International Journal of Energy, Information and Communications
Vol.11, No.1 (2020), pp.7-14
http://dx.doi.org/10.21742/ijeic.2020.11.1.02

Industry Energy Consumption Prediction Using Data Mining Techniques

Sathishkumar V E, Jonghyun Lim, Myeongbae Lee, Kyeongryong Cho, Jangwoo Park, Changsun Shin, and Yongyun Cho*

Dept. of Information and Communication Engineering, Suncheon, Sunchon National University, Korea
**yycho@scnu.ac.kr*

Abstract

Predicting energy consumption is an essential part of the electricity company supply. This paper presents and explores energy consumption prediction models using data mining approach for the steel industry. DAEWOO steel industry energy consumption data is used in this study. Data used include lagging and leading current reactive power, lagging and leading current power factor, carbon dioxide (tCO_2) emission, and load types. The prediction models are trained with its best hyperparameters selected using repeated cross-validation and are evaluated using a test set: (a) General Linear Regression, (b) Classification and Regression Trees (c) Support Vector Machine with Radial Basis Kernel (d) K Nearest Neighbor, (e) Random Forest. Four evaluation indices such as Root Mean Squared Error, Mean Absolute Error, Mean Absolute Percentage Error and Coefficient of Variation are used to measure the prediction efficiency of regression models. The results show that the Random Forest model can best predict energy consumption and outperforms other conventional algorithms in comparison.

Keywords: *Energy consumption, Data mining, Predictive analytics, Data analysis*

1. Introduction

Due to the steady rise in population, energy is deemed one of the most significant and scarce resources. Energy saving is sought not only for advancing a green environment for future efficiency but also for domestic consumers and energy production companies. The rapid economic growth adds to the emerging call for electrical power energy worldwide. In the meantime, electricity is deemed to be among the most fundamental pushing element of economic growth and is held vital in our domestic lives [1]. Consequently, for a Nation or region, the forecast of electrical energy usage becomes pressing and crucial [2]. Many countries are continuing research to optimize the energy consumption and deliver electricity in a consistent way. The energy consumption sectors are classified as Industrial, Transportation, Residential and Commercial. Out of these four sectors, industrial sector energy consumption is comparatively higher than other sectors. So there arises a need to control energy consumption mainly in the Industrial sector.

Meanwhile, Third Scientific and Technological Revolution's constructive accomplishments made life simpler for people and boosted technological innovation and structural

Article history:
Received (October 12, 2019), Review Result (November 16, 2019), Accepted (December 28, 2019)

Print ISSN: 2293-9655, eISSN: 2652-1989 IJEIC

transformation in traditional industries. The manufacturing industry is the vital sector and the key index of a nation or region's financial level. Many developed nations already have advanced manufacturing sectors, but they proceed to pursue fresh possibilities and redesign their manufacturing industries to guarantee an unconquerable place in the presence of technological growth and modernization. A typical instance is Germany, as its "Industry 4.0" relies on smart growth, with a focus on output effectiveness, material use, and energy consumption [3].

Since the 1990s, the manufacturing industry in South Korea has proceeded to evolve at an elevated speed and has become the primary leading power of the South Korean economy's ongoing fast growth. Primary energy usage increased in the 1990s at an annualized pace of 7.5%, greater than the annualized financial development rate of 6.5% in the same time. This is due to the sharp progression in energy-intensive industries, including petrochemical industries. The steep rise in industrial electricity consumption facilitated the hike in loss of energy conversion, further undermining energy intensity. The rise in energy-industry production after 2009 considerably buffered the nation in resistance to the global economic crisis, but it adversely affected the gross energy efficiency of the nation [4]. Many uncertain factors, such as industrial structure, technology level, energy price, economic scale and national policy, influence the energy consumption of the industries. Considering the need for optimizing and managing the energy consumption within Industrial sector, this paper suggests a data mining based approach for Industry energy consumption.

The paper is further structured as follows. Section 2 offers a thorough overview of the literature, followed by Section 3 which discusses the recorded data and description. Section 4 details the results and discussion. Section 6 concludes the paper.

