Power Consumption Prediction for Steel Industry Using Machine Learning Algorithms

Submitted by:
Sai Charith Ghanta
Sree Birudala

Objectives

The type of problem we are solving here is regression. The project's goal is to create a machine learning-based predictive model for estimating energy consumption in the steel sector, taking into account variables such as reactive power, power factors, and CO2 emissions for various load types. The basic goals are to identify important contributing elements, improve understanding of energy usage trends, and empower the steel sector to optimize energy consumption and decrease costs.

1. Background

Since 1981, the U.S. Department of Energy's Industrial Assessments Centers (IACs) have audited and analyzed energy usage data in many industries, giving useful insights and recommendations for improving energy efficiency. Several studies have demonstrated the effectiveness of machine learning (ML) and deep learning (DL) methodologies, such as Multiple Linear Regression (MLR), Decision Tree (DT), Artificial Neural Networks (ANN), and Recurrent Neural Networks (RNN), in accurately predicting power consumption in the industry sector using extensive datasets encompassing historical power usage, meteorological conditions, energy consumption patterns, and lighting data.

2. Dataset Overview

The dataset for our project is *"Steel Industry Energy Consumption"*
https://archive.ics.uci.edu/dataset/851/steel+industry+energy+consumption
The answer variable in the aforementioned dataset is "Usage_kwh". There are ten features, including date, lagging/leading reactive power, lagging/leading power factor, CO2, weekStatus, and day of week. Table 1 explains all of the features. The dataset has 35040 observations.

Feature	Description
Date	Data collected in real time on the first of the month
Usage_kWh	Energy Consumption in Industry kWh continuous
Lagging Current	Reactive energy kVarh Continuous
Leading Current	Reactive energy kVarh Continuous
CO2	CO2 Continuous ppm
NSM	Minutes and seconds since midnight S Continuous
Week status	Weekday or Weekend
Day of week	Sunday, Monday ..etc
Load Type	Light Load, Medium Load, Maximum Load

The screenshot showing the first six samples of the dataset is attached below.

```
> head(data)
              date Usage_kWh Lagging_Current_Reactive.Power_kVarh Leading_Current_Reactive_Power_kVarh CO2.tCO2.
1 01/01/2018 00:15      3.17                                 2.95                                     0         0
2 01/01/2018 00:30      4.00                                 4.46                                     0         0
3 01/01/2018 00:45      3.24                                 3.28                                     0         0
4 01/01/2018 01:00      3.31                                 3.56                                     0         0
5 01/01/2018 01:15      3.82                                 4.50                                     0         0
6 01/01/2018 01:30      3.28                                 3.56                                     0         0
  Lagging_Current_Power_Factor Leading_Current_Power_Factor  NSM WeekStatus Day_of_week  Load_Type
1                        73.21                         100  900    Weekday      Monday Light_Load
2                        66.77                         100 1800    Weekday      Monday Light_Load
3                        70.28                         100 2700    Weekday      Monday Light_Load
4                        68.09                         100 3600    Weekday      Monday Light_Load
5                        64.72                         100 4500    Weekday      Monday Light_Load
6                        67.76                         100 5400    Weekday      Monday Light_Load
```

3. Procedures/Methods

3.1 Data Preprocessing

i)Remove null values: We tested the dataset for null values and found none.
We extracted the feature from the date column and used it as a distinct feature.
ii)Remove any duplicate or similar features-In the dataset context, we discovered that one of the features, "Load_Type," is equivalent to the answer variable "Usage_kWh." The "Load_Type" attribute categorizes energy usage, whereas "Usage_kWh" measures the actual quantity of electricity consumed. Because our goal is to estimate energy, we decided to leave out the "Load_Type" attribute when building the predictive model.
Furthermore, after extracting the month from the Date feature, we removed the "Date" feature from the dataset. This decision was based on the availability of existing features such as "Week Status," "Day of the Week," and "Month," which give essential information without the necessity for the "Date" feature.
iii) Convert category features into numerical-We changed certain categorical features to numerical ones using the function as.numeric(factor()).
iv)Train-Test Splits-We've split the dataset into a "70% training set" and a "30% test set." This section evaluates the model's real-world performance by evaluating its ability to make accurate predictions on data that it did not encounter during training, demonstrating its practical value.

