A Comparative Study of Energy Big Data Analysis for Product Management in a Smart Factory

Rahman A. B. M. Salman, Sunchon National University, South Korea

Lee Myeongbae, Sunchon National University, South Korea

Lim Jonghyun, Sunchon National University, South Korea

Yongyun Cho, Sunchon National University, South Korea

Shin Changsun, Sunchon National University, South Korea

ABSTRACT

Energy is one of the key inputs for a country's economic growth and social development. Analysis and modeling of industrial energy are currently time-intensive processes because more and more energy is consumed for economic growth in a smart factory. This study aims to present and analyse the predictive models of the data-driven system to be used by appliances and find the most significant product item. With repeated cross-validation, three statistical models were trained and tested in a test set: 1) general linear regression model (GLM), 2) support vector machine (SVM), and 3) boosting tree (BT). The performance of prediction models were measured by R2 error, root mean squared error (RMSE), mean absolute error (MAE), and coefficient of variation (CV). The best model from the study is the support vector machine (SVM) that has been able to provide R2 of 0.86 for the training data set and 0.85 for the testing data set with a low coefficient of variation, and the most significant product of this smart factory is Skelp.

KEYWORDS

Big Data Analysis, Boosting Tree, Correlation, Data Mining, Energy Consumption, General Linear Regression, Principal Component Analysis, Product Management, Support Vector Regression

INTRODUCTION

Energy is the most significant and vital requirement for all living things on earth to survive and grow. Energy has been seen as one of the key inputs for a country's economic growth and social development. The rise of industrialization raises energy demand, which is a critical component of national strategy. Moreover, energy consumption rises in tandem with economic development and human progress. Nowadays, more and more energy is being used for economic growth and population growth. Facilities of industrial customers and the use of electricity to process various types of machinery, manufacture or assemble products, include such diverse industries as production, mining, and construction (Liye Xiao et al., 2016). Ultimately, more than one-third of electrical energy is used by those industrial sectors from total energy for a country. As a result, they begin to collect huge databases in order to gather valuable information. Authorities plan to use this data to improve the industry's standards

DOI: 10.4018/JOEUC.291559

and long-term viability. As a result, industrial authorities optimize the utilization of diverse energy resources to minimize energy consumption expenses.

The industrial sector is one of the primary sectors that need energy stability. Since the 1990s, South Korea's manufacturing industry has continued to expand rapidly and has become the main driving force of South Korean economies. Primary energy consumption rose at an annual rate of 7.5% in the 1990s, which in the same period was higher than the annual economic growth rate of 6.5%. This was due to the rapid growth of energy-intensive factories and petrochemical industries as well. The sharp increase in industrial electricity consumption helped increase energy conversion loss, further reducing the energy intensity (Kim et al., 2001). The increase in energy industry output after 2009 greatly buffered the nation against the global financial crisis, but it negatively affected the overall energy performance of the country. Several unpredictable factors influence the energy usage of industries, such as the nature of the market, the level of technology, energy costs, economic size, and national policy.

The achievements of the third scientific and technological revolution have simplified life. Industrial production is a significant sector for both the country and the nation, and it acts as a major financial indicator. In traditional sectors, it has boosted new technologies and systemic transformation. In traditional industries, the fruitful successes of the third scientific and technical revolution have encouraged people's lives and promoted technological progress and institutional change. The production sector is a vital industry and a primary predictor of a nation or region's economic level. Many advanced manufacturing countries already have advanced industries. Still, they continue to explore new opportunities and overhaul their manufacturing industries in order to ensure an unstoppable role in the face of modernization and technological growth. Germany is a common example, as the 'Industry 4.0' focuses on intelligent growth, emphasizing product quality, resource use, and energy use (Liye Xiao et al., 2016).

Many studies have shown that improving energy efficiency is very important for economic growth (David G.Ockwell., 2008, Chirs Bataille et al., 2017). The relationship between economic development, trade, and resources in Asia was examined by Nasreen and Anwar (Nasreen et al., 2014), and they found that economic growth and trade transparency had a positive effect on the use of energy. While several researchers have confirmed the one-way relationship between economics and energy (Lee, C.C., 2005, Tasni, S.Z., 2010), others have shown a two-way relationship (Cheng et al., 2004, Stern, D.I.A., 2000). Industrial factory owners are also beginning to realize that analysing and forecasting the energy data with the production data is very important for the benefit of their companies or plants. This issue is caused by unregulated energy use, such as overconsumption, weak systems, and waste energy. Energy is regarded as one of the most important and precious resources due to the continual growth in demand. It is imperative to engage with the management board of industrial companies as a supporting technical hand to enhance their energy usage. (A.B.M. Salman Rahman et al., 2019).

A Smart Factory is an IoT idea that envisions a production process as a completely automated and intelligent network of technologies that allows facilities, equipment, and logistical chains to be operated without the need for human interaction. Furthermore, a smart factory is a location where all of these things occur due to data interchange between all elements in the production technology chain, not only between production tools and machines (Peng Lin et al., 2018). The above drives turn machine learning, allowing operations to run more effectively and save money than they might if humans only supervised manufacturing processes. Data collection and analysis are fundamental in the smart factory concept because it unlocks the potential buried in equipment, resources, and people. Data may reach the appropriate location in the production chain at the right moment in the smart factory without the intervention of a human supervisor. This is for a more rapid interaction paradigm in which machines and tools communicate information to achieve better efficiency (Tongtong Geng et al., 2020). Data from various production settings must be collected, combined, and analysed to yield valuable insights to achieve better efficiency. The components of a smart factory are depicted in Figure 1. The Internet of Things (IoT) is the driving force behind smart factories, which connect smart devices and sensors with the factory to make industrial operations data-driven and data-enabled.

The progress of industrialization raises energy demand, which is a critical component of national strategy. Additionally, energy consumption rises in tandem with economic development and human advancement. This paper aims to present and analyse the predictive models of a data-driven system used by appliances to find out the best prediction model and determine the most important product item using principal component analysis of the Daewoo steel factory in South Korea to improve an industrial factory's energy utilization rate.

The remaining part of the paper is laid out as follows. The second section is devoted to related works. The methodology for providing services is described in Section 3. Section 4 describes the recorded energy and production statistics. The exploratory analysis is described in Section 5. Model selection is covered in Section 6. Evaluation indices are presented in Section 7. Results and discussion is covered in Section 8, and the paper is concluded up in Section 9.

Related Works

The expression "Internet of Things" is defined by a variety of professional and scientific research sources. Kevin Ashton, co-owner and CEO of Auto-ID Center, coined the term "internet of things" in 1999. Various professional standards agencies, organizations, and associations in IK technologies and numerous scholars have defined the idea of IoT as it has evolved and expanded in use (Ivan Cvitić et al.,2021).

