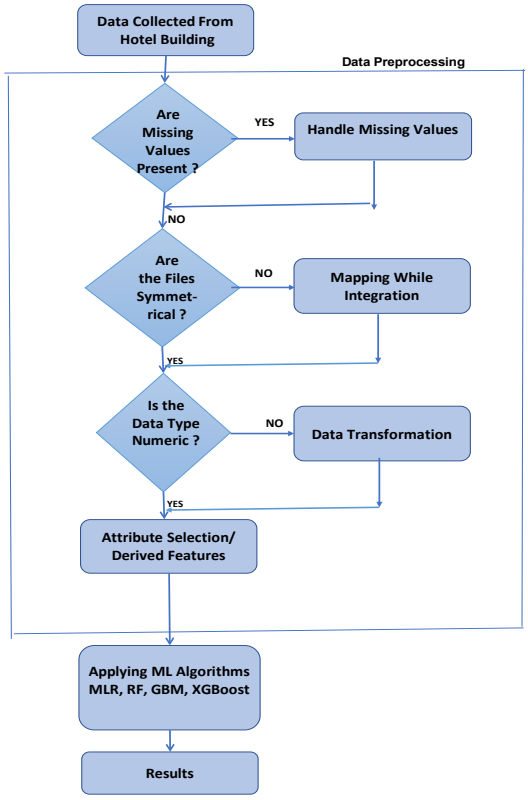

Figure 5. Work Flow of the proposed methodology

Figure 6 depicts the energy consumption pattern from December to April where energy consumption is less as there is no humidity in the weather. Similarly, Figure 7 depicts the energy pattern for the remaining months of the year which consume comparatively more energy due to humidity in the weather. It can be noticed from Figure 6 and Figure 7 that in the Month of April maximum energy consumption is approximately 10000 KWH, whereas the energy consumption in August is more than 40000 KWH. Therefore, in this research, the data of April and August were considered for analysis & prediction. April represents non-humid weather and August represents humid weather conditions particularly in the NCR region of India. It must be mentioned thatHVAC consumes more energy in humid conditions to counter humidity rather than summer or winters.

The attributes of the Cooling Tower data include Inlet temperature, Outlet temperature and energy consumed by Cooling Tower, Dry Bulb Temperature and Relative Humidity. Energy consumed by Cooling tower depends on Wet Bulb Temperature, so WBT is calculated using DBT and RH. The experiments for this research were performed on the Cooling Tower data. Table 1 describes the parameters along with the units in which each parameter is measured for the data used:

Table 1. Dataset Parameter description

Parameter	Description	Unit
DBT(Dry Bulb Temperature)	Ambient temperature	°Celcius
RH(Relative Humidity)	Amount of water vapour in air relative to its temperature	%age
WBT(Wet Bulb Temperature)	Temperature brought down by water evaporation	°Celcius
CT_INLET	Temperature of water entering into Cooling Tower	°Celcius
CT_OUTLET	Temperature of water exiting from Cooling Tower	°Celcius
CT_POWER	Power consumed by Cooling Tower	Kilo Watts