3.2 Data Visualization

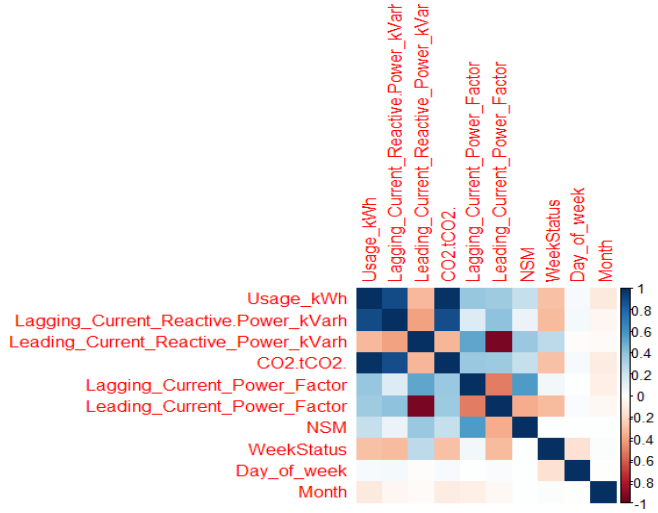


Figure 1 depicts the correlation matrix, which provides insights into the correlations between distinct variables and their links to the response variable. The investigation reveals significant connections between output responsiveness and key parameters such as lagging reactive power (kVarh), leading reactive power (kVarh), and CO2 emissions. Furthermore, moderate connections exist between the Lagging Power Factor, Leading Power Factor, and WeekStatus. Conversely, Days of the week, NSM, and month had weaker correlations with the response variable. Below is a screenshot of the computed correlation coefficients. Notably, "CO2" exhibits the greatest

```
> print(correlation_results)
Lagging_Current_Reactive.Power_kVarh Leading_Current_Reactive_Power_kVarh                CO2.tCO2.
                          0.89614990                          -0.32492178               0.98817977
       Lagging_Current_Power_Factor       Leading_Current_Power_Factor                      NSM
                          0.38596046                           0.35356571               0.23461033
                         WeekStatus                         Day_of_week                    Month
                         -0.29547483                           0.03986516              -0.11396069
```

correlation coefficient of 0.9881, followed by "Lagging Current Reactive Power" wi

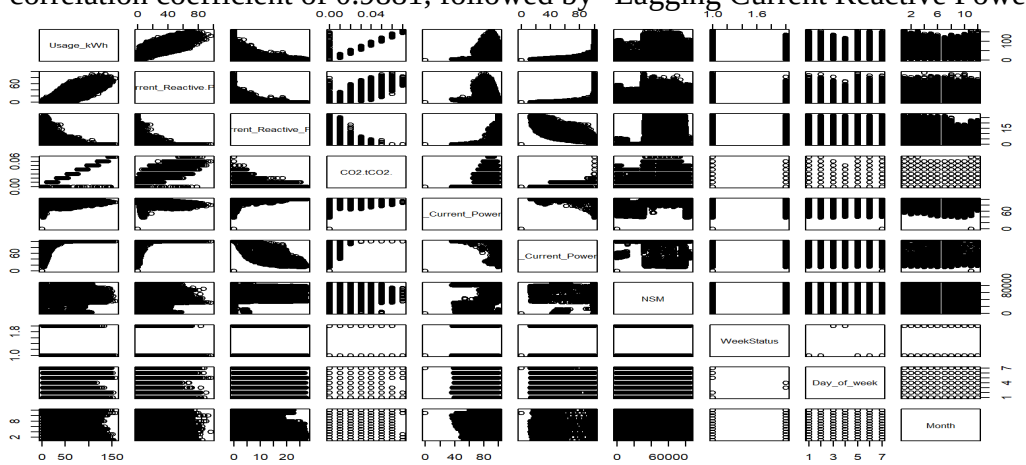


Figure 2 depicts a scatterplot matrix for each pair of variables, which represents a pairs plot for our dataset. Because this Figure is insufficient, we have created other scatterplots displaying numerous properties plotted against the response variable in Figures 3 through 6.