Many researchers have long looked at the subject of predicting energy usage using data-mining techniques. Several statistical and Artificial Intelligence (AI) approaches have been developed to estimate energy usage trends. Machine learning algorithms are beneficial and convenient for a normal operator to utilize after constructing the model; they are increasingly common in various applications (Sathishkumar V E et al., 2020).

Since it is commonly acknowledged that big data opens up new business prospects, many companies are working to build and improve their big data analytics capabilities in order to uncover and better comprehend the valuable information hidden in their own data sources. The notion of continual development is reflected in these enormous databases gathered from big industrial locations. The problem is that as big data volumes grow, data becomes more sophisticated and diverse, necessitating advanced algorithms. This enables for changes to be analyzed, perhaps exposing previously hidden and undiscovered viewpoints that might help enhance industrial operations (Esa Hämäläinen et al.,2019).

Researchers have recently explored that the use of machine learning techniques for predicting energy consumption. Several analytical and artificial intelligence approaches have been developed to model the trends in energy consumption. Machine learning algorithms are very convenient and easy to use after the model is established by an ordinary operator and are more popular in many applications (O. Simeone., 2018). Reducing the use of electricity in the steel industry is a global issue where actions are taken vigorously by the government. If modeling and forecasting demand is practical, then a steel plant can do better control on uses of energy (Chen et al., 2019). Conservation of energy is a vital challenge since the three major economic industries, manufacturing, transport, and development, have heavy energy consumption. The heating load calculation is the first step of the construction process for the HVAC system (Chou et al., 2014).

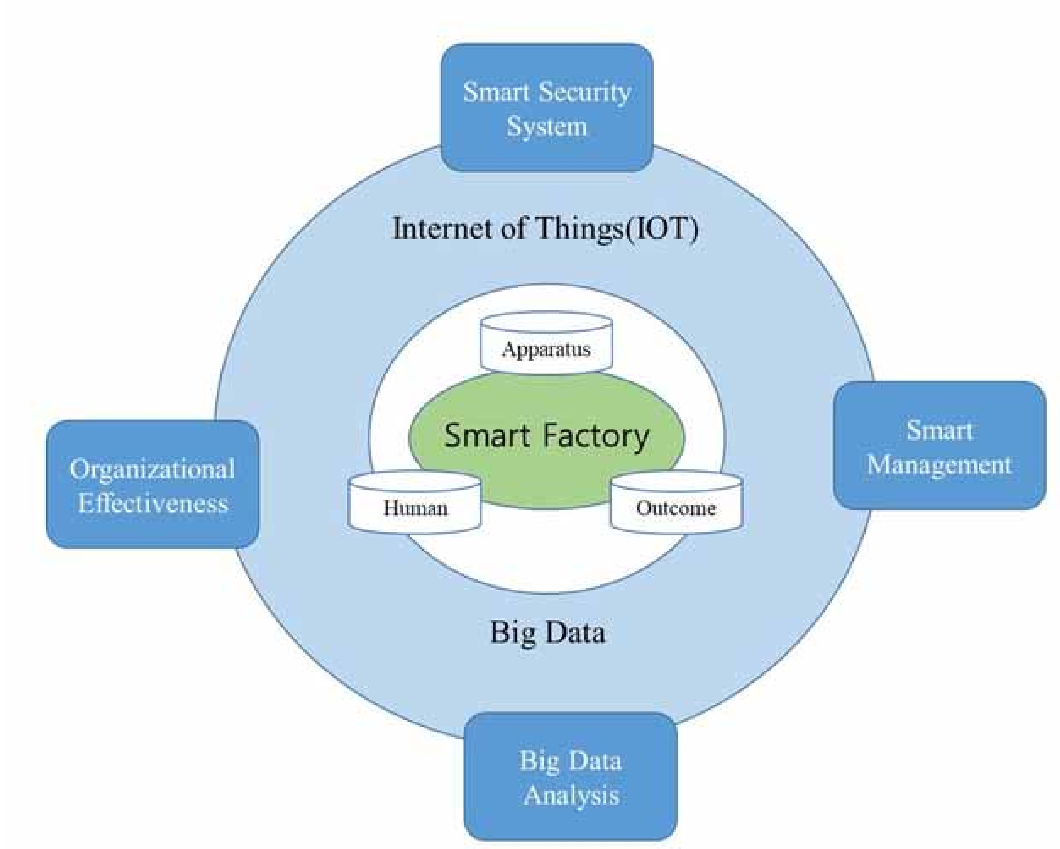
Electrical equipment and machinery dominated the overall structure and strength effect and the sub-sectors of the raw chemical materials and components. The sectors of gas and petroleum refining and coking production and supply contributed the least (Zha et al., 2009). Technological innovation has provided large opportunities for researchers in diverse fields to use artificial intelligence. Various attempts have been made in the manufacturing and development fields to use machine learning methods (Paturi et al., 2020).

Production is critical to the global economy's success. However, natural resources have been quickly exhausted in recent years, resulting in a slew of environmental and societal issues, primarily due to the expansion of the industry. According to reliable data, the industry sector utilizes more energy than other sectors, accounting for roughly 37% of total supplied energy globally. Manufacturing consumes a significant amount of energy in the industrial sector (Fei Tao et al., 2016).

Building Information Modeling is mostly pushed and employed in the cost management of civil and public buildings, but it is rarely employed in the power construction business. Many professions are involved in the electric power construction project and many types of buildings and structures. These characteristics, when combined with Chinese particular quantity computation regulations, resulted in certain roadblocks. When combined with Chinese particular quantity computation standards, these variables resulted in certain challenges when it came to BIM implementation (Rui Lui et al., 2016). Xiaofei has published a paper entitled a modern identification system based on enhanced Spatio-temporal functionality and AdaBoost-SVM classifiers (Xiaofei et al., 2015).

From automated production to "smart manufacturing," the manufacturing industry is undergoing a fundamental change. During this progression, the Internet of Things (IoT) plays a critical role in linking manufacturing's physical environment to the cyberspace of computing platforms and decision-making algorithms, resulting in the formation of a Cyber-Physical System (CPS) (Hong-Ning Dai et al.,)

Figure 1. Components of the smart factory.



Methodology

This study uses three different models to find out the best prediction model for industrial factories. Three methods are 1) General Linnear Regression model, 2) Support Vector Regression Model, and 3) Boosting Tree. In this part, we describe these three regression models in detail step by step.

General Linear Regression

In order to evaluate the association among a scalar dependent variable y as well as one or several independent variables X, linear regression is known as a method of modeling. Simple linear regression is understood to be the case with single explanatory parameters. Multiple linear regression is considered the technique with two or more explanatory variables (A.B.M. Salman Rahman et al., 2018). Data in linear regression is built using linear predictor functions, and unspecified model parameters are determined from the data. Linear regression models are called prototypes (Achilles D. Boursianis et al., 2020).