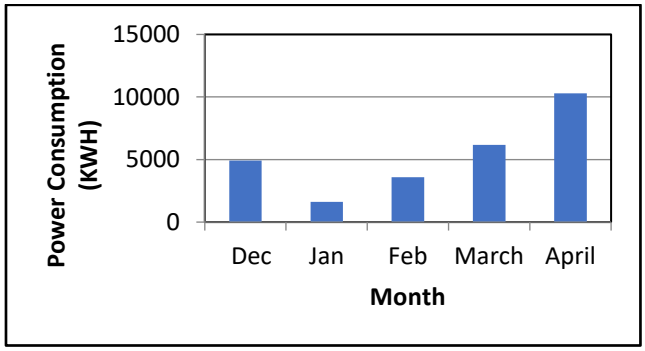

Figure 6. Energy consumption Dec 2017-April 2018

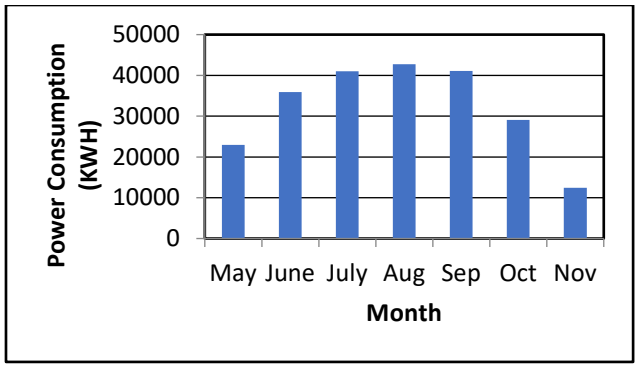

Figure 7. Energy consumption May2018-Oct 2018

Figure 8 represents the day-wise pattern of energy consumed by Cooling Tower in the months of April and August. Here x-axis denotes the days of month and y-axis denotes the power consumed in Kilo Watt Hour. The analysis of this graph shows that energy consumption in the month of August is nearly similar for each day with very few deviations only, whereas energy consumption in April is comparatively less similar and it shows a sudden

rise towards the end of the month. The pattern of energy consumption for the two months is entirely different as the temperature in August is similar throughout the month while in April temperature is moderate initially and starts rising towards the end.

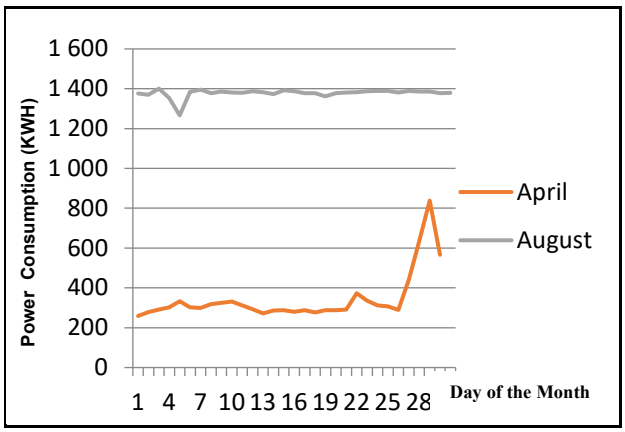

Figure 8. Energy consumption in April and August

4.2 Data Pre-processing

Data pre-processing consists of, filling the missing values, removing any outliers, transforming it into a form suitable for algorithm application, feature selection etc. [31]. The dataset used for this research was pre-processed in the following manner:

The data has been recorded by sensors at every 5-minute intervals. The power consumption dataset (Excel files) for the month of April 2018 and August 2018 consists of8639 and 9019 instances respectively. Similarly, Dry Bulb Temperature and Relative Humidity consist of 8926 and 8642 instances respectively. The Cooling Tower Inlet and Outlet instances were 8665 and 8960 instances respectively. The reason behind this inequality in the number of instances is that the data has been recorded by different sensors installed at different places so some sensors misrecorded some of the readings at 1-minute or 2- minute intervals instead of 5-minute interval. Some instances were not complete and most of their fields were blank.

There are approximately 288 instances for each day (24 hrs) if data has been recorded at 5-minute intervals. The days on which the number of instances was more than 288 had several extra readings, which were removed manually from the files to avoid any kind of flaw in result calculation. Thereafter the values of different parameters have been compiled in a single excel file such that all the entries of an instance represent values recorded at the same time.

The missing values in the incomplete instances were filled with the value of the previous record in the respective columns.

The original dataset consists of Dry Bulb Temperature and Relative Humidity as features, but the energy consumption of Cooling Tower also depends on Wet Bulb Temperature, therefore one more feature namely Wet Bulb Temperature has been derived from Dry Bulb Temperature and Relative Humidity for the experiments.

Since the type of data was factor, Data Transformation was done to convert it into numeric data to make it suitable for applying Regression algorithms.

4.3 Experiments, Results and Discussion

Four Machine Learning algorithms, "Multiple Linear Regression", "Random Forests", "Gradient Boosting Machines" and "Extreme Gradient Boosting" were applied using R. The algorithms were applied on both the datasets of April and August. The models evaluated using three well-known performance measures namely RMSE, MSE and R Squared.