2. Related works

Several statistical and artificial intelligence techniques are developed to model the patterns of energy consumption. Recently, scientists examined the prediction of energy consumption using Machine learning and Data Mining techniques. Data Mining approaches are very useful and convenient to use by normal users in the field and are more popular in many applications hence used in this research [5].

To predict the electricity requirement, researches typically use traditional algorithms such as multiple regression [6], support vector machines [7], support vector machines with Principal Component Analysis [8], time series techniques [9], Gradient Boosted Machines [10], Artificial Neural Networks [11] and Deep Learning Methods [12]. The commonly regarded parameters or features for energy consumption include period of the day, outside temperature, month, weekend, holidays, utilization of yesterday electricity consumption, global solar radiation, rainfall index, wind speed and occupancy details [13][14]. Since all patterns of energy consumption are generally linked with certain modes of time cycles, the characteristics obtained from date and time stamp are favorable in designing predictive models.

A study shows that a Tree-based model called Classification and Regression Trees (CART) gives a precise forecast of building energy demand with minimum error [15]. This technique classifies and foresees categorical factors with an interpretable flowchart like tree structures allowing consumers to obtain helpful data quickly. As diverse CART models are produced and used as foundation models, Random Forest can be regarded as an ensemble of multiple Classification and Regression Tree (CART). RF gains considerable focus and in many areas, becomes a popular prediction algorithm. For instance, to a Classification problem in gene

selection a new method was proposed using Random Forest [16]; Culter et al. [17] presented Random Forest in the domain of ecology to distinguish various plant species; Rodriguez-Galiano et al. [18] used Random Forest in the domain of land remote sensing to distinguish varying land coverage and Sun et al. [19] used Random Forest to forecast solar radiation from various areas. The findings of this study shows Random Forest model favorable efficiency in tackling issues of classification and regression. Besides, earlier research programs compare Random Forest to other prominent methods such as linear discriminant analysis [20], decision tree [21], and Support Vector Machine [22]. The study findings indicate that Radom Forest outperformed these challengers in fixing the study issues, showing their ability as a promising tool for solving the issue of building energy prediction.

Although the findings of this literature are useful, the drawback is that the discussion is comparatively macroscopic, and there is a lack of research in Industry sector. This article centres on South Korea's energy consumption of Small scale steel industry. The above researches accept that all models of Machine learning performs sufficiently in predicting patterns of energy consumption. This work aims to comprehend the links between industrial energy consumption with varying predictors. As well to explore the efficiency of various machine learning models namely General Linear Regression (GLR), Classification and Regression Trees (CART), K Nearest Neighbor (KNN), Support Vector Machine with Radial Basis Kernel (SVM) and Random Forest (RF) in predicting energy consumption of steel industry.

3. Recorded data and description

The information is gathered from DAEWOO steel.co. Ltd in Gwanyang, South Korea. It produces several types of coils, steel plates and iron plates. The information on electricity consumption is stored in a cloud-based system and made available on the website, Korea Electric Power Corporation (pccs.kepco.go.kr) and the detailed insights about on daily, monthly and yearly energy consumption patterns are computed and displayed.

Table 1. Data variables

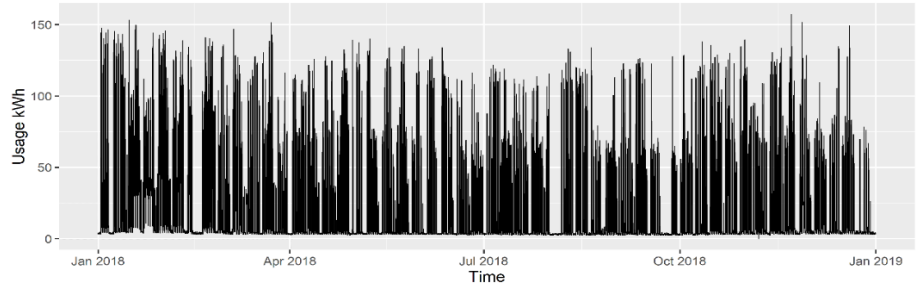
Data Variables	Type	Measurement
Industry Energy Consumption	Continuous	kWh
Lagging Current Reactive Power	Continuous	kVarh
Leading Current Reactive Power	Continuous	kVarh
tCO2(CO2)	Continuous	Ppm
Lagging Current Power Factor	Continuous	%
Leading Current Power Factor	Continuous	%
Number of Seconds from Midnight	Continuous	S
Week status	Categorical	(Weekend (0) or a Weekday(1))
Day of week	Categorical	Sunday, Monday …. Saturday
Load Type	Categorical	Light Load, Medium Load, Maximum Load