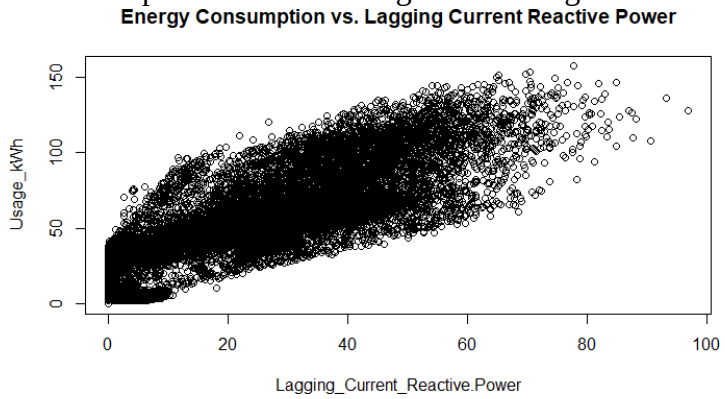
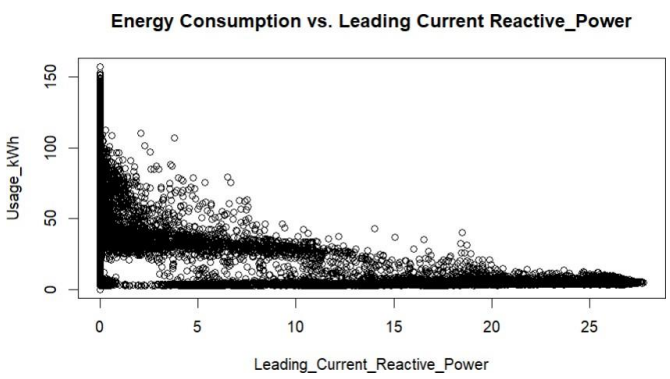


Figure 3 shows the linear relationship between the consumption of energy and lagging current reactive power.



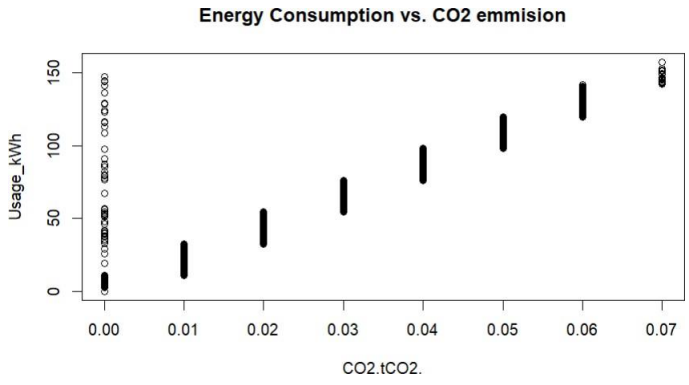


Figure 4 shows a noticeable relationship between energy use and leading current reactive power. As leading Current Reactive Power increases, Energy Consumption decreases.Figure 5 Energy use versus CO2 emission shows a clear linear association between energy consumption and CO2 emissions.

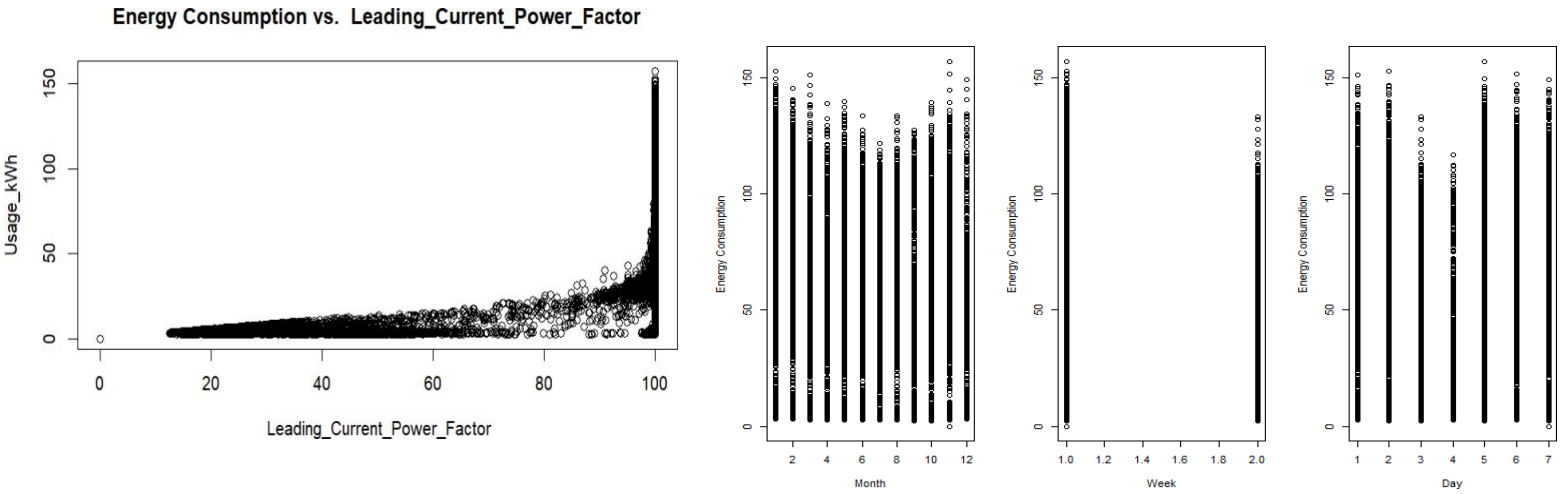


Figure 6 shows a clear trend: an increase in the Leading Current Power Factor corresponds to a decrease in energy usage.

