Predicting Steel Industry Power Consumption Using Machine Learning

I. Introduction

The steel manufacturing industry is crucial to global infrastructure development, providing key materials for sectors such as construction, automotive, and machinery. It ranks among the largest industrial sectors worldwide, characterized by significant energy requirements. Managing power consumption is a major operational cost for steel companies, which face the dual challenges of fluctuating energy prices and stringent environmental regulations while striving to remain profitable and competitive[1].

Steel production involves a multi-stage process, including raw material extraction, ironmaking, steelmaking, casting, rolling, and finishing[2]. Each stage demands substantial energy, relying on electricity, natural gas, and other fuels to operate machinery and maintain high temperatures. Given the scale of steel manufacturing, even minor inefficiencies in energy use can lead to considerable cost and environmental impacts.

In this dynamic environment, where energy demands vary based on numerous factors, accurately predicting these fluctuations is crucial. Effective prediction can help companies make informed decisions regarding energy procurement and usage, thereby enhancing operational efficiency and reducing unnecessary costs. This report leverages machine learning techniques to:

i. Develop predictive model for power consumption using historical data.

ii. Analyze the factors affecting power usage to identify key drivers of energy demand.

By analyzing power consumption data systematically, this study seeks to offer actionable insights that improve energy management strategies, promote sustainable practices, and support the overall profitability of the steel industry.

II. DATASET OVERVIEW

The data was sourced from the UCI Machine Learning Repository[3], originally provided by DAEWOO Steel Co. Ltd, located in Gwangyang, South Korea. The characteristics of the dataset are:

- Dataset Type: Multivariate.
- Associated Task: Regression.
- Feature Types: Real, Categorical.
- Number of Instances: 35,040.
- Number of Features: 9.

A detailed description of the features is shown in TABLE II.

TABLE II: Dataset Features

Feature	Type	Unit
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	%
No of Seconds from Midnight (NSM)	Continuous	s
Week Status	Categorical	Weekend, Weekday
Day of Week	Categorical	Sunday, Monday…
Load Type	Categorical	Light, Medium, Maximum

Exploratory data analysis (EDA) was performed to extract insights from the dataset and understand the relationships between features and the target variable. The analysis was guided by specific questions.

A. Univariate Analyisis

Statistical summaries were performed to analyze the distribution and spread of each continuous feature. Violin plots were used to visualize these distributions, revealing patterns that served as a basis for further analysis. Figure 1 illustrates the distribution for usage_kWh.

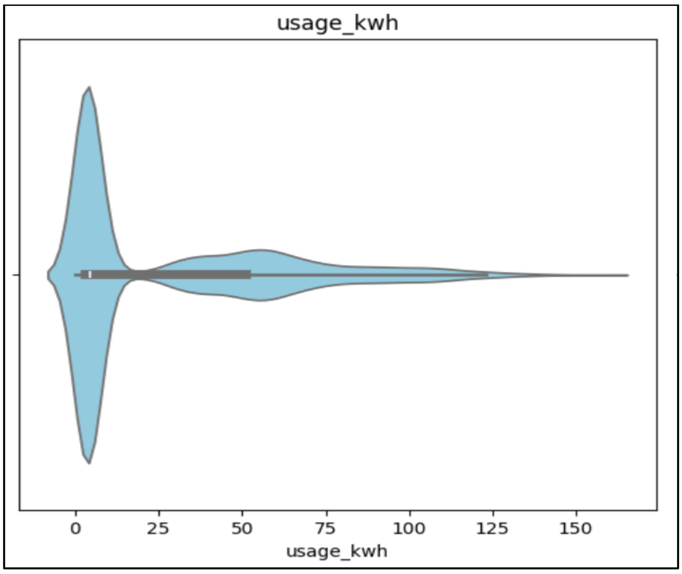

Fig.1: Distribution plot for power usage using violin chart. Observations on the distribution plot shows that most values are near 0 kWh, reflecting periods when machinery is off. Most data falls within the 0 to 50 kWh range, with only a few higher values.

B. Multivariant Analysis

This analysis, which examines multiple variables at once, was carried out to gain insights from the dataset and understand the relationships between features and the target variable. This was driven by specific questions.