Linear regression is associated with the nursing method for evaluating the association between the scalar dependent quantity parameters y and one or several independent quantity parameters X. The general linear regression model (GLM) in statistics is essentially a versatile generalization, like its basic simple linear regression model (GLM) (Jui-Sheng et al., 2014) that also implies a uniform distribution of the data points.

Assume model,

$$G\left(M\left(y\right)\right) = x_j \times \beta_0 + \mathrm{O} > \mathrm{y} \sim \mathrm{K}$$

Here,
G(M) = Identified relation function,
β_0 = Regression coefficient,
x_j = Predictor Variable
y = Predicted output,
O = Variable offset
K = Model of y distribution.

The GLM employs the newton-raphson method to achieve a continual approximation such that ($x \times \beta + O$) approaches g(E(y)). The posterior final equation is developed mostly as the relational representation of a(X-Y). Although additional GLM parameters improve model variance, GLM has a wider implementation scope and a more functional interaction model than Linear Regression.

Support Vector Regression

The area unit supports vector machines (SVMs) in data science and supervises learning models with related learning algorithms analyzing information. Support Vector Machine has also been used as the regression technique, holding with most options characterizing the formula (maximum margin) (Bernhard Schölkopf et al., 1995).

Support Vector Regression (SVR) uses a similar concept for classification due to SVM with only a few minor variations. First, since the performance is an actual number, estimating the knowledge at hand becomes very difficult, which has endless possibilities. There is a margin of tolerance (epsilon) for the regression scenario, which approximates the SVM, which may have been previously asked from the matter. An additional complex interpretation is also available, but the formula is more difficult to understand (Bing Dong et al., 2005). Normally, the primary strategy is still the same: to minimize error and individualize the hyperplane that maximizes the margin, considering that a portion of the error is tolerated.

Equation of SVR;

$$f\left(x\right) = \sum_{i=1}^{m}\left(\alpha_i^* - \alpha_i\right)k\left(x_i, x\right) + b$$

Throughout this situation, the cost of K(x_i,x) in the function space ф(xi) and ф(xj) is similar to the outer combination of the two vectors xi and xj, K(xi,xj) = ф(xi)·ф(xj). Indeed, all the necessary simulations can be done directly in the function space and do not need the ф (x) map to be measured using kernels. K(xi,xj) = xi·xj the linear kernel, K(xi,xj) = (xi·xj + 1)d the polynomial kernel and the radial base function(RBF) kernel tend to be several common kernel functions,

Equation is,

$$k\left(x_i, \gamma_i\right) = \exp\left(-\gamma x_i - x_j^2\right), \gamma > 0$$

Where the kernel parameters are respectively d and γ (Gareth James et al., 2013), there are numerical benefits to the radial centered function, and it was used in that analysis. The radial based approach has numerical advantages and has been used in the analysis

Boosting Tree

Boosting Tree is a machine learning algorithm for classification and regression challenges and produces a simulation function in the form of such an ensemble of strong estimation methods. Boosting Trees (BT) is an exponential regression model that consists of an aggregate of decision trees. A standard decision tree has the issue of over-fitting, but the BT algorithm can overcome this by combining hundreds of poor decision trees consisting of several leaf nodes (Jerome H. Friedman., 2002).

The Boosting Trees (BT) is also known as blackboost. To optimize conditional loss functions where regression trees are used as simple learners called the Gradient boost. This approach follows the traditional gradient boosting of foundation learners using regression trees. The primary distinction between GBM and BT is that the family claim to blackboost will assign arbitrary loss functions to be optimized, whereas GBM utilizes a hard-coded loos feature. In comparison, the foundation learners (conditional inference trees) are a bit more stable. The regression design seems to be a projection engine for a data recorder and is thus barely interpretable. In this table 1, all of the variables are included along with their explanations.

Assume the model,

$$F_m\left(X\right) =)F_{m=1}\left(X\right) + v\sum_{i=1}^{m} Y_m l\left(x \in R_{jm}\right)$$

RECORDED ENERGY AND PRODUCTION DATA DESCRIPTION

This section describes the energy and smart factory production data used in this study. This research analyzes hourly energy use data to develop prediction models utilizing machine learning methods like GLM, SVM RBF, and BT. There are two types of data set available in this study gained from DAEWOO Steel Co. Ltd in Gwangyang, South Korea. Between these two data sets, one data set is for the energy consumption of the steel factory and the other data set for the productions of the steel factory. On the Korea Electric Power Corporation's website (pccs.kepco. - go.kr), data on industry energy usage is recorded, and daily, monthly, and yearly data views are calculated and shown. This study focuses on the industry's energy (kWh) statistics for every one hour. A one-hour reporting interval has been chosen to capture the fast fluctuations in energy use. This factory manufactured different types of steel, rods, and plates, and the manufacturing products names are "Sheet," "Skelp," and "Cyongkg." For the year 2017, the data collection period is 365 days (or 12 months). R has used to do all of the data analysis.

Table 1. Notation table

System of Symbols	Description
$g(\bullet)$	Link function
x	Predictor
y	Response variable
o	Offset variable
β	Regression coefficient
i and j	Dependent field categories
y_i	Actual measurement (energy consumption),
$\hat{Y}_j$	Predicted value from the regression algorithms
n	The number of measurements.
$\overline{y}$	Average energy consumption

Figure 2 depicts the suggested design's overall flow. In smart factory flow charts, energy consumption for manufacturing products and other uses of energy is measured by IoT-based smart meters. Moreover, data related to energy use is coming to the structural and data modeling stage. The structured and data modeling stages are pre-processed data for data modeling, and from this stage, energy data is saved and stored in the cloud system. Production data is also collected and stored on the same cloud-based system. Big data analytics algorithms are employed to estimate demand, which is used for successful energy policy and product management.

Figure 2. The suggested system's overall flow chart

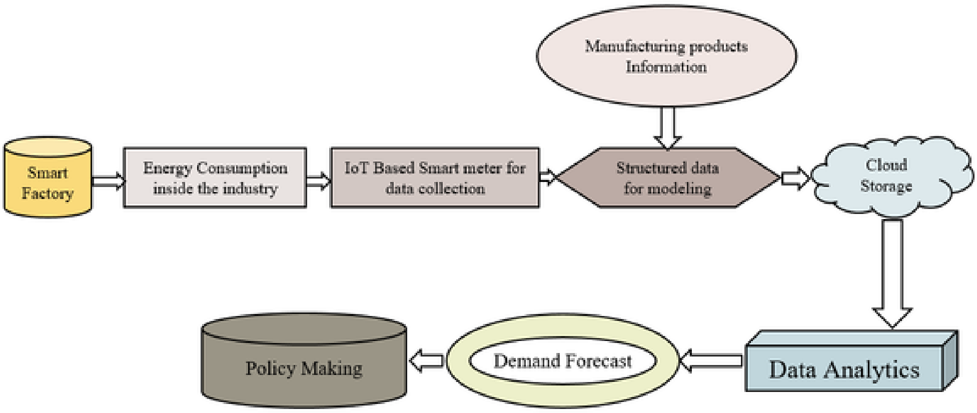

Manufacturing products Information
Smart Factory
Energy Consumption inside the industry
IoT Based Smart meter for data collection
Structured data for modeling
Cloud Storage
Data Analytics
Demand Forecast
Policy Making