This research focuses on the electricity (kWh) data recorded for the industry every 15min. The reporting interval of 15min is selected to track the rapid energy consumption variations. Data timespan is 365 days (12 months). Data stored in the website include electricity consumed, lagging and leading current reactive power, lagging and leading current power factor, carbon dioxide (tCO_2) emission, and load types. Load types vary across different time and months. Three types of Load categories such as Light, Medium and Maximum are used. During Light load timing, the energy consumption is less and during maximum load the energy consumption is high. When medium load, the energy consumption range is between light and maximum load. Weather data and occupancy details are used as predictors to predict energy consumption [35][36]. But in this research, the temperature variables have no impact on energy consumption as the steel industry is in open space and has no heaters or cooling systems. Table 1 shows the full overview of the dataset with its corresponding data type and measurement. The date/time variable produces other additional characteristics: the number of seconds from midnight for each day (NSM), week status and Day of the week [19]. This process of creating additional variables from existing variables known as Feature Engineering.

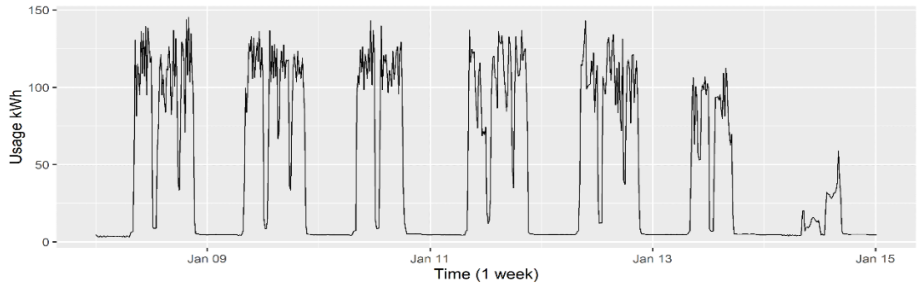
Table 2. Training and Testing Dataset Dimensions

Dataset	Number of observations
Training	26281 and 10 variables
Testing	and 10 variables

The full one-year data set is divided randomly into training and test validation dataset. Training the models uses 75% of the data and the remaining 25% is used for testing purposes. Table 2 displays the dimensions of the training and testing set.



(A)



(B)

Figure 1. (A) Industry Energy Consumption measurement for the whole period; (B) Industry Energy Consumption for the First week

[Figure 1] (A) shows the plot for industry energy consumption for the total period of 1 year and [Figure 1] (B) shows the industry energy consumption for the first one week. As can be seen from the figures that the industry energy consumption shows high variability. It can be also noted that from [Figure 1] (B) that industry energy consumption is high during weekdays and less during the weekends.

4. Results and discussion

All the models are developed using the best hyperparameters selected using the Grid Search with repeated 10-fold cross-validation. Four evaluation indices such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Coefficient of Variation (CV) are used to measure the performance of the regression models. The models with lower values of RMSE, MAE, MAPE and CV is considered as the best performing model.

[Table 3] provides the performance of the trained regression models in the training and testing sets. As can be seen from the Table that the RF has the smallest values of RMSE, MAE, MAPE and CV values. Other than RF, SVM performance is good compared to other models. Out of all the models chosen in this research, the GLM model produce highest values of RMSE, MAE, MAPE and CV values in the test set. This shows the data is not linearly related to any of the independent variables. It can be also noted that ensemble strategies applied to multiple CART made the RF model to perform the best and so CART model is used for comparison in this study.

