# 1 Introduction

Energy is a primary raw material essential for the steel industry, known for its high energy demands. Traditionally, statistical models have been employed to predict energy usage. Now though, machine learning algorithms and advanced prediction models have revolutionized our approach, enabling us to process extensive datasets and generate more accurate predictions. In this project, I aim to analyze the relation between different features of the dataset and create a predictive model for energy consumption estimation for the steel industry based on the provided dataset on Kaggle.

# 2 Data Exploration

As we delve into the data, at first, we focus on exploring the dataset's features. The Table 1 offers a comprehensive overview of the features of this dataset. In addition, as shown in Figure

Table 1: Description of the features used in the dataset.

| **Features** | **Description** |
|---|---|
| Date | Date recorded for one year. |
| Usage (kWh) | Energy consumption recorded in real time. |
| Lagging Current | Reactive energy in kVarh. |
| Leading Current | Reactive energy in kVarh. |
| CO2 (tonn) | CO2 ppm emitted. |
| NSM | Number of Seconds from Midnight. |
| WeekStatus | Indicates whether the day is a weekday or weekend. |
| Day of Week | Weekdays from Monday to Sunday. |
| Load Type | Type of Load (min, med, max) during production. |

1, the heat map provides a visual representation of the relation between features of the dataset.

## 2.1 Data Prepration

By checking the dataset, a serious problem with date and time was founded e.g after '2018-01-01 23:45:00' the date and time '2018-01-01 00:00:00' is given while normally it should be '2018-01-02 00:00:00'. Therefore, an additional function called ***adjust_date*** is used to reformat the date and time and shift the days with '00:00:00' time to the next day.

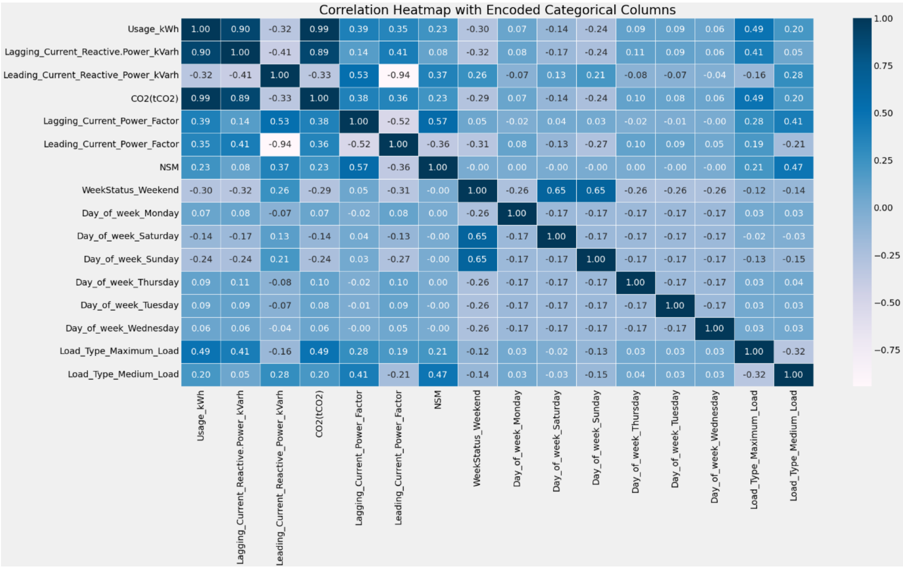


Figure 1: Heat map visualization of the dataset.