Root Mean Square Error

Root Mean Square Error can be viewed as the standard deviation of residuals, where residuals indicate the distance of data points from the line of best fit i.e. these are the difference between actual and predicted values. Lower the value of RMSE better is the prediction.
Following equation defines the formula for RMSE:

$$RMSE = \sqrt[2]{\sum_{i=1}^{N} \frac{(\hat{Y}_i - Y_i)^2}{N}} \quad (12)$$

Where
Y_i:is the observed value for the i^{th}observation
$\hat{Y}_i$:is the predicted value
N :is sample size

Mean Square Error

Mean Square Error can be defined as the average of the error squared. Itis used as the loss function in least squares regression. MSE is the sum of the square of the difference between predicted and actual target variables, spanning over all the data points, divided by the total number of data points.
Following equation defines the formula for MSE:

$$MSE = \sum_{i=1}^{N} \frac{(\hat{Y}_i - Y_i)^2}{N} \quad (13)$$

R Squared

R Squared is used to statistically measure the closeness of data points to the fitted regression line. It is also known as the Coefficient of Determination. R squared can be defined by the following equation:

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{y})^2} \quad (14)$$

Where$\hat{Y}_i$: is the predicted value of Y
$\bar{Y}$:is the mean value of Y

The results of the aforementioned algorithms for the months of April and August 2018 are shown in table 2 and table 3 respectively.

It is evident from the results shown in Table 2 that RMSE and MSE have come out to be lowest when Extreme Gradient Boosting is applied. For the month of April, RMSE is 0.43 with XGBoost as compared to 5.32 with GBM, 5.08 with RF and 6.14 with MLR. The value of R Squared is 0.99 with XGBoost which is better as compared to GBM, RF and MLR which have the values 0.65,0.5 and 0.23 respectively.

Table 2. Results for the month of April 2018

ML Algorithm → Performance Metric ↓	MLR	RF	GBM	XG Boost
RMSE	6.14	5.08	5.32	0.43
MSE	37.78	25.88	28.3	0.19
R Squared	0.23	0.5	0.65	0.99

The results of the algorithms for the month of August are shown in Table 3.RMSE is lowest with value 2.81 when XGBoost is applied as compared to 3.72, 3.09 and 3.44 when GBM, RF and MLR are applied respectively. A similar difference can be seen in the values of MSE for all algorithms. R Squared value with XGBoost is also better than RF and MLR and approximately equal to GBM.

Table 3. Results for the month of August 2018

ML Algorithm → Performance Metric ↓	MLR	RF	GBM	XG Boost
RMSE	3.44	3.09	3.72	2.81
MSE	11.89	9.57	13.83	7.89
R Squared	0.05	0.43	0.57	0.5

Figure 9 represents the comparison chart of RMSE values obtained from experiments for datasets of April and August. It is evident from the chart that Ensemble techniques are better in terms of lower RMSE values as compared to normal Machine Learning techniques.

A similar comparison of April and August in terms of obtained R Squared values is shown in Figure 10. It again proves that Ensemble techniques are better in terms of higher R Squared values.

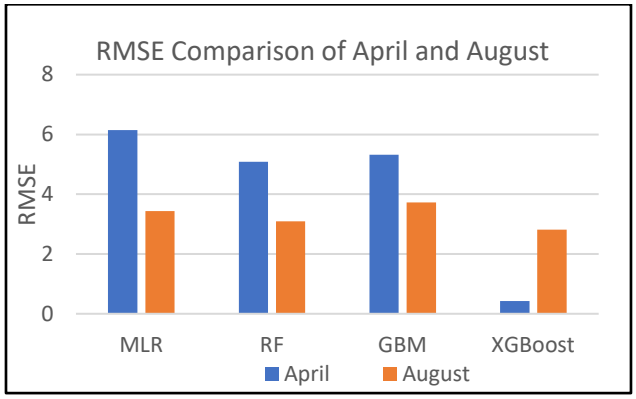

Figure 9. RMSE Comparison of April and August

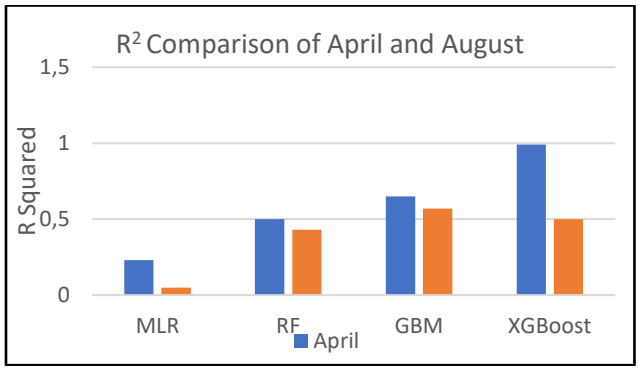

Figure 10. R^2 Comparison of April and August

As per the experiments performed, Extreme Gradient Boosting is the most appropriate algorithm for this research. Therefore, to validate the results Extreme Gradient Boosting algorithms was also applied on benchmark dataset obtained from the UCI repository [32]. This dataset [32] was also used by researchers [20] to perform regression. The dataset consisted of eight independent variables describing various building parameters and two dependent variables: Heating Load(Y1) and Cooling Load(Y2).