Although the results show RF model performance is good, RF also suffers the instability problem in testing and training process. The performance of RF in testing phase is more than double of the error rate in the training phase. Still the performance of RF is best with reduced metrices such as RMSE, MAE, MAPE and CV. Also, the performance has been improved by RF compared with conventional algorithms: GLM, CART, SVM and KNN [Table 3]. As a powerful tool for modelling ensemble-based models, RF is capable of balancing the relationship between the accuracy in prediction and the intelligibility requirements.

Table 3. Models performance

Models	Training				Testing			
	RMSE	MAE	MAPE	CV	RMSE	MAE	MAPE	CV
GLM	4,61	2.53	16.01	16.87	4.85	2.56	16.13	17.67
CART	3.27	1.98	13.88	11.99	3.46	2.04	14.10	12.59
SVM	1.89	1.51	4.67	6.92	1.97	1.71	4.87	7.16
KNN	1.59	0.75	2.89	5.81	2.99	1.75	5.33	10.90
RF	0.50	0.14	2.45	1.85	1.12	0.36	1.28	4.09

5. Conclusion

This study focused on predicting energy consumption of a small-scale steel industry. The main purpose of this research is to determine the best performing regression algorithm to best predict the energy consumption with reduced error. Results indicate that the RF model enhances RMSE, MAE, MAPE and CV values of prediction relative to other regression models such as GLM, CART, SVM and KNN. The analysis is conducted for just one industry, which is one of this study's primary limitation. Analyzing several industries can lead to

several significant information. Additional features with the energy consumption of the industry can be explored in association with the products they produce, the geometry of the building, sort of equipment, etc. In the process of both the exploratory analysis and developing prediction models, the data analysis shows thought-provoking outcomes. Future work could include predicting the individual machine's energy consumption for optimizing the energy consumed by each machine in the steel industry and to develop IoT based technology to optimize the energy consumption.

Acknowledgements

This work is supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry & Energy (MOTIE) of the Republic of Korea (No. 20172010000730).

References

[1] Xiao L., Shao W., Liang T., and Wang C., "A combined model based on multiple seasonal patterns and modified firefly algorithm for electrical load forecasting," Applied Energy, vol.167, pp.135-153, **(2016)** DOI: 10.1016/j.apenergy.2016.01.050

[2] Wang Z.X., Li Q., and Pei L.L., "A seasonal GM (1, 1) model for forecasting the electricity consumption of the primary economic sectors," Energy, vol.154, pp.522-534, **(2018)** DOI: 10.1016/j.energy.2018.04.155

[3] Schwab Klaus, The fourth industrial revolution, Currency, **(2017)**

[4] Lee Seung-moon, "Mid-term Korea Energy Demand Outlook," Korea Energy Economics Institute, May **(2014)**

[5] Simeone O., "A very brief introduction to machine learning with applications to communication systems," IEEE Transactions on Cognitive Communications and Networking, vol.4, no.4, pp.648-664, **(2018)** DOI: 10.1109/TCCN.2018.2881442

[6] Candanedo L.M., Feldheim V. and Deramaix D., "Data driven prediction models of energy use of appliances in a low-energy house," Energy and buildings, vol.140, pp.81-97, **(2017)** DOI: 10.1016/j.enbuild.2017.01.083

[7] Cherkassky V. and Ma Y., "Practical selection of SVM parameters and noise estimation for SVM regression," Neural networks, vol.17, no.1, pp.113-126, **(2004)** DOI: 10.1016/S0893-6080(03)00169-2

[8] Jinhu L., Xuemei L., Lixing D., and Liangzhong J., "Applying principal component analysis and weighted support vector machine in building cooling load forecasting," In 2010 International Conference on Computer and Communication Technologies in Agriculture Engineering, vol.1, pp.434-437, **(2010)** DOI: 10.1109/CCTAE.2010.5543476