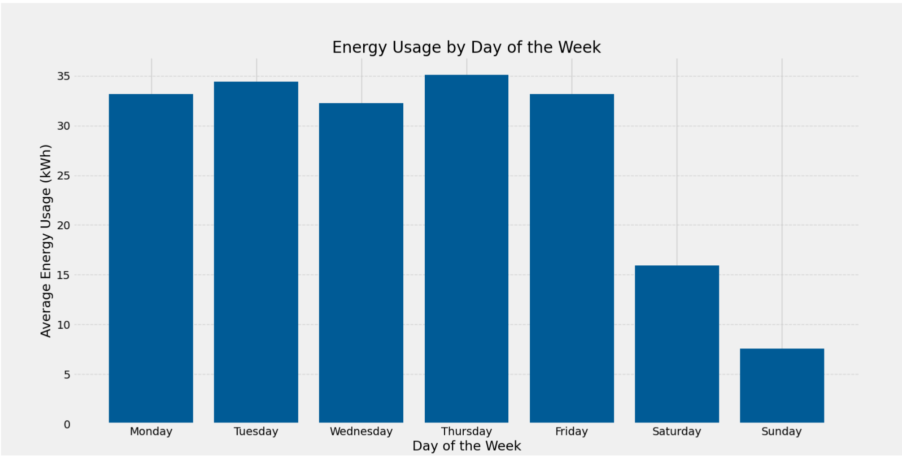


Figure 2: Average Daily Energy Usage

## 2.2 Daily Energy Usage Insights

The energy usage power probably is not the same for each day of the week. Therefore, the pattern should be identified first. This would help determine the days on which the most energy is used. Fig. 2 shows the average usage of power on a daily basis. In addition, the total energy used in each day is visible on figure 3. Also, referring to figure 4, a clear and concise visual representation of the distribution and variability of energy usage across the days of the week is presented. Box plots help to identify the patterns, trends, and anomalies in power consumption with ease and help to assess how usage fluctuates.

As clearly illustrated in the presented graphs, an unusual pattern in energy consumption is evident during weekends. This suggests that the operational hours or workload likely decrease significantly on Saturdays. Sunday could be a non-operational day for this particular industry. Also, between days 0 to 2 (Monday-Wednesday) and day 4 (Friday) energy consumption ap-