Table 4 depicts the comparison of the results of the algorithms applied to both the datasets. The value of RMSE is 0.43 and 2.81 for the months of April and August respectively when XGBoost was applied on the hotel building dataset while 0.55 and 0.96 for Heating Load and Cooling Load respectively when XGBoost was applied on the dataset obtained from UCI repository.

Similarly, MSE values are 0.19 and 7.89 for April and August for the Hotel dataset and 0.3 and 0.92 for Heating Load and Cooling Load for UCI dataset.

Additionally, the values of R Squared are 0.99 and 0.5 for April and August months for the Hotel dataset and 0.99 and 0.98 for Heating Load and Cooling Load for the dataset obtained from the UCI repository.

Table 4. Comparative Results of both datasets

Dataset → Performance Metric ↓	Hotel dataset April 2018	Hotel dataset August 2018	UCI dataset, Y1	UCI dataset, Y2
RMSE	0.43	2.81	0.55	0.96
MSE	0.19	7.89	0.3	0.92
R Squared	0.99	0.5	0.99	0.98
No. of Instances	8638	8926	768	768

Figure 11 also shows the graphical representation of the comparison of results of experiments performed on the hotel dataset and building dataset collected from the UCI repository.

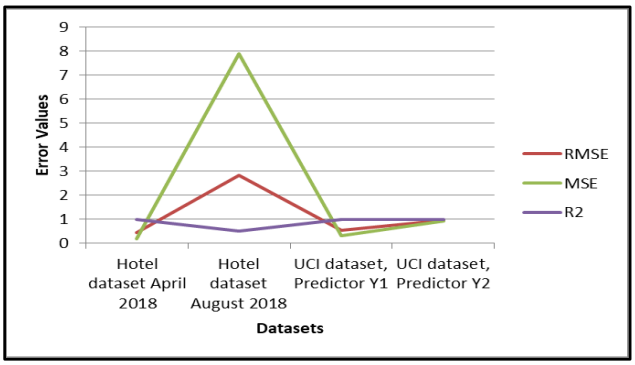

Figure 11. Comparative analysis of both datasets

5. Conclusion and Future Work

Energy being a precious resource needs to be utilized in the most efficient manner so that it is conserved while the comfort of consumers is also not compromised. Buildings are the largest consumer of energy globally and within a building HVAC accounts for the maximum energy consumption. In this paper energy consumption of HVAC plant was targeted for the prediction.

Four well known Regression algorithms namely Multiple Linear Regression, Random Forests, Gradient Boosting Machines and XGBoost algorithm were experimented. The findings of the experiments performed in this paper strongly recommend the use of Ensemble Machine Learning techniques as they perform better as compared to traditional Machine Learning techniques for the prediction of energy consumption.

The experimental results reveal that the XGBoost prediction model for cooling tower data achieves high accuracy and low over fitting for energy consumption in HVAC plants.

The scope of this paper was limited with the Cooling Tower data of HVAC plant. However, a Chiller is also an important component of HVAC, which contribute towards energy consumption in buildings. Therefore, the future work of this research will also include the data of Chiller for better prediction and energy optimization. In this research only two months data(April and August) were used for the prediction, future research will also focus on entire data for both the categories (humid and non humid data) of dataset.

References

[1] Goyal M., Pandey M. Energy Optimization in Buildings Using Machine Learning Techniques: A Survey. IJISMS.2018; 1(2).

[2] Jain R. K., Smith K. M., Culligan P. J., Taylor J. E. Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. APPL ENERG.2014; 123:168-178.

[3] Naganathan H., Chong W. O., Chen X. Building energy modeling (BEM) using clustering algorithms and semi-supervised machine learning approaches. AUTOMAT CONSTR.2016; 72: 187-194.

[4] Ahmad M. W., Mourshed M., Rezgui Y. Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. ENERG BUILDINGS.2017; 147: 77-89.

[5] Chou J. S., Bui D. K.Modeling heating and cooling loads by artificial intelligence for energy-efficient building design. ENERG BUILDINGS.2014; 82: 437-446.

