Journal of Science and Engineering Management
Vol. 1, No. 2 (2020), pp.9-16
http//dx.doi.org/10.33832/jsem.2020.1.2.02

A PREDICTION MODEL FOR STEEL FACTORY MANUFACTURING PRODUCT BASED ON ENERGY CONSUMPTION USING DATA MINING TECHNIQUE

A. B. M. Salman Rahman[1], Myeongbae Lee[2], Jonghyun Lim[3], Yongyun Cho[4] and Changsun Shin[5*]

[1]Department of Information & Communication Engineering, Sunchon National University, South Korea
[1]salman01717@gmail.com, [2]lmb@scnu.ac.kr, [3]sshb56@s.scnu.ac.kr, [4]yycho@sunchon.ac.kr, [5]csshin@sunchon.ac.kr

***Abstract*— Energy has been obtained as one of the key inputs for a country's economic growth and also for social development as well. Analysis and modeling of industrial energy are currently a time-insertion process because more and more energy is consumed for economic growth in an industrial factory. Industrial energy consumption analysis and predictions play a very important role in improving energy utilization rates to make profitable things for industrial companies or factories. This study is aimed to present and analyze the predictive models of the data-driven system for the uses of appliances. This paper is intended to address the filtering of data to use non-predictive parameters and ratings of features. 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 with the radial kernel (SVM RBF) 3) boosting tree (BT). The performance of prediction models was measured by R^2 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 R^2 of 0.86 for the training data set and 0.85 for the testing data set with a low coefficient of variation**

***Keywords*— Energy Consumptions, Correlation, General Linear Regression, Support Vector Regression, Boosting Tree**

1. 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, and nowadays more and more energy is being used for both economic growth and population growth. Facilities of industrial customers and the use of electricity to process various types of machinery, manufacture or assemble products, including such diverse industries as production, mining, and construction [1]. Ultimately, more than one-third of electrical energy is used by those industrial sectors from total energy for a country.

Since the 1990s, South Korea's manufacturing industry has continued to expand at a high pace and has become the main driving force of South Korean economies, with rapid growth. Primary energy consumption rose at an annualized rate of 7.5% in the 1990s, which in the same period was higher than the annualized economic

Received: October 26, 2020
Reviewed: December 14, 2020
Accepted: December 17, 2020
* Corresponding Author

ISSN: 2652-8045 JSEM

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 to increase the loss of energy conversion, further reducing the energy intensity [2]. Many studies have shown that improving energy efficiency is very important for economic growth [3],[4]. Industrial factory owners are also beginning to realize that analyzing and forecasting the energy data with the production data is so important for the benefit of their companies or plants.

This paper focused on energy consumption based on the productions of the Daewoo steel factory in South Korea The rest of this paper is arranged in the following way. Section two discusses the related works. Section three discusses the analysis methodology. Section four presents the reported data and description, exploratory analysis, filtering, and significance of data features. Section five concentrated on findings and discussion, and the paper is concluded in section six.

2. RELATED WORKS

Reducing energy use in the steel sector is a global problem where the government is aggressively taking steps. A steel plant can handle its resources better if it is possible to model and estimate consumption [5]. The calculation of the heating load is the first step of the construction process for iterative heating, ventilation, and air conditioning (HVAC). [6]. The highest contributors to the overall structure and strength effect came from the electrical equipment and machinery and raw chemical materials and chemical components sub-sectors, and the smallest contributions were from the gas and petroleum refining and coking manufacturing and supply industries. [7]. Technological innovation has provided large opportunities for researchers in diverse fields to use artificial intelligence. In the manufacturing and development fields, various attempts have been made to use machine learning methods [8].

3. MATERIALS AND METHODS

3.1. DATA DESCRIPTION

There are two types of data set available in this study that have been gained from Daewoo Steel Factory, South Korea. Between these two data sets, one data set for the energy consumption of the steel factory and the other data set for the productions of the steel factory. This factory manufactured different types of steel, rods, and plates. The period of collecting data is 365 days (12 months) for the year 2017. Fig. 1 shows the total usages of energy for the year 2017.