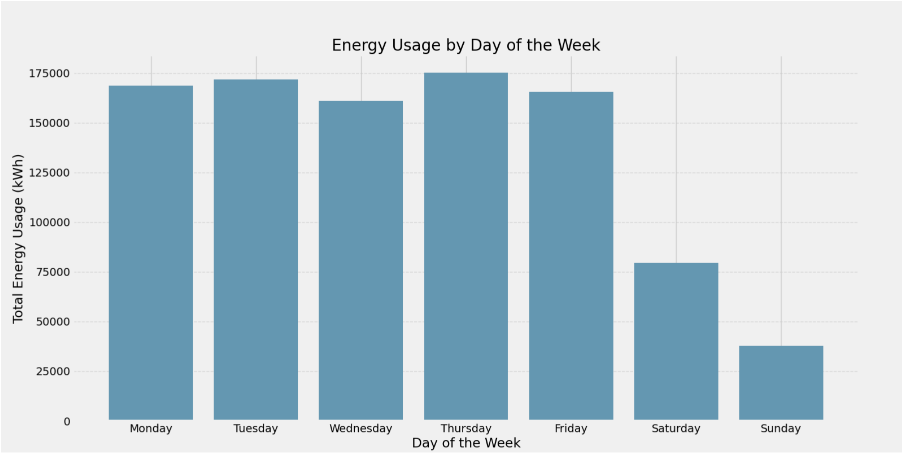


Figure 3: Total Energy used on Each Day

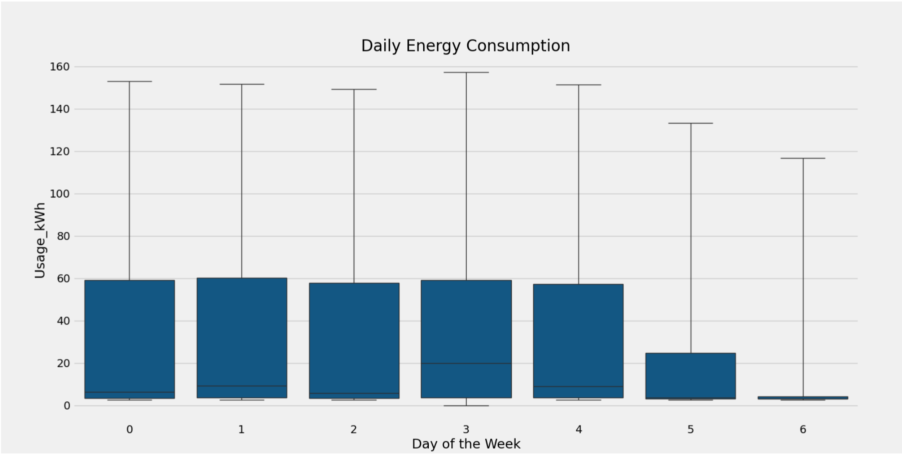


Figure 4: Distribution of Daily Energy Consumption Starting from Monday (No. 0)

pears relatively stable suggesting consistent energy usage with moderate variability. However, there's a slight decrease in median energy consumption (around 20 kWh) on day 3 (Thursday), indicating less variability in usage compared to earlier days.

## 2.3 Hourly Energy Usage Insights

Identifying the usage pattern during a single day is also important. So, high consumption patterns during the day should be located. Box plot on figure 6 would represent the general overview of the whole dataset which contains the whole year. The following graphs show a clear daily cycle of energy consumption with high usage in the morning (06:00-09:00) and late afternoon/early evening (15:00-18:00), moderate usage in the late morning and evening, and very low usage at night and during a notable dip from 12:00-14:00. This pattern aligns with a typical industry weekday routine.

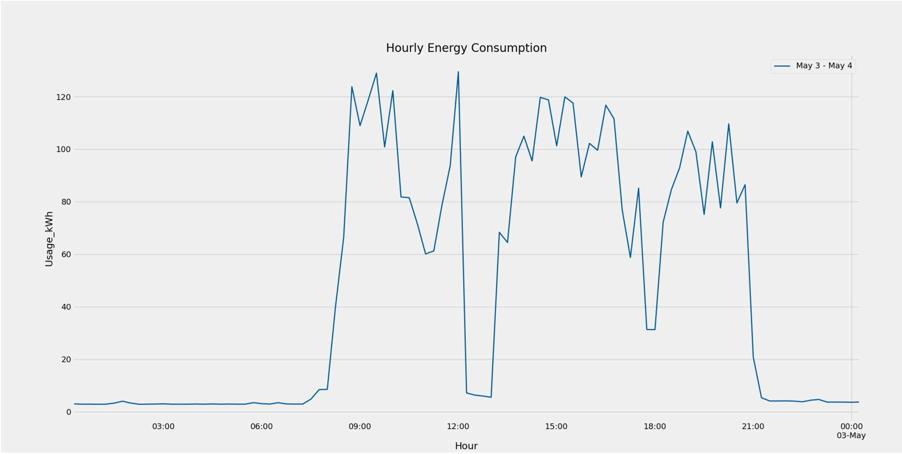


Figure 5: Hourly Usage of Energy on a Random Day (3th May - 4th May 2018)