Figure 3. Energy consumption measurement in 2017

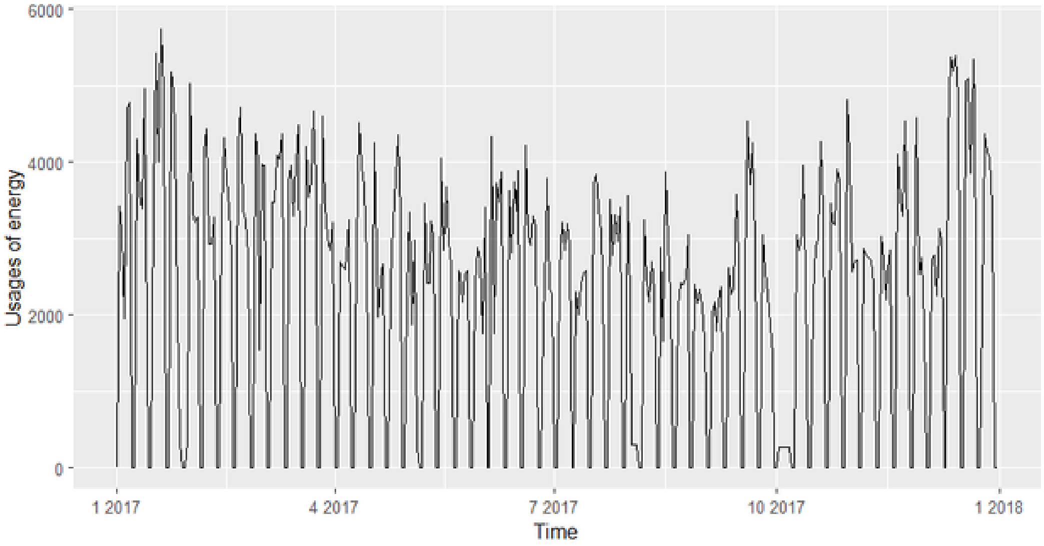


Figure 3 represents the total usages of energy for the Daewoo steel factory, South Korea, in the year2017, and the energy consumption pattern has a lot of variation. Figure 3 shows the energy consumption profile for the period and shows high variability. Since the steel industry is in open space and has no heaters or cooling systems, the temperature factors do not affect energy consumption. Figures 4 and 5 display the details of the energy usage of the steel factory in 2017 through a histogram and box plot. The histogram plot displays the degree of energy use at the time, and the boxplot indicates the median location with the black line. The outliers are identified only with circles well

Figure 4. Histogram of allocation of the energy consumption of appliances. The figure indicates the frequency (bar width) of energy consumption in the interval.

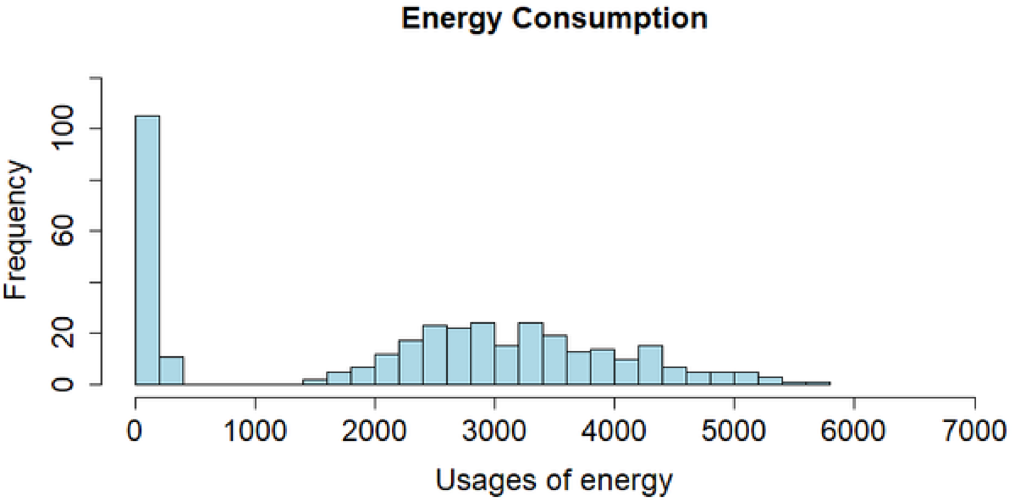
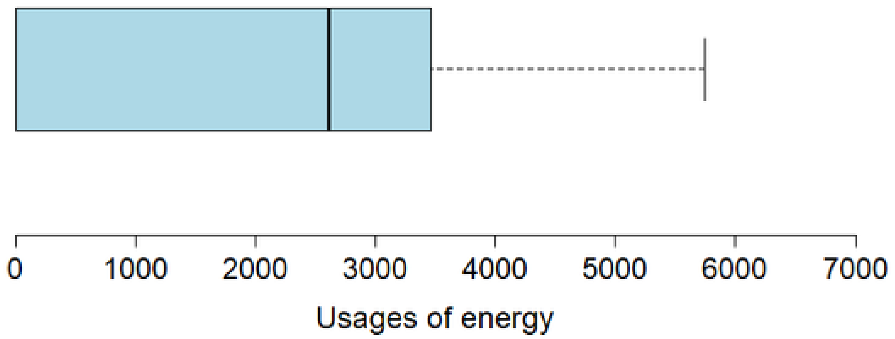


Figure 5. Appliances energy consumption distribution of the box plot. The diagram indicates where the median is located with the black line.



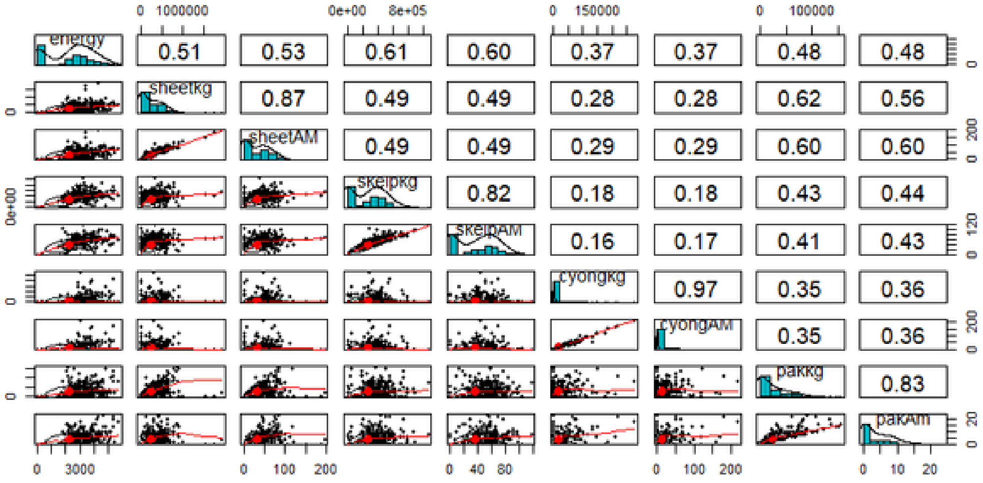
above upper feathers. Since the steel industry is in open space and does not have heaters or cooling systems, energy consumption is not influenced by temperature or other environmental factors.