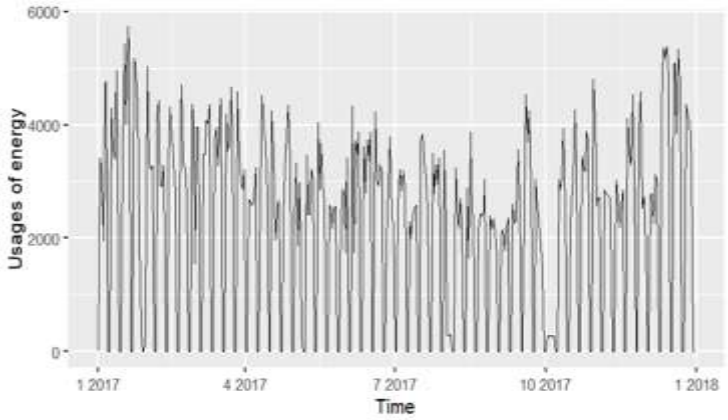

Fig. 1 Energy Consumption Measurement for the Year 2017

3.2. EXPLORATORY ANALYSIS

In this study, the data set has a total of 18222 entries with 12 variables. The final dataset is deveined into two-part one training validation part and another one is testing validation part by using CCARET's data partition. For training the models 75 present data is used and the rest data is used for testing purposes. Fig. 2 shows the relationship among all variables with the usages of total energy in the training data sets by using the pairs plot function. This diagram shows below the directional histogram plots the bivariate scatter plots around the directional and the spearman connection above. This is the estimate of the two variables' monotonic relations. The correlation of 1 is a positive overall correlation and -1 is negative overall correlations and 0 does not reflect a correlation between variables. Fig. 2 shows the positive correlation between energy and skelpkg (.61). This means for manufacturing skelp this Daewoo steel factory used more energy than other manufacturing products in the year 2017. The relation between Sheet and usages of energy is (.51) and also the relation between manufacturing product cyong and usages of energy (.37).

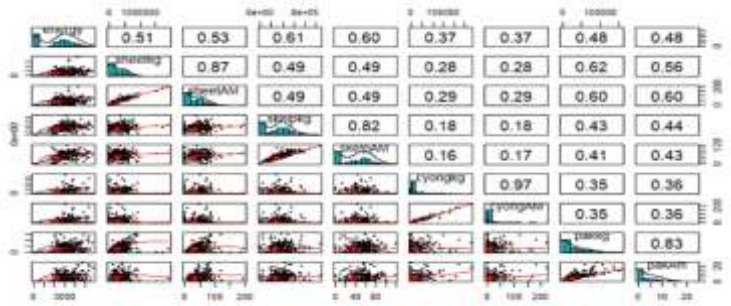

Fig. 2 Pairs Plot. Relationship between the Energy Consumption of Industrial Resources with the Production

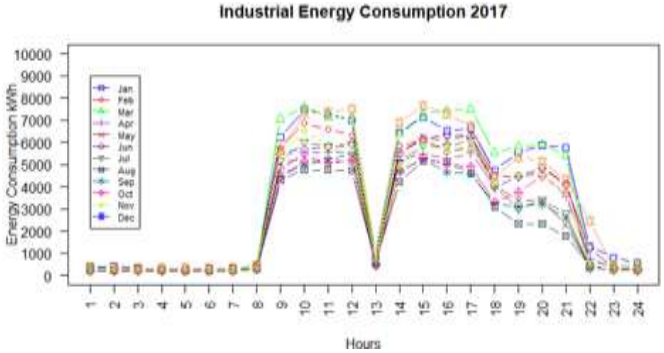

Fig. 3 Hourly Usages of Energy Measurement for Every Month of the Year 2017

Fig. 3 shows every hour total usages of energy for working days consumption of every month in the year 2017. The x-axis shows the time in hours in the table, and the y-axis shows in kWh for overall energy usages. From the figure, we can see that every month from morning 8 am to 10 pm energy consumption is so high and we can also see that after 11 pm to till 8 am energy consumption is low. Fig. 4 shows the variable importance of the Daewoo steel factory by 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.