[9] Alsharif M.H., Younes M.K. and Kim J., "Time series ARIMA model for prediction of daily and monthly average global solar radiation: The Case Study of Seoul," South Korea, Symmetry, vol.11, no.2, p.240, **(2019)** DOI: 10.3390/sym11020240

[10] Candanedo Luis M., Véronique Feldheim, and Dominique Deramaix, "Data driven prediction models of energy use of appliances in a low-energy house," Energy and buildings, vol.140, pp.81-97, **(2017)** DOI: 10.1016/j.enbuild.2017.01.083

[11] Li K., Su H., and Chu J., "Forecasting building energy consumption using neural networks and hybrid neuro-fuzzy system: A comparative study," Energy and Buildings, vol.43, no.10, pp.2893-2899, **(2011)** DOI: 10.1016/j.enbuild.2011.07.010

[12] Mocanu E., Nguyen P.H., Gibescu M., and Kling W.L., "Deep learning for estimating building energy consumption," Sustainable Energy, Grids and Networks, vol.6, pp.91-99, **(2016)** DOI: 10.1016/j.segan.2016.02.005

[13] C. Sandels, J. Widén, L. Nordström, and E. Andersson, "Day-ahead predictions of electricity consumption in a Swedish office building from weather, occupancy, and temporal data," Energy Build, vol.108, pp.279-290, **(2015)** DOI: 10.1016/j.enbuild.2015.08.052

[14] P. Bacher, H. Madsen, H.A. Nielsen, and B. Perers, "Short-term heat load forecasting for single family houses," Energy Build, vol.65, pp.101-112, **(2013)** DOI: 10.1016/j.enbuild.2013.04.022

[15] Yu Z., Haghighat F., Fung B.C., and Yoshino H., "A decision tree method for building energy demand modeling," Energy and Buildings, vol.42, no.10, pp.1637-1646, **(2010)** DOI: 10.1016/j.enbuild.2010.04.006

[16] Díaz-Uriarte R. and De Andres S.A., "Gene selection and classification of microarray data using random forest," BMC bioinformatics, vol.7, no.1, p.3, **(2006)** DOI: 10.1186/1471-2105-7-3

[17] Cutler D.R., Edwards Jr T.C., Beard K.H., Cutler A., Hess K.T., Gibson J., and Lawler J.J., "Random forests for classification in ecology," Ecology, vol.88, no.11, pp.2783-2792, **(2007)** DOI: 10.1890/07-0539.1

[18] Rodriguez-Galiano V.F., Ghimire B., Rogan J., Chica-Olmo M., and Rigol-Sanchez J.P., "An assessment of the effectiveness of a random forest classifier for land-cover classification," ISPRS Journal of Photogrammetry and Remote Sensing, vol.67, pp.93-104, **(2012)** DOI: 10.1016/j.isprsjprs.2011.11.002

[19] Sun H., Gui D., Yan B., Liu Y., Liao W., Zhu Y., Lu C. and Zhao N., "Assessing the potential of random forest method for estimating solar radiation using air pollution index," Energy Conversion and Management, vol.119, pp.121-129, **(2016)** DOI: 10.1016/j.enconman.2016.04.051

[20] Cutler D.R., Edwards Jr T.C., Beard K.H., Cutler A., Hess K.T., Gibson J., and Lawler J.J., "Random forests for classification in ecology," Ecology, vol.88, no.11, pp.2783-2792, **(2007)** DOI: 10.1890/07-0539.1

[21] Rodriguez-Galiano V.F., Ghimire B., Rogan J., Chica-Olmo M., and Rigol-Sanchez J.P., "An assessment of the effectiveness of a random forest classifier for land-cover classification," ISPRS Journal of Photogrammetry and Remote Sensing, vol.67, pp.93-104, **(2012)** DOI: 10.1016/j.isprsjprs.2011.11.002

[22] Khalilia M., Chakraborty S. and Popescu M., "Predicting disease risks from highly imbalanced data using random forest," BMC medical informatics and decision making, vol.11, article no.51, **(2011)** DOI: 10.1186/1472-6947-11-51

This page is empty by intention.