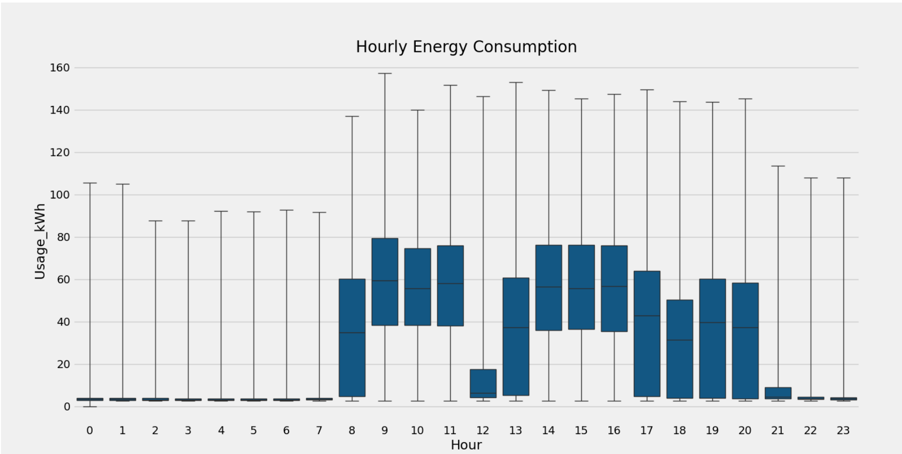


Figure 6: Distribution of Hourly Energy Consumption

## 2.4 Power Quality and Efficiency Analysis

In energy efficiency, there is a concept called ”Power Factor”. It simply describes how efficiently electricity is being used and it is the ratio of useful power to the total power. In our dataset, two columns ”Lagging_Current_Power_Factor” and ”Leading_Current_Power_Factor” provide the data about power factors resulting from lagging and leading current. The mean of these values is calculated and then reported through the whole year and a random weak in figures 7 and 8. Also, based on the equation 1, the total efficiency of the power consumption is calculated. and is equal to $87.14\%$.

$$\text{Efficiency Percentage} = \left( \frac{\sum_{\text{Power\_Factor}>0.9} \text{Usage\_kWh}}{\sum \text{Usage\_kWh}} \right) \times 100 \tag{1}$$