EXPLORATORY ANALYSIS

In this study, the data sets have a total of 18222 entries with 12 variables. The final dataset deveined into two-part training validation, and another one is testing validation by using CCARET's data partition. In training the models, 75 present data were used, and the rest of the data was used for testing purposes.

Figure 6 shows the relationship among variables with total energy usage in the training data sets using the pairs plot function. The psych package was used to create the figure. This diagram shows how the directional histogram plots the bivariate scatter plots around the directional and the spearman

Figure 6. Pairs plot. Relationship between the consumption of industrial energy with Sheet, Skelp, Cyong, and packaging.

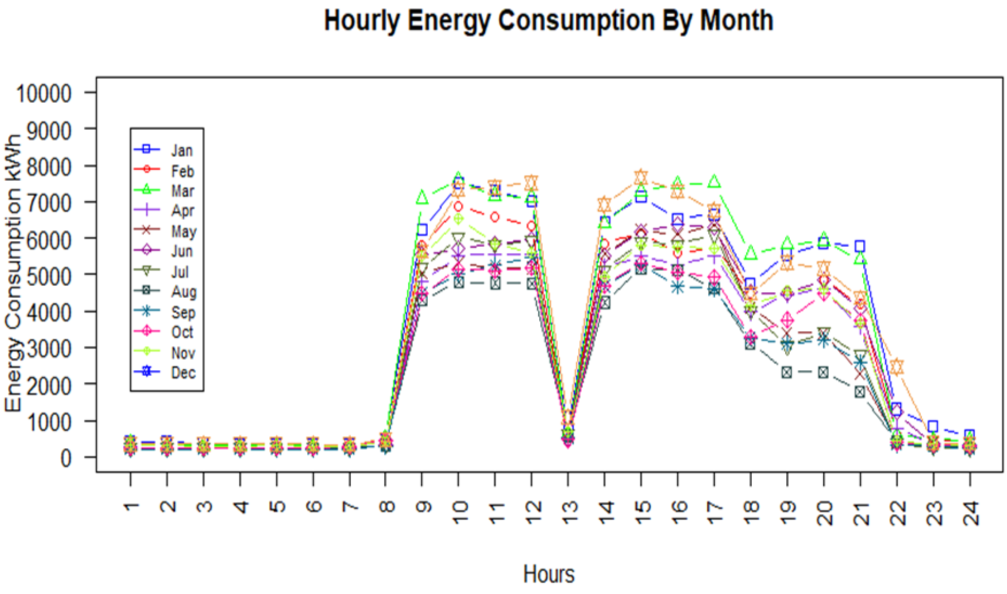


connection above. This is the estimate of two variables' monotonic relations. The correlation value of 1 is the positive overall correlation, and -1 is the overall negative correlation, and 0 does not reflect a correlation between variables. For each pair, the linear regression redline is seen.

Figure 6 shows the positive correlation between energy and skelp kg is (.61). This that the manufacturing skelp of Daewoo steel factory used more energy than other manufacturing products in 2017. The relation between sheet and usages of energy is (.51), and the relation between manufacturing product cyong and usages of energy (.37). These correlation values show the relation among all manufacturing products with total energy usage in the Daewoo steel factory.

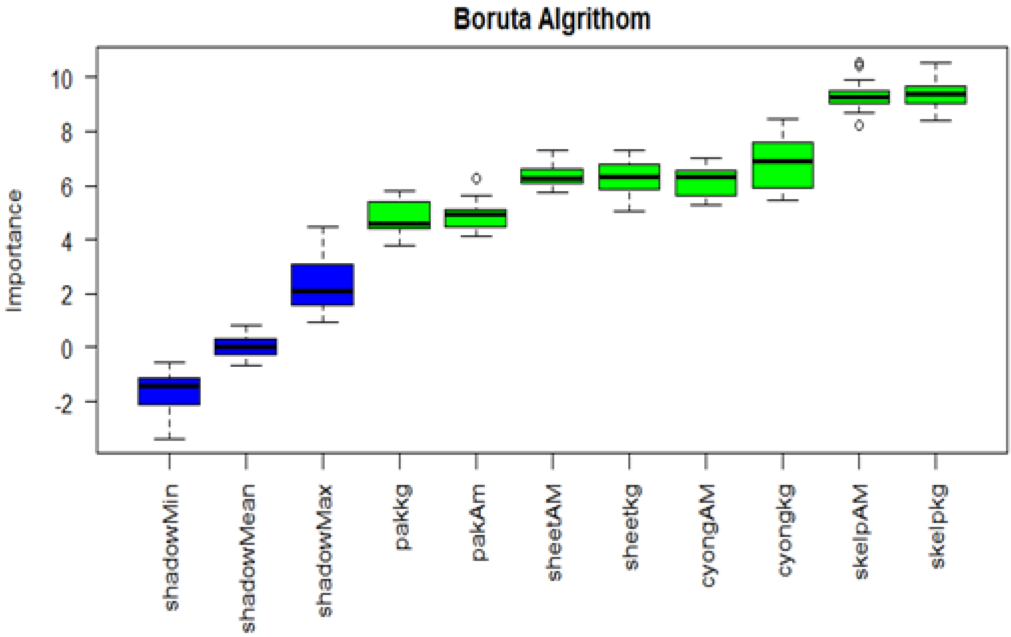
Based on the energy consumption dataset in 2017, we evaluate the hourly uses of energy for every month to find out per-hour uses of energy to calculate the time-period of highest and lowest energy consumption. For every month of the year 2017, Figure 7 displays the hourly cumulative consumption of electricity for the number of working days. The x-axis shows the time in hours in the table, and the y-axis shows the overall energy usages in kWh. From figure 7, we can see that every month from 8 AM to 10 PM, energy consumption is so high, and we can also see that after 11 PM to till 8 AM the next day, the energy consumption is low. Based on the figure, we can easily say that the Daewoo steel factory working hours are from 8 AM to 10 PM, but sometimes they work until 11 AM.

Figure 7. Hourly usages of energy measurement for every month of the year 2017.