- **How do current power factors affect power usage?**

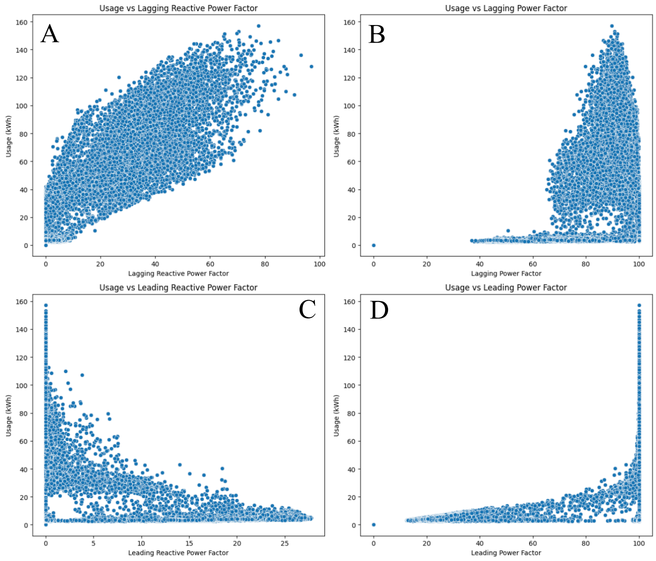

Fig. 2: Relationship between power usage and current power factors: (a) Lagging Reactive Power Factor, (b) Lagging Power Factor, (c) Leading Reactive Power Factor, and (d) Leading Power Factor.

A clear linear relationship is observed between Lagging Reactive Power and power usage, as shown in Fig. 2.

- **What is the power usage pattern across different days of the week and week status?**

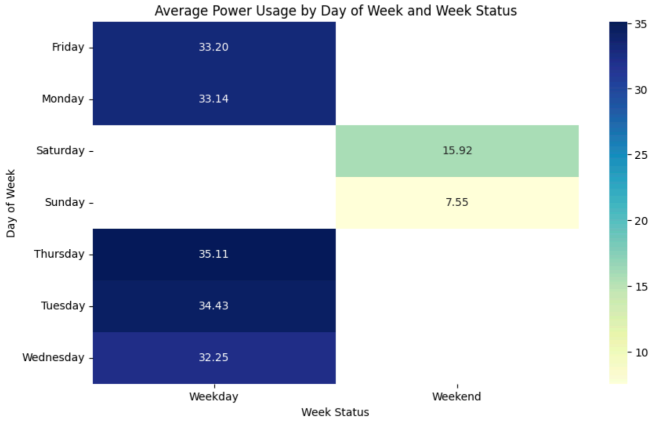

Fig. 3: Average Power Usage by Day of Week and Week Status

Insights from the heatmap analysis reveal that power usage is generally higher during weekdays compared to weekends[4]. This pattern reflects typical usage behavior, with increased consumption on business days and reduced usage on weekends.

- **What are the trends in power usage over time?**

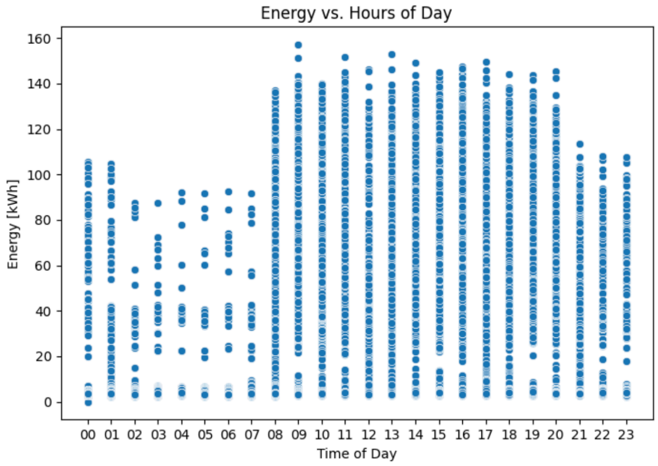

Fig. 4: Power Usage vs Hours of Day

Insights reveal that data predominantly clusters around typical business hours, from 8 a.m. to 8 p.m. Additionally, monthly variations in energy usage are minimal, indicating that the month is unlikely to be a significant predictor.

III. PROBLEM DEFINITION

The problem is a supervised learning regression task. Using a set of input features, the goal is to accurately predict power usage (usage_kWh). The objective is to develop a model that can reliably estimate energy consumption based on these predictors[5].

A cleaned, scaled feature set was created, and several models were built and evaluated. Each model was fitted and assessed based on performance metrics. The analysis of these metrics helped identify the best model for predicting real-time energy consumption[6], ensuring the most accurate and reliable predictions[7].

Algorithm Selection

Two main factors were considered.

- Nature of the Problem: Since the goal was to predict a continuous variable (usage_kwh), the models considered were regression models, capable of predicting numeric values.

- Model Interpretability: explainable models were required to provides interpretable insights into feature importance, which helped to understand the key drivers of energy usage.

Below traditional algorithm and ensemble techniques were selected for evaluation:

- *Linear Regression* models the relationship between a continuous dependent variable and one or more independent variables by fitting a linear equation to observed data. It estimates the dependent variable as a weighted sum of the independent variables, aiming to minimize the difference between predicted and actual values.
- *Ridge Regression* is a type of linear regression that includes an L2 penalty term to address multicollinearity and prevent overfitting. The penalty term, which is the squared magnitude of the coefficients, shrinks the coefficient estimates towards zero, improving model stability and generalization.
- *Lasso Regression* applies L1 regularization to the linear regression model, penalizing the absolute values of the coefficients. This leads to sparse solutions where some coefficients are exactly zero, thereby performing automatic feature selection and enhancing model interpretability.
- *Random Forest* is an ensemble learning technique that constructs multiple decision trees during training and outputs the mean prediction of the individual trees. By averaging the results from a diverse set of trees, it improves predictive accuracy and robustness against overfitting. It handles large datasets and complex relationships effectively through its bagging and feature randomness.
- *Gradient Boosting* builds models sequentially, where each new model corrects errors made by the previous models. It focuses on residual errors and combines weak learners to create a strong predictive model. This approach enhances performance through iterative refinement, leading to high accuracy and capturing complex patterns in the data.
- *XGBoost* (Extreme Gradient Boosting) is an optimized version of gradient boosting that enhances performance through regularization, handling missing data efficiently, and parallel processing. It is particularly effective for complex regression tasks, offering high predictive accuracy and scalability.

Evaluation methodology

The performance of the regression models was assessed using several key metrics. The Mean Absolute Error (MAE) was calculated as:

$$\mathbf{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Where y_i represents the actual values and $\hat{y}_i$ represents the predicted values. This metric provides a straightforward indication of the average magnitude of errors in the predictions.

The Mean Squared Error (MSE) was used to highlight the variance in errors by squaring them:

$$\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

This gives more weight to larger discrepancies. The Root Mean Squared Error (RMSE), derived from the MSE, provides error values in the same units as the target variable:

$$\mathbf{RMSE} = \sqrt{MSE}$$

This metric facilitates direct interpretation of the prediction errors.

R-squared (R^2) was computed to indicate the proportion of variance in the target variable that was explained by the model:

$$\mathbf{R^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where $\bar{y}$ is the mean of the actual values. This offers insights into the model's explanatory power.

The evaluation included data splitting for unbiased assessment, residual analysis for prediction biases, and feature importance analysis to understand key factors in energy consumption predictions

IV. ANALYSIS & EVALUATION

Below steps were taken for the model task (Table. 1)

Step	Description
Data Preprocessing	Data Cleaning: Addressed missing values, and inconsistencies to ensure data integrity.
	Feature Scaling: Standardized continuous features using z-score normalization for equal feature contribution during model training.
	Train-Test Split: Divided the dataset into an 70-30 split to evaluate model performance on unseen data.
Model Selection	**Define model set and tune hyperparameter for evaluation:** Linear Regression, Ridge, Lasso, Random Forest, Gradient Boosting, XGBoost.
Model Training	Trained models on the training dataset; hyperparameters were tuned
Model Evaluation	Evaluated models on test data using MAE, MSE, RMSE, and R^2 metrics.
	Generated residual plots to inspect errors.
	Conducted feature importance analysis to interpret feature contributions.
Model Comparison and Selection	Compared models based on evaluation metrics to identify the best-balanced model.

Model Selection & Evaluation

Regressor	Train R^2	Test R^2	Parameters
Linear	0.981	0.977	
Ridge	0.981	0.977	alpha=1
Lasso	0.981	0.977	alpha=0.1
Random Forest	1.000	0.999	n_estimators=200, random_state=42
Gradient Boosting	0.993	0.992	n_estimators=100, max_depth=3, random_state=42
XGBoost	0.997	0.996	n_estimators=50, max_depth=4, random_state=42

Table 2. Regression models for predicting power usage

Based on the evaluation metrics, we conducted a comprehensive analysis of various models to identify the most suitable one for predicting usage_kwh. The models evaluated include Linear Regression, Ridge Regression, Lasso Regression, Random Forest, Gradient Boosting, and XGBoost.