[6] Carreira P., Costa A. A., Mansu V., Arsénio A. Can HVAC really learn from users? A simulation-based study on the effectiveness of voting for comfort and energy use optimisation. SUSTAIN CITIES SOC.2018.

[7] Drgoňa, J., Picard D., Kvasnica M., Helsen L. Approximate model predictive building control via machine learning. APPL ENERG.2018; 218:199-216.

[8] Banihashemi S., DingG., Wang J. Developing a hybrid model of prediction and classification algorithms for building energy consumption. Energy Procedia.2017; 110:371-376.

[9] Chen T., Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acmsigkdd international conference on knowledge discovery and data mining; ACM;2016. p. 785-794.

[10] Araya D. B., Grolinger K., ElYamany H. F., Capretz M. A., Bitsuamlak G. An ensemble learning framework for anomaly detection in building energy consumption. ENERG BUILDINGS.2017; 144:191-206.

[11] Cho J., Kim Y., Koo J., Park W. Energy-cost analysis of HVAC system for office buildings: Development of a multiple prediction methodology for HVAC system cost estimation. ENERG BUILDINGS.2018.

[12] Deng H., Fannon D., Eckelman M. J. Predictive modeling for US commercial building energy use: A

comparison of existing statistical and machine learning algorithms using CBECS microdata. ENERG BUILDINGS.2018; 163:34-43.
[13] Jain R. K., Smith K. M., Culligan P. J., Taylor J. E. Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. APPL ENERG.2014; 123:168-178.
[14] Naganathan H., Chong W. O., Chen X. Building energy modeling (BEM) using clustering algorithms and semi-supervised machine learning approaches.AUTOMAT CONSTR.2016; 72:187-194.
[15] Rahman I., Kuzlu M., Rahman S. Power disaggregation of combined HVAC loads using supervised machine learning algorithms. ENERG BUILDINGS.2018; 172:57-66.
[16] Singaravel S., Geyer P., Suykens J. Component-based machine learning modelling approach for design stage building energy prediction: weather conditions and size. In: Proceedings of the 15th IBPSA conference; 2017. p. 2617-2626.
[17] Candanedo L. M., Feldheim V., Deramaix D. Data driven prediction models of energy use of appliances in a low-energy house. ENERG BUILDINGS.2017; 140:81-97.
[18] Nayak S. C. Escalation of Forecasting Accuracy through Linear Combiners of Predictive Models. EAI, Scalable Information Systems.2019.
[19] Sethi J. S., Mittal M. Ambient Air Quality Estimation using Supervised Learning Techniques. Scalable Information Systems.2019.
[20] Tsanas A., Xifara A. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. ENERG BUILDINGS.2012; 49:560-567.
[21] Krzywinski M., Altman N. Multiple linear regression: when multiple variables are associated with a response, the interpretation of a prediction equation is seldom simple. Nat. Methods.2015; 12(12):1103-1105.
[22] Nimon K. F., Oswald F. L. Understanding the results of multiple linear regression: Beyond standardized regression coefficients. Organ Res Methods.2013; 16(4):650-674.
[23] Peng Y. A novel ensemble machine learning for robust microarray data classification. COMPUT BIOL MED.2006; 36(6):553-573.
[24] Mendes-Moreira J., Soares C., Jorge A. M., Sousa J. F. D. Ensemble approaches for regression: A survey. ACM COMPUT SERV.2012;45(1): 10.
[25] Prasad A. M., Iverson L. R., Liaw A. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems.2006; 9(2):181-199.
[26] Breiman L. Random forests. Machine learning.2001; 45(1):5-32.
[27] Cutler A., Cutler D. R., Stevens J. R. Random forests. In Ensemble machine learning. Springer, Boston, MA.2012;157-175.
[28] Meir R., Rätsch G. An introduction to boosting and leveraging. In Advanced lectures on machine learning. Berlin, Heidelberg: Springer; 2003.p. 118-183.
[29] Friedman J. H. Stochastic gradient boosting. COMPUT STATDATA AN.2002; 38(4):367-378.
[30] Sosvilla-Rivero S., Rodríguez P. N. Linkages in international stock markets: evidence from a classification procedure. APPL ENERG.2010; 42(16):2081-2089.
[31] Fan C., Xiao F., Wang S. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. APPL ENERG.2014; 127:1-10.
[32] https://archive.ics.uci.edu/ml/datasets/Energy+efficiency accessed on 01/04/2019