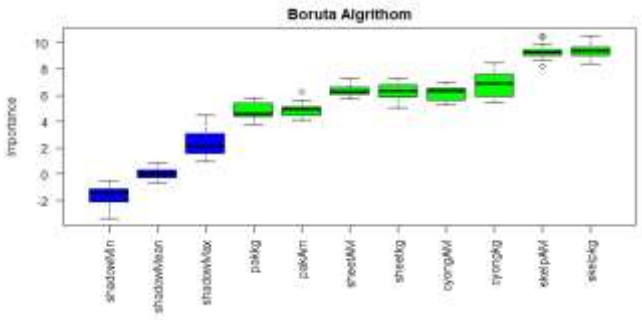

Fig. 4 Component Importance and Preference from Boruta's Algorithm

3.3.1. LINEAR REGRESSION: Linear regression is associated with the nursing method for analyzing the relationship between a scalar dependent quantity variable y and one or more independent quantity variable X denoted. In statistics, the general linear regression model (GLM), is a flexible generalization of the simple linear regression model (GLM) [9] and assumes that the data points are distributed randomly.

We assume a model,

$$G(M(y)) = x_j \times \beta_0 \quad + 0 > y \sim K \tag{1}$$

Here G(M) is the identified relation function, β_0 is the regression coefficient, x_i is the predictor, y is the predicted output, O is the offset variable [9], and K is the distribution model of y.

3.3.2. SUPPORT VECTOR REGRESSION: In data science, the area unit supports vector machines (SVMs) and supervises learning models with related learning algorithms analyzing information. Support Vector Machine can also be used as a regression tool, holding all the most options characterizing the formula (maximum margin) [10]. Support Vector Regression (SVR) uses a similar concept for classification due to SVM with only a few minor variations. First, since the output is a real number, predicting the information at hand becomes very complicated, which has infinite possibilities. For the case of regression, an approximation to the SVM, there is a margin of tolerance (epsilon) which may have already been requested from the matter, but there is also an additional sophisticated explanation, so the formula is more complicated to consider[11].
Equation of SVR;

$$f(x) = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) k(x_i, x) + b \tag{2}$$

In this case, the value of $K(x_i, x)$ is similar to the outer combination of two vectors x_i and x_j in the function space $\phi(x_i)$ and $\phi(x_j)$, that is, $K(x_i,x_j) = \phi(x_i)\cdot\phi(x_j)$. All required simulations can indeed be performed directly in the feature space, and no need to calculate the map $\phi(x)$, by the use of kernels.

3.3.3. BOOSTING TREE: Boosting Tree is a regression and classification problem machine learning technique and generates a prediction model in the form of an ensemble of strong prediction models. Boosting Trees (BT) consists of an ensemble of decision trees and is an additive regression model. A single decision tree has the problem of over-fitting, but by integrating hundreds of weak decision trees consisting of a few leaf nodes, the BT algorithm can solve this [12].

Assume model is,

$$F_m(x) = F_{m-1}(x) + v\sum_{i=1}^{J_m}\gamma_m I(x \in R_{jm}) \quad (3)$$

3.4. EVALUATION INDICES

All prediction models are equipped to select the finest tuning parameters with a 10-fold cross-validation scheme. To compare the regression model performance, multiple measurement parameters are used. The performance measurement indices used here are R square value, root mean square error (RMSE), mean absolute error (MAE), and coefficient of variance (CV).

RMSE is a scale-dependent metric and it results in values with the same units of the measurements and R^2 is the coefficient of determination, which ranges from 0 to 1, reflecting the goodness-of-fit.

Equations are,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}} \qquad R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \ddot{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2}$$

Mean absolute error (MAE) to evaluate the prediction acuteness and MAE is a scale-dependent metric. We can calculate MAE by using the following equation,

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

Here, $Y_i$is the actual measurement value, $\hat{Y}_i$ is the predicted value and n is the number of performance measures.