It is beneficial now to identify which characteristics are the most significant and those that do not boost the estimation of the electrical energy consumption of the appliances. For this job, all the necessary variables were selected from the Boruta kit. Many researchers have used this kit to filter variables. Figure 8 gives a lot of information on the relevance of characteristics for prediction, but it doesn't provide any information on the efficacy of the components used for the RMSE values. Recursive Feature Elimination (RFE) is used to find the ideal variable count needed to keep the increase to a minimum (Fan et al., 2014). In the regression process, the RMSE is used to assess the effectiveness of a method. The RFE method from the CARET package was used in this study (Kuhn, 2015). The RFE approach in CARET requires creating dummy variables from factor/category variables constructed using the R program. In the data sets, two random variables were used for evaluating the Boruta algorithm. In addition, this function or variable collection assists in model interpretability and decreases model complexity. For example, figure 8 shows the variable importance of the Daewoo steel factory using Boruta's algorithm. From the figure, we can easily find out that Skelp is the most important variable for the Daewoo steel factory of South Korea.

Figure 8. Component importance and preference from Boruta's algorithm.



EVALUATION INDICES

All prediction models are equipped to select the finest tuning parameters with a 10-fold cross-validation scheme. Multiple measurement factors were utilized to compare the regression model's performance. The performance measurement indices used here are the R square value, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Variance (CV).

To calculate the performance of prediction models evaluating criteria are utilized. The Root Mean Squared Error (RMSE) is used to discover the square error relative to real values and compute the prediction's square error compared to actual values and the square root of the summation factor. RMSE is a level-dependent variable that consists of values of the same measuring units.

R-squared is a mathematical measurement of just how near the fitted regression line is to the results. It is also known as the coefficient of determination or the coefficient of multiple determination for multiple regression. R^2 is the determination coefficient that varies from 0 to 1, and the higher value represents the goodness-of-fitness. So, if the R^2 value is close to one means it gives good prediction results.

The equations,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n}}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{y}_{data}\right)^2}$$

The Mean Absolute Error (MAE) evaluates the prediction acuteness. It is a scale-dependent metric, which effectively reflects the prediction error by preventing the offset between positive and negative errors.

We can calculate MAE using the following equation,

$$MAE = \frac{\sum_{i=1}^{n}\left[Y_i - \hat{Y}_i\right]}{n}$$

Here, the real value of the calculation is Y_i, the expected value is $\hat{Y}_{j,}$ And the number of success measurements is z.

The coefficient of variation

The coefficient of variance (CV) is used for measuring the calculation of relative uncertainty. CV is used for evaluating the ratio of the standard deviation to the mean. The CV is extremely useful when comparing findings from two different surveys or studies with different measurements or values (Candanedo et al., 2017). An example is when we compare the findings from two tests that have different scoring mechanisms. If sample A has an 11 percent CV and sample B has a 23 percent CV, then assume that sample B has more variance than its mean.

The equations of the coefficient of variation,

$$CV = \frac{\sqrt{\frac{\sum_{j=1}^{z}\left(Y_j - \hat{Y}_j\right)^2}{z}}}{\overline{y}} * 100$$

Where,

Y_j = Actual measurement value,

$\hat{Y}_j$ = Predicted value

z = Number of performance measures

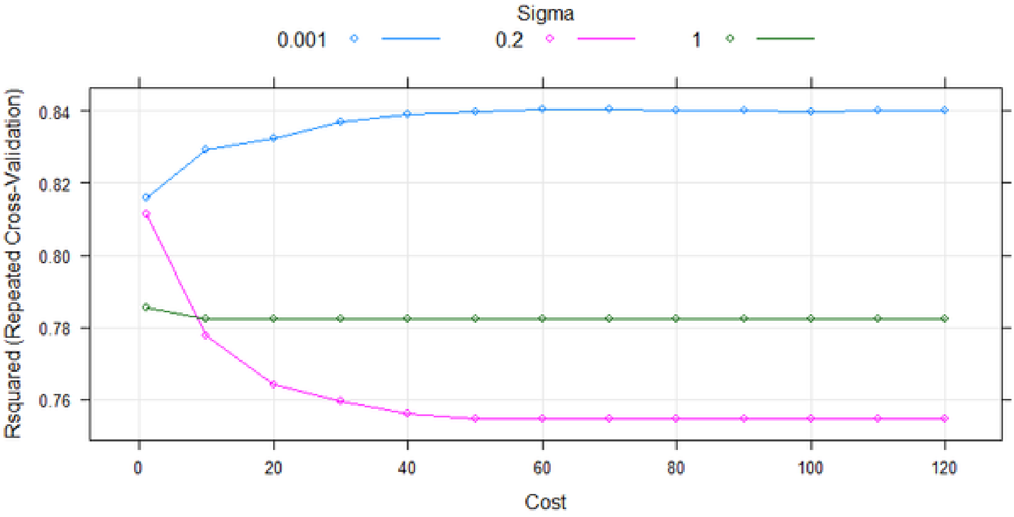
$\overline{y} = \vec{X}$ is mean value

And are multiplied by the coefficient with 100 to get a percentage.

RESULT AND DISCUSSION

To figure out the optimum controller parameters in each regression algorithms defining and minimizing the error values is necessary. The caret kit offers a grid search feature to find the right parameter values for a model (A.B.M. Salman Rahman et al., 2020). In our study, we used three statistical models to find the best prediction model among these three. The three statistical models are the general GLM, SVM Radial, and BT, and we find the performance of all prediction models by measuring R^2, RMSE, MAE, and CV% value.

Figure 9. Results for appeasement values of sigma and cost for the SVM-radial model using the grid search function



In addition to the predictors, the SVM- radial model involves two tuning parameters, namely sigma and cost. The optimal values of sigma for this study (1) and cost (60) variables were obtained with grid search results shown in figure 9. After the cost value of 60, the R^2 value remains constant.

For the Boosting Tree model, the optimum tuning parameters include the highest possible tree depth of 3 and most top trees 42, and the grid results are shown in Figure 10. It can be shown that the R square value stays unchanged after tree value 41. Table 1 displays the model's performance outcomes for both data sets in training and testing. Table 2 can easily find the best prediction among the statistical model based on R^2, RMSE, MAE, and CV% value.

Figure 10. Results for appeasement values of sigma and cost for boosting Tree using the grid search function.