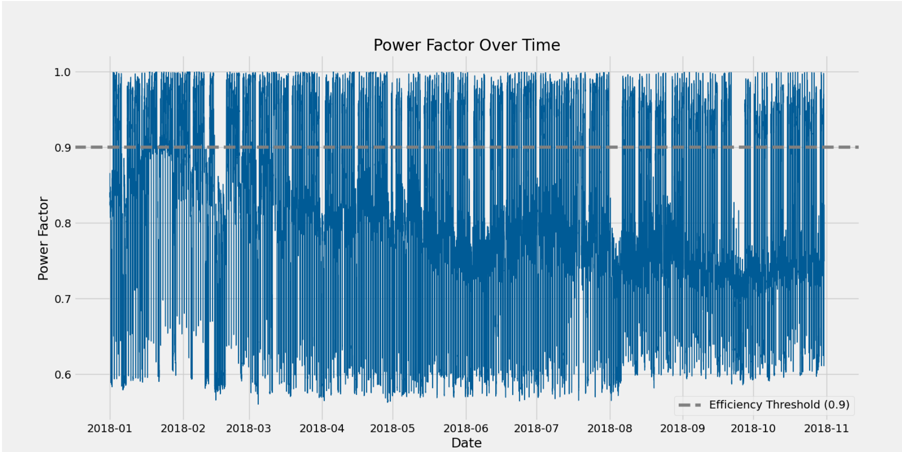


Figure 7: Power Factor Changes During the Period of Measurement

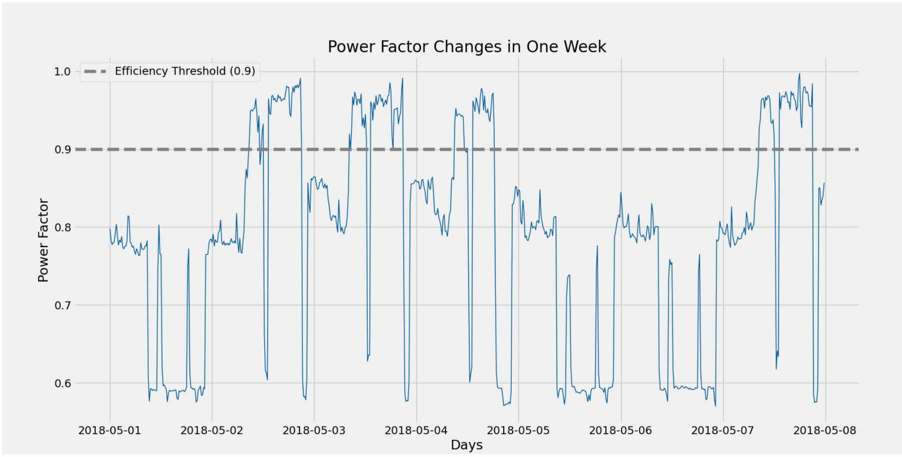


Figure 8: Power Factor Changes During the Period of One Random Week

### 2.5 Carbon Footprint and Sustainability

The dataset indicates three different load categories: Maximum, Medium, and Light load. Each load type is likely to have different levels of influence on $CO_2$ emissions. To visually assess the impact of each load type, a pie chart can be used. Figure 9 demonstrates a pie chart that represents the proportionate effects of these different load categories on $CO_2$ emissions. Considering the pie chart, higher load conditions (maximum and medium) are the main drivers of $CO_2$ emissions, with maximum load being the most significant contributor.

## 3 Model Creating

With a fundamental understanding of the entire dataset and system, now it is very beneficial to have a predictive model capable of estimating the energy consumption based on the specified features given within our dataset. In the following, the procedure to create a predictive model based on Gradient Boosting algorithm is explained. Gradient boosting algorithms work itera-

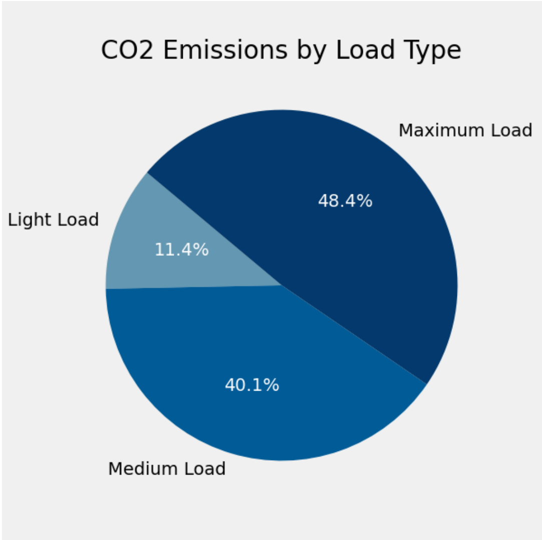


Figure 9: Proportionate Effects of Different Load Types on $CO2$ emissions

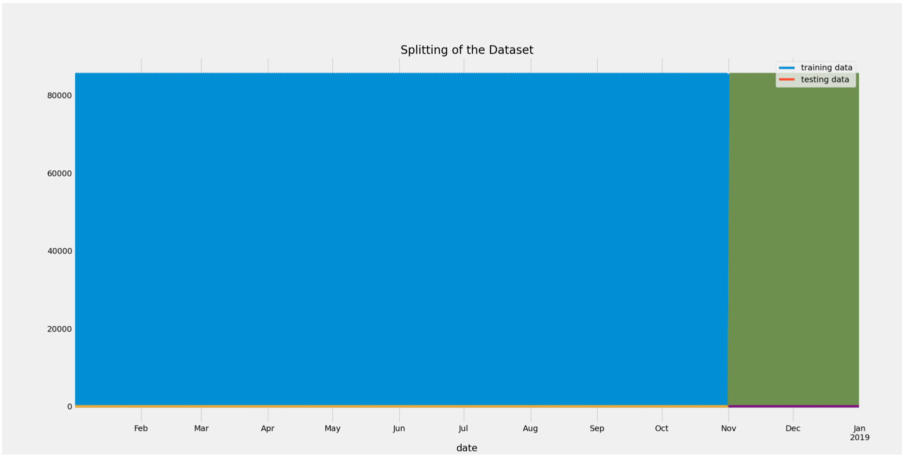


Figure 10: Splitting the Dataset to Training and Testing Sets

tively by adding new models sequentially, with each new addition aiming to resolve the errors made by the previous ones.

### 3.1 Preparing the Dataset

It is obvious that to train our model, we need to split our dataset into two different sections. One section is for training and the other part is for testing our model. Figure 10 presents how the dataset is splitter into train and test sets. It can be mentioned as a percentage, but it is much easier to say that 2 months of our dataset is selected as a testing dataset.

### 3.2 Train the Model

As previously indicated, the Gradient Boosting algorithm provided by the XGBoost library is used in this study. Given that this prediction is conducted via a regression task, the ***XGBRe-***

***gressor*** specifically from the XGBoost library is utilized. In Table 2 the most critical hyperparameters related to the prediction model configuration are presented.

| **Hyperparameter** | **Value** | **Description** |
|---|---|---|
| `n_estimators` | 50000 | Number of boosting rounds (trees) to build. |
| `learning_rate` | 0.02 | Step size shrinkage to prevent overfitting; controls contribution of each tree. |
| `max_depth` | 5 | Maximum depth of a tree; limits complexity to prevent overfitting. |
| `min_child_weight` | 3 | Minimum sum of instance weight (hessian) needed in a child node. |
| `subsample` | 0.85 | Fraction of training data sampled for each tree; reduces overfitting. |
| `colsample_bytree` | 0.85 | Fraction of features sampled for each tree; adds randomness. |

Table 2: Hyperparameters of the `xgb.XGBRegressor` model.

# 4 Results and Model Metrics

In regression model analysis, it is common to report RMSE (Root Mean Square Error) or MSE (Mean Square Error) to report the accuracy of the model. However, as shown in table 3, both the Mean Absolute Error and R-squared are also presented as additional metrics. In terms of

| **Metric** | **Value** | **Description** |
|---|---|---|
| MAE | 0.3202 | Mean Absolute Error: Average absolute difference between predicted and actual values. |
| MSE | 0.4181 | Mean Squared Error: Average squared difference between predicted and actual values. |
| RMSE | 0.6466 | Root Mean Squared Error: Square root of MSE; measures error in the same units as the target. |
| $R^2$ | 0.9996 | R-squared: Proportion of variance in the dependent variable explained by the model. |

Table 3: Performance metrics of the model.

the model's performance metrics, as well as the characteristics and size of the dataset, it can be claimed that the model shows a very good performance. Following, a visual representation is provided to illustrate the difference between the actual and predicted values.

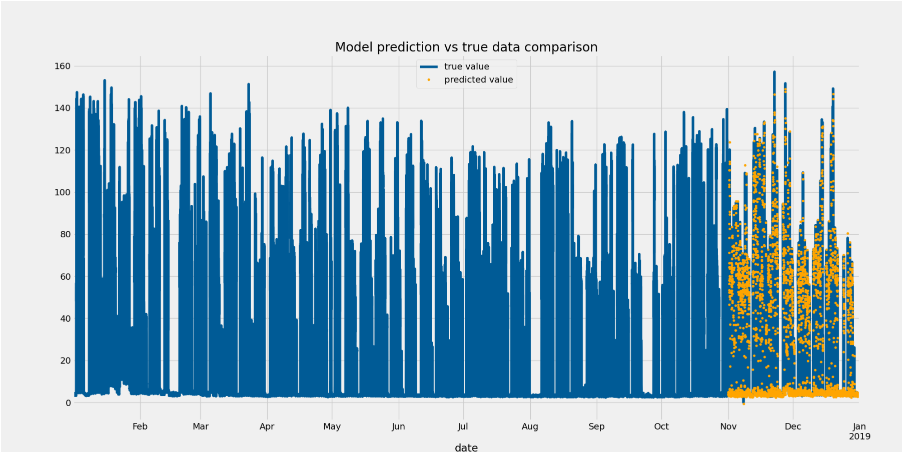