To calculate the measure of relative variability the coefficient of variation (CV) is utilized. CV is used to find out the ratio of the standard deviation to the mean.

Equation of coefficient of variation:

$$CV = \frac{\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}}}{\bar{y}} * 100$$

Here, Y_i= Actual measurement value, $\hat{Y}_i$ = Predicted value n = Number of performance measures $\bar{y}$= $\vec{X}$ is mean value Multiplying the coefficient by 100 to get a percentage.

5. RESULT AND DISCUSSION

To figure out the optimum controller parameters in each of the regression algorithms it is necessary to define and minimize the error values. The caret kit offers a grid search feature to find the right parameter values for a model. In our study, we use three statistical models to find the best prediction model among these three. Three statistical models are the general GLM, SVM Radial, and BT and we find out the performance of all predictions model by measuring R^2, RMSE, MAE, and CV% value. Fig. 5 shows the grid search results for optimal values for the SVM Radial kernel. For SVM- radial kernel model we need two tuning parameters namely sigma and cost. Fig. 6 shows a grid search result for BT. Table 1 displays the model's performance outcomes for both data sets in training and testing. From table 1 we can easily find out the best prediction among the statistical model.

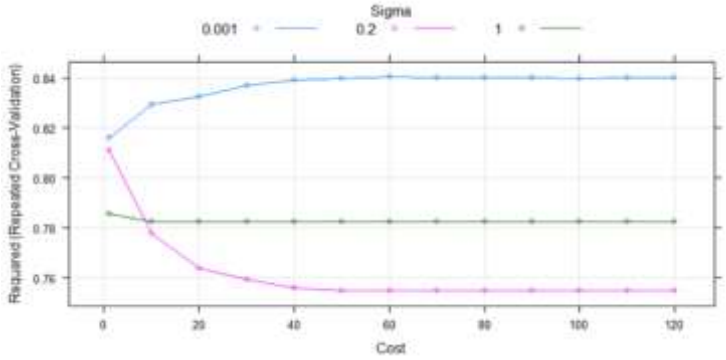

Fig. 5 Results for Appeasement Values of Sigma and Cost for the SVM-radial Model using the Grid Search Function

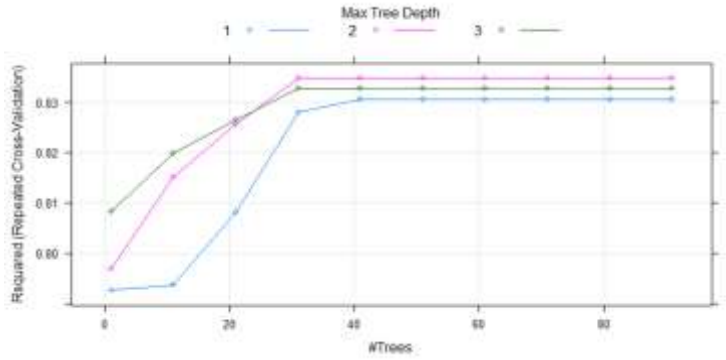

Fig. 6 Results for Appeasement Values of Sigma and Cost for Boosting Tree using the Grid Search Function

Table 1. 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. This instruction is being used by CARET along with the confidence intervals to plot R^2 and RMSE values for each model together. The model with close to 1 of R square and 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. Fig. 7 shows the variable importance for the GLM, SVM Radial, and BT models, and from Fig. 7 we can easily justify that skelp is the most important parameter or product for using energy in the Daewoo steel factory, South Kore.

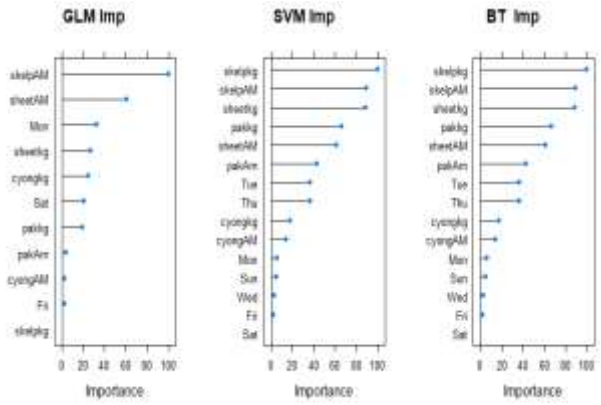

Fig. 7 Variable Importance for GLM Square, SVM, and BT

All tests and analyses give us the acuteness between energy consumptions and different types of manufacturing products Daewoo steel factory. As shown in Fig. 1 The Daewoo steel industry's energy consumption structure is extremely complex, with almost continuous cycles of demand followed by high spikes. In fig. 2. there are strong associations between energy and skelp usages (0.61). Fig. 3 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 also indicated that almost all the key parameters are important in the problem of prediction in fig. 4. From table 1 we find SVM Radial is the best fit model based on R square error .86 for taring and 0.85 for testing data sets for steel factory energy prediction among three models. Regarding the variable importance functions in fig. 7 we find out skelp is the most important factor for energy consumption.

6. CONCLUSION

In the case of both the prediction models in exploratory analysis and the data analysis revealed a thought-provoking result. The pairwise plots displayed different types of parameter relationships that could be concealed in the initial statistical models. The GLM, SVM Radial, and BT models boost the predictions of R square value, RMSE, and MAE to compared with the models and considered from all three models SVM Radial based model give the best result for predictions. For all regression models, Skelp was considered as the most important product in the Daewoo steel factory, and also for predicting energy consumption skelp is the most important factor. Future work could include analyzing the energy consumption for every equipment of the Daewoo steel factory and try to find out product wise energy consumption.

ACKNOWLEDGMENTS

This work was 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). This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korea government (MOTIE) (20202020900060), The Development and Application of Operational Technology in Smart Farm Utilizing Waste Heat from Particulates Reduced Smokestack). This work was 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 (20194210100230).

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, 167, 2016, pp.135-153.
[2] Lee, Seung-moon, “Mid-term Korea Energy Demand Outlook”, Korea Energy Economics Institute, May 2014.
[3] David G.Ockwell, “Energy and economic growth: Grounding our understanding in physical reality”, Energy Policy, 36, 2008, pp.4600-4604.
[4] Chirs Bataille, Noel Melton, “Energy efficiency and economic growth: A retrospective CGE analysis for Canada from 2002 to 2012”, Energy Economics,64, 2017, pp. 118-130.
[5] Chen, Chong, Ying Liu, Maneesh Kumar, Jian Qin, and Yunxia Ren. “Energy consumption modelling using deep learning embedded semi-supervised learning” Computers & Industrial Engineering, volume 135,2019,pp. 757-765.
[6] Chou, J.S. and Bui, D.K., 2014, “Modeling heating and cooling loads by artificial intelligence for energy-efficient building design”, Energy and Buildings, volume 82, 2014,pp.437-446.
[7] Zha, D, Zhou, D, Ding, N. “The contribution degree of sub-sectors to structure effect and intensity effects on industry energy intensity in China from 1993 to 2003”. Renew. Sustain. Energy Rev. 13, 2009, 895–902.
[8] Paturi, Uma Maheshwera Reddy, and Suryapavan Cheruku. "Application and performance of machine learning techniques in manufacturing sector from the past two decades: A review." Materials Today: Proceedings (2020).
[9] Jui-Sheng Chou, and Dac-Khuong Bui, “Modeling heating and cooling loads by artificial intelligence for energy-efficient building design”. Energy and Buildings, 82, 2014, pp.437-446.
[10] Bernhard Schölkopf, Chris Burges, and Vladimir Vapnik, “Extracting support data for a given task” Proceedings of first international conference on knowledge discovery and data mining. 1995.
[11] Bing Dong, Cheng Cao ,Siew Eang Leea, “Applying support vector machines to predict building energy consumption in tropical region” Energy and Building, volume 37,2005 pp.545-553.
[12] Jerome H. Friedman, “Stochastic gradient boosting”, Computational Statistics & Data Analysis, volume 38, 2002, pp. 367-378.