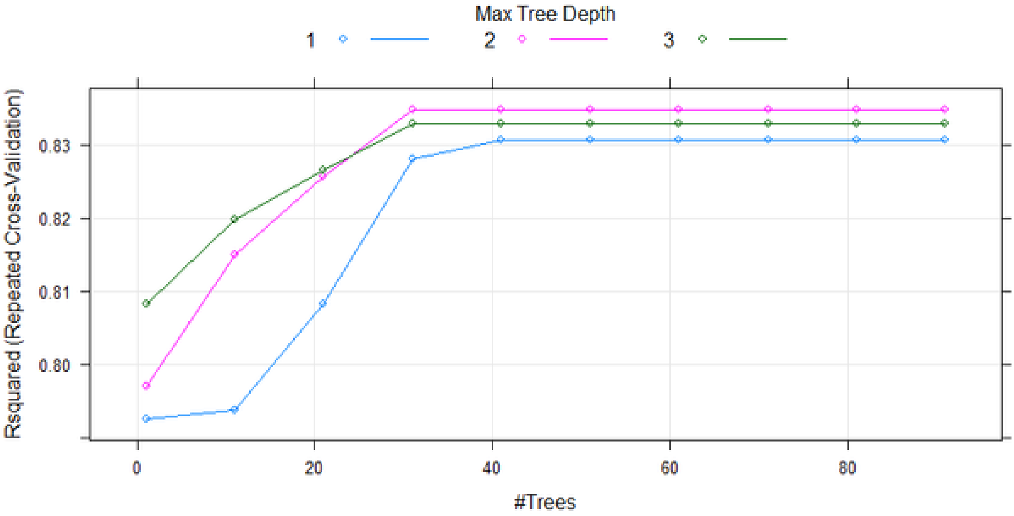
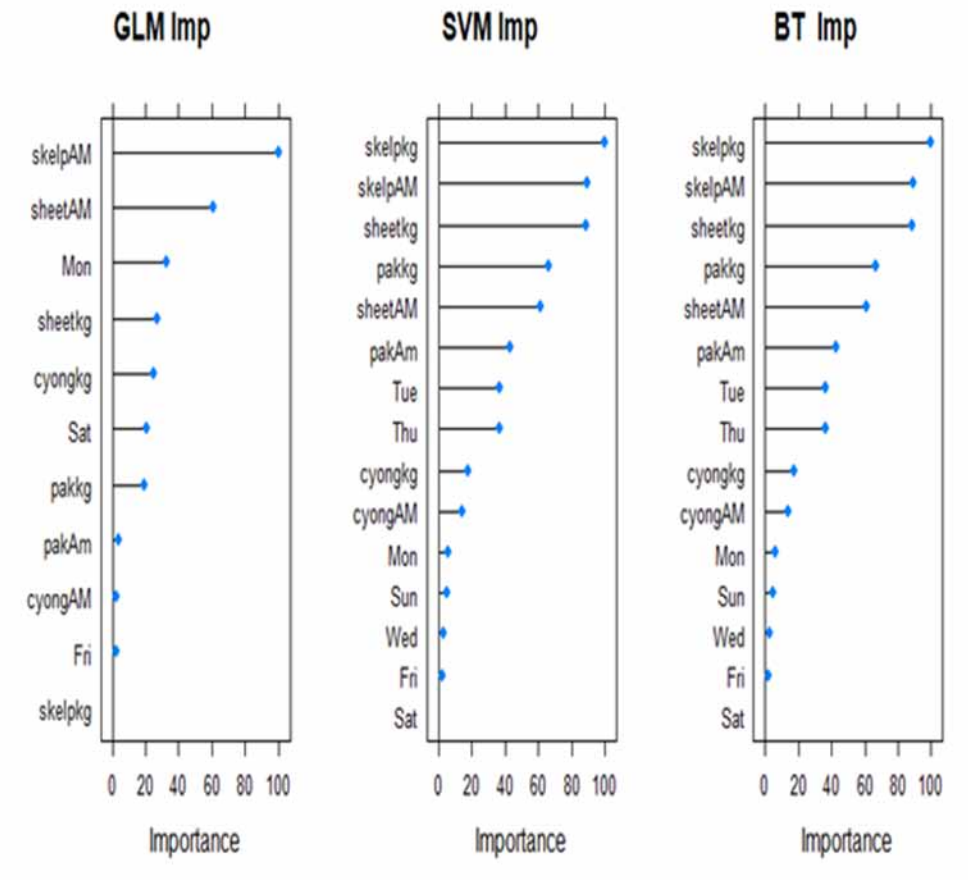


Table 2. Models Performance for taring and testing data sets

Models.	Training.				Testing.			
	R square	RMSE	MAE	CV (%)	R square	RMSE	MAE	CV (%)
GLM	0.79	7.40	5.01	17.25	0.76	7.00	5.32	18.62
SVM RBF	0.86	6.61	4.59	14.36	0.85	6.13	4.30	15.48
BT	0.84	6.17	4.82	15.51	0.84	6.57	4.67	16.32

Every method includes thirty outcomes of ten-fold cross-validation (CV) sets and three repeats after training of each regression model. CARET and the confidence intervals are used to plot R^2 and RMSE values for each model together. The model with close to 1 of R^2 and the lowest RMSE value is considered as the best one among these three models for prediction. As we can see from table 1, SVM Radial has the best R square and RMSE value among these three predictive models. So, SVM Radial is the best predictive model consider with GLM and BT. Figure 11 shows the variable importance for the GLM, SVM Radial, and BT models. Figure 11 can easily justify that Skelp is the most important parameter or product for using energy in the Daedwoo steel factory.

Figure 11. Variable Importance for GLM, SVM, and BT



All tests and analyses give us the acuteness between energy consumption and different manufacturing products in the Daewoo steel factory. As shown in Figure 2, the Daewoo steel industry's energy consumption structure is extremely complex, with almost continuous cycles of demand followed by high spikes. In figure 6, there are strong associations between energy and skelp usages (0.61). Figure 7 shows hourly usages of energy measurement for every month of the year 2017. The Boruta algorithm observed that the dataset has two random variables and then suggested how nearly all the primary factors are relevant in the prediction problem in Figure 8. From table 2, we find SVM Radial is the best fit model based on R square error .86 for training and 0.85 for testing data sets for steel factory energy prediction among three models. Regarding the variable importance functions in figure 11, we find out Skelp is the most important factor for energy consumption

Conclusion

It is critical for a smart factory to have a full grasp of industrial energy usage trends. There are two levels to the optimization process. The case of both the prediction models in exploratory analysis and the data analysis revealed a thought-provoking result. The pairwise plots revealed various types of relationships between variables hidden in the initial statistical models. The GLM, SVM Radial, and BT models boost R square value, RMSE, and MAE predictions to compare with the models. From the three models, SVM Radial based model give the best result for predictions. Skelp was considered the most important product in the Daewoo steel factory for all regression models and the most important factor in predicting energy consumption. Future work could include analysing the energy consumption for every piece of equipment of the Daewoo steel factory to determine the product-wise energy consumption.

ACKNOWLEDGMENT

This work was assisted 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). The Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korean government (MOTIE) (20202020900060) have been sponsored this work. This research has been funded by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2020-0-01489) supervised by the Institute for Information & communications Technology Planning & Evaluation (IITP). Interdisciplinary Program in IT-Bio Convergence System, 21Sunchon National University, 255, Jungang-ro, Suncheon-si, Jeollanam-do 57922, Republic of Korea

REFERENCES

Bataille, C., & Melton, N. (2017). Energy efficiency and economic growth: A retrospective CGE analysis for Canada from 2002 to 2012. *Energy Economics*, *64*, 118–130. doi:10.1016/j.eneco.2017.03.008

Boursianis, Papadopoulou, Diamantoulakis, Liopa-Tsakalidi, Barouchas, Salahas, Karagiannidis, Wan, & Goudos. (2020). Internet of Things (IoT) and Agricultural Unmanned Aerial Vehicles (UAVs) in smart farming: A comprehensive review. *Internet of Things*. 10.1016/j.iot.2020.100187

Candanedo, , Feldheim, & Deramaix. (2017). Data-driven prediction models of energy use of appliances in a low-energy house. *Energy and Building*, *140*, 81–97.