Upon extracting the Month from the "date" feature, we utilized this newly derived "Month" feature to analyze energy consumption trends across different timeframes. The resulting plot in Figure 7. showcases energy consumption patterns for each Month, week status, and day. Notably, the plot reveals higher power consumption during November and over the weekends.

Figure 7: Energy Consumption by Month, Week Status, and Day

3.3 Evaluation methods/criteria

We evaluated the performance using k-fold cross-validation, R^2, and MSE.

i) K-folds Cross Validation.

We picked k = 5 because we had more extensive sample data.

Figure 8 depicts the functionality of 5-fold cross-validation (CV). This approach divides the training dataset into 5 folds. The first fold is the test set, while the remaining data is used to train the first model. Afterwards, the second fold serves as the test set, and the rest of the data serves as the training set for building the second model. This method is repeated until it includes all five folds. The average error produced from these iterations serves as the final evaluation statistic, offering a full assessment of the model's performance across various subsets of the data.

Figure 8. 5 Fold Cross- Validation

ii)The coefficient of determination (R^2)

iii)Mean Square Error (MSE)

3. 4. Statistical Learning Methods

We used numerous processes and techniques that we learned in class to find the best model that fits the goals of this project. These methods covered a variety of approaches.

i) Multiple Linear Regression Model
ii) Subset Selection
iii) LASSO Regression Model
iv) Ridge Regression Model
v) Principal Component Analysis (PCA)
vi) Random Forest Model
vii) Random Forest Model with CV

4. Results and Discussion

i) Linear Regression Model

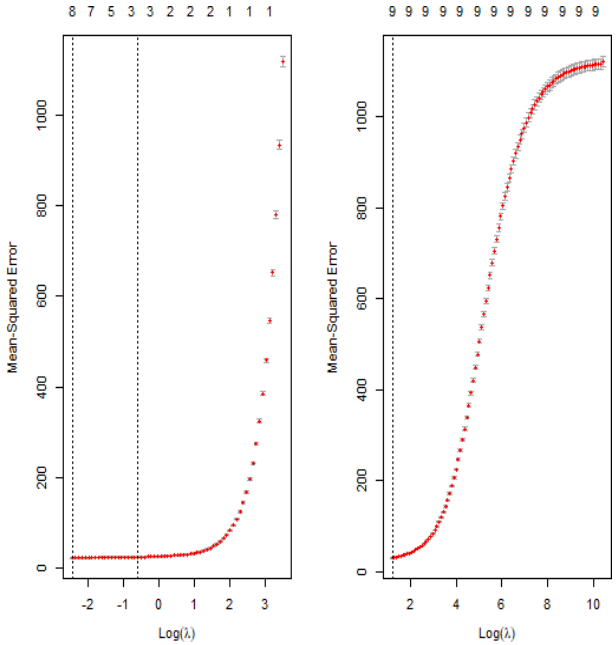
We have applied multiple linear regression models to all features. Based on the summary table below for this model, we observed that all features are significant as their p-value is less than the significant threshold of 0.05. We observed an MSE of 22.85371 and an R-squared of 0.979435.

ii) Subset Selection

For subset selection, we applied the forward subset selection method. Based on adjusted R-squared, the minimum number of variables is 8. The plot for Adjusted R-squared vs. the number of variables and best subset coefficients are provided below. From subset selection, almost all features are relevant. With the eight features based on subset selection, we applied a linear regression model and achieved an MSE of 22.85 and an R-squared of 0.97.

iii) LASSO and Ridge Regression Models

We formatted the data to be compatible with the *glmnet* package, which includes ridge and lasso functions. Using the training data, these models were trained using cross-validation to find the best lambda value (minimum value). Figure 10. depicts MSE for Lasso and Ridge's model based on different lambda values are provided below. We got the optimal lambda of 0.078 and 3.309 for the Lasso and Ridge models, respectively. Also, the Lasso model achieved an MSE of 2 2.988, whereas the ridge model achieved an MSE of 36.484. Hence, the lasso model was better as it achieved less MSE than the ridge model. Furthermore, the lasso model has one Zero coefficient indicating all relevant features except Leading_Current_Reactive_Power_kVarh.