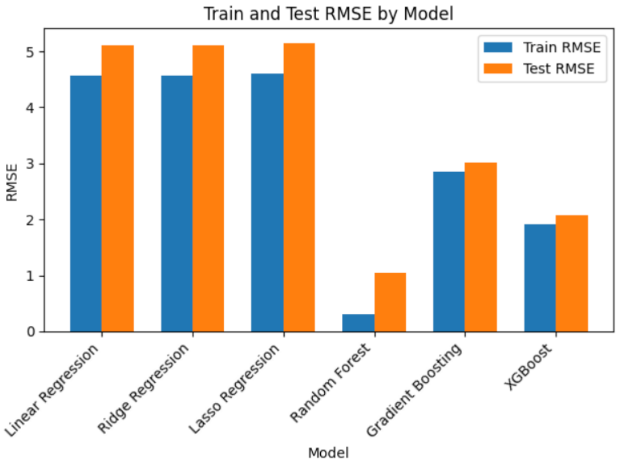

Fig. 5. Bar plot for Train and Test RMSE by Models

Linear Regression and Ridge Regression demonstrated similar performance, with high R² values and low RMSE scores, though the slight increase in RMSE from train to test sets indicates a minor overfitting tendency. Lasso Regression, while similar, showed marginally lower performance, likely due to its feature selection mechanism which may have excluded some predictive features.

Random Forest stood out with near-perfect R² scores and exceptionally low RMSE, but the almost zero train RMSE suggests significant overfitting, which raises concerns about its generalization capability. XGBoost, on the other hand, balanced model complexity with performance, showing strong R² and RMSE metrics with better generalization, making it a robust choice. Gradient Boosting, while effective, was slightly outperformed by XGBoost in terms of both RMSE and R².

In conclusion, XGBoost is recommended as the best model for predicting usage_kwh due to its strong predictive accuracy and lower risk of overfitting, making it reliable for real-world applications where data variability is expected.

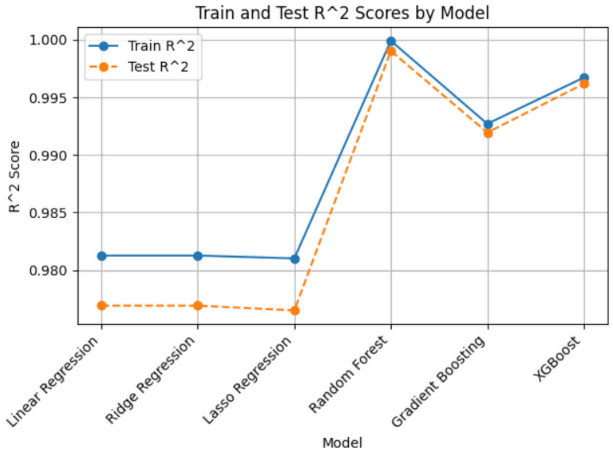

Fig. 6. Line plot for Train and Test R² scores by Models

Feature importance analysis was conducted to understand the contribution of each feature to the model's predictions. This step is crucial in interpreting the model, especially when dealing with complex algorithms like ensemble methods. The analysis reveals which features have the most significant impact on the target variable, allowing us to focus on the key drivers of power usage.

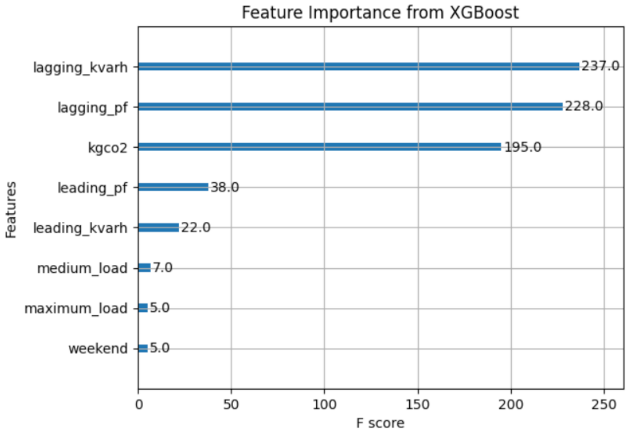

Fig. 7. Plot to show feature importance

Key insights from the feature importance analysis, as shown in Fig. 7, revealed that Lagging Reactive Power [kvarh] emerged as a critical predictor with high importance scores, indicating its significant influence on power usage variations[8]. The Power Factor also demonstrated substantial importance, underscoring its crucial role in predicting energy consumption and its impact on overall efficiency.

The Prediction Error Plot for the XGBoost model was generated to visually assess the accuracy of the predictions made by the model against the actual values of the target variable (usage_kWh).

The prediction plot shown in Fig. 8 below provides insights into the performance and reliability of the model across different ranges of predicted values.

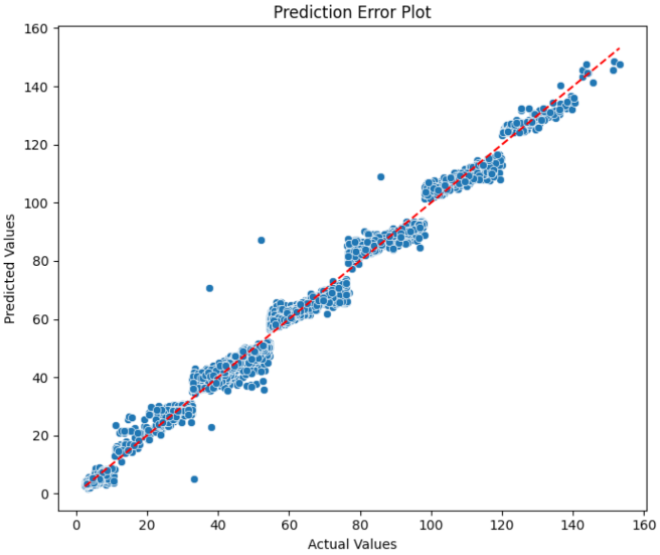

Fig. 8. Prediction Error Plot

The distribution of residuals percentage plot shown in Fig. 9 below provides insight into the accuracy of the model's predictions. Residuals represent the difference between the actual and predicted values, and when expressed as a percentage, they highlight the relative magnitude of prediction errors. It was observed that 21% of the predictions deviated by 10% or more from the actual power usage, indicating instances of overestimation or underestimation.

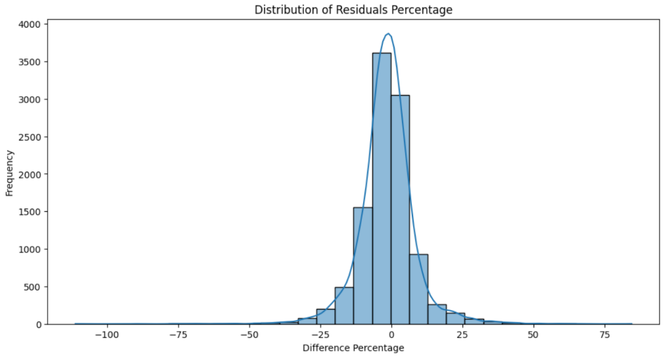

Fig. 9. Distribution of Residuals Percentage

V. CONCLUSION

In this study, various regression models were developed and evaluated to predict real-time energy consumption based on a set of features[9]. The analysis revealed that more complex models, while providing higher accuracy, often lacked interpretability. However, simpler models provided clearer insights into the key drivers of energy usage[10]. The final model selection was based on a balance between accuracy and interpretability, ensuring reliable predictions and a better understanding of the factors influencing energy consumption.

REFERENCES

[1] Margarita M., Carlos D., Iñigo C., Luis J.M.,, Iñaki A., and Fernando F. (2013), "The transition towards renewable energies: Physical limits and temporal conditions", *Energy Policy*, vol. 52, pp. 297-311.

[2] Vibhav S., Syed I., and Mirza B. (2012). "Energy-efficient communication methods in wireless sensor networks: A critical review", *International Journal of Computer Applications*, vol. 39, pp. 35-48, 02.

[3] Uci.edu. (2019). *UCI Machine Learning Repository*. [online] Available at: http://archive.ics.uci.edu [Accessed 21 Aug. 2024].

[4] Hai-xiang Z., and Frédéric M. (2012). "A review on the prediction of building energy consumption", *Renewable and Sustainable Energy Reviews*, vol. 16, no. 6, pp. 3586-3592.

[5] Kadir A., and Nora M. E. (2018). "A review of data-driven building energy consumption prediction studies", *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1192-1205.

[6] Yong X., Yonghua C., Jiaojiao X., and Zheyou C. (2022). "Research on sustainability evaluation of green building engineering based on artificial intelligence and energy consumption", *Energy Reports*, vol. 8, pp. 11378-11391.

[7] Mel K., Nor A. R., and Lilik J.A. (2021). "Energy consumption prediction by using machine learning for smart building: Case study in malaysia", *Developments in the Built Environment*, vol. 5, pp. 100037.

[8] Ran W., Shilei L., and Wei F. (2020). "A novel improved model for building energy consumption prediction based on model integration", *Applied Energy*, vol. 262, pp. 114561.

[9] Zhe W., Tianzhen H., Han L., and Mary A.P. (2021). "Predicting city-scale daily electricity consumption using data-driven models", *Advances in Applied Energy*, vol. 2, pp. 100025.

[10] VE S., Changsun S., and Yongyun C. (2021). "Efficient energy consumption prediction model for a data analytic-enabled industry building in a smart city", *Building Research & Information*, vol. 49, no. 1, pp. 127-143.