Chen, C., Liu, Y., Kumar, M., Qin, J., & Ren, Y. (2019). Energy consumption modeling using deep learning embedded semi-supervised learning. *Computers & Industrial Engineering*, *135*, 757–765.

Cheng, B. S., & Lai, T. W. (2004). An investigation of co-integration and causality between energy consumption and economic activity in Taiwan. *Energy Econ*, *21*, 435–444.

Chou, J. S., & Bui, D. K. (2014). Modeling heating and cooling loads by artificial intelligence for energy-efficient building design. *Energy and Building*, *82*, 437–446.

Chou, J.-S., & Bui, D.-K. (2014). Modeling heating and cooling loads by artificial intelligence for energy-efficient building design. *Energy and Building*, *82*, 437–446.

Cvitić, Perakovic, Periša, & Stojanović. (2021). Novel Classification of IoT Devices Based on Traffic Flow Features. *Journal of Organizational and End User Computing, 33*(6), 1-20.

Dai, , Wang, Xu, Wan, & Imran. (2019). Big Data Analytics for Manufacturing Internet of Things: Opportunities, Challenges and Enabling Technologies. *Enterprise Information Systems*, *14*(9), 1–25.

Dong, , & Cao, , & Leea. (2005). Applying support vector machines to predict building energy consumption in the tropical region. *Energy and Building*, *37*, 545–553.

Fan, C., Xiao, F., & Wang, S. (2014). Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. *Applied Energy*, *127*, 1–10.

Geng, T., & Du, Y. (2020, February 4). The business model of intelligent manufacturing with Internet of Things and machine learning. *Enterprise Information Systems*, 1–19. Advance online publication. doi:10.1080/17517575.2020.1722253

Hämäläinen & Inkinen. (2019). Industrial applications of big data in disruptive innovations supporting environmental reporting. *Journal of Industrial Information Integration, 16*.

James, Witten, Hastie, & Tibshirani. (2013). *An introduction to statistical learning.* Springer.

Ji, X., Zhou, L., & Wu, Q. (2015, May). A Novel Action Recognition Method Based on Improved Spatio-Temporal Features and AdaBoost-SVM Classifiers. *International Journal of Hybrid Information Technology*, *8*(5), 165–176.

Kim, S. H., Kim, T. H., Kim, Y., & Na, I. G. (2001). Korean energy demand in the new millennium: Outlook and policy implications, 2000–2005. *Energy Policy*, *29*(11), 899–910. doi:10.1016/S0301-4215(01)00018-0

Kuhn, M. (2015). Caret: Classification and regression training. Astrophysics Source Code Library.

Lee, C. C. (2005). Energy consumption and GDP in developing countries: A cointegrated panel analysis. *Energy Econ*, *27*(3), 415–427. doi:10.1016/j.eneco.2005.03.003

Lin, P., Li, M., Kong, X., Chen, J., Huang, G. Q., & Wang, M. (2017). Synchronisation for smart factory - towards IoT-enabled mechanisms. *International Journal of Computer Integrated Manufacturing*, *31*(7), 624–635. doi:10.1080/0951192X.2017.1407445

Liu, R., & Hu, X. (2016, October). Case Study of Construction Cost Estimation in China Electric Power Industry Based on BIM Technology. *International Journal of Grid and Distributed Computing*, *9*(10), 173–186.

Nasreen, S., & Anwar, S. (2014). Causal relationship between trade openness, economic growth and energy consumption: A panel data analysis of Asian countries. *Energy Policy*, *69*, 82–91. doi:10.1016/j.enpol.2014.02.009

Ockwell, D. G. (2008). Energy and economic growth: Grounding our understanding in physical reality. *Energy Policy*, *36*(12), 4600–4604. doi:10.1016/j.enpol.2008.09.005

Paturi, U. M. R., & Cheruku, S. (2020). Application and performance of machine learning techniques in the manufacturing sector from the past two decades: A review. *Materials Today: Proceedings*.

Rahman, , Ragu, Lee, Park, Cho, Lee, & Shin. (2018, October). An Analysis Study Based on Linear Regression Model for Changes of Fruit Size over Plum Diseases. *Journal of Knowledge Information Technology and Systems*, *12*(5), 509–519.

Rahman, , Lee, Lim, Cho, & Shin. (2020). A prediction model for steel factory manufacturing product based on energy consumption using data mining technique. *Journal of Science and Engineering Management*, *1*(2), 9–16.

Salman Rahman, A. B. M. (2019, May 21). Identification of High Significance Product Items Through the Analysis of Energy Consumption in Steel Factory. *Journal of Knowledge Information Technology and Systems*, *14*(3), 275–289. doi:10.34163/jkits.2019.14.3.007

Sathishkumar, V. E. (2021, January 2). Changsun Shin., & Yongyun Cho., (2020), Efficient energy consumption prediction model for a data analytic-enabled industry building in a smart city. *Building Research and Information*, *49*(1), 127–143. Advance online publication. doi:10.1080/09613218.2020.1809983

Schölkopf, B., Burges, C., & Vapnik, V. (1995), Extracting support data for a given task. *Proceedings of first international conference on knowledge discovery and data mining*.

Simeone, O. (2018). A very brief introduction to machine learning with applications to communication systems. *IEEE Transactions on Cognitive Communications and Networking*, *4*(4), 648–664.

Stern, D. I. A. (2000). Multivariate cointegration analysis of the role of energy in the U.S. macroeconomy. *Energy Econ*, *22*(2), 267–283. doi:10.1016/S0140-9883(99)00028-6

Tao, , Wang, Zuo, Yang, & Zhang. (2016). Internet of Things in product life-cycle energy management. *Journal of Industrial Information Integration.*, *1*, 26–39.

Donglan, Zhou, & Ding. (2009). The contribution degree of sub-sectors to structure effect and intensity effects on industry energy intensity in China from 1993 to 2003. *Renewable & Sustainable Energy Reviews*, *13*, 895–902.

Tsani, S. Z. (2010). Energy consumption and economic growth: A causality analysis for Greece. *Energy Econ*, *32*(3), 582–590. doi:10.1016/j.eneco.2009.09.007

Xiao, L., Shao, W., Liang, T., & Wang, C. (2016). A combined model based on multiple seasonal patterns and modified firefly algorithm for electrical load forecasting. *Applied Energy*, *167*, 135–153. doi:10.1016/j.apenergy.2016.01.050