iv) Principal Component Analysis (PCA)

We have applied PCA for feature selection. In the summary table, we sorted the features by their contribution using PCA, where the NSM has a higher contribution, and weekStatus has less contribution. However, all the features are significant, having a p-value less than the threshold of 0.05. In addition, we used the top seven features shown in the summary table, ranked by PCA, to

train and test linear models. We achieved an MSE of 22.869 and an R-squared of 0.9794.

v) Random Forest Model

After exploring the above models, we planned to use Random Forest to achieve our project's goal. We applied RF on training data to train a model and tested that model on test data by calculating MSE and R-square. We achieved an MSE of 2.752 and an R-squared of 0.997. We can see that these results are better than all previous models.

vi) Random Forest model with 5-fold Cross-validation

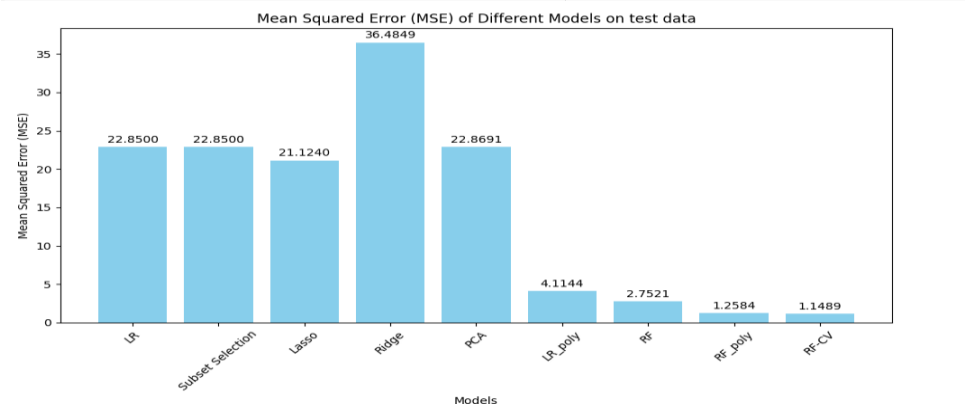
Based on the results from the RF model. We have sought to improve it by using cross-validation for validation purposes on train data and the trained model to test on test data. We opted for a 5-fold cross-validation at the end, as we tried using 3-fold, which didn't significantly improve the last model to justify the computational cost. We achieved an MSE of 1.1488 and an R-squared of 0.999. Based on the results, we can say that the cross-validation increased the accuracy of the model while mitigating the overfitting problem we faced in the regular RF model

vii) polynomial Linear Regression Model

Furthermore, this project employed a Polynomial Linear Regression model of degree 2 to forecast energy consumption accurately. By incorporating higher-order polynomial terms in the regression equation, the model effectively captured intricate nonlinear relationships between the predictors and the target variable ('Usage_kWh'). The model's robustness was evidenced by its remarkable performance metrics, achieving an MSE of 4.1144 and an outstanding R-squared value of 0.9962.

viii) Random Forest Model with polynomial features

This project also employed a Polynomial Random Forest model of degree 2 to forecast energy consumption accurately. The model's robustness was evidenced by its remarkable performance metrics, achieving an MSE of 1.2584 and an outstanding R-squared value of 0.9988.



The results of the above bar chart suggest that models incorporating polynomial features (linear regression and random forest) outperformed simpler linear models. The Random Forest incorporated with cross-validation also performed better than traditional linear models. The Random Forest model with cross-validation provided the most accurate predictions, achieving a test MSE of 1.148. Based on our results, we selected RF with CV model as a proposed model for predicting energy consumption.

Conclusion

Throughout this project, our comprehensive exploration of diverse statistical learning methods on the Steel Industry Energy Consumption dataset has been fruitful. Among the models employed to detect power consumption, our proposed model notably emerged as the leading solution. The Random Forest model, further strengthened by 5-fold Cross-Validation, showcased exceptional performance, boasting a remarkable R-squared value of 0.999082 and achieving a notably low test

Mean Squared Error (MSE) of 1.148884. This outstanding success in accurately predicting energy consumption represents a pivotal advancement toward fostering heightened efficiency and substantial cost reduction within the steel industry. It signifies a promising stride towards optimizing energy utilization and potentially revolutionizing energy management practices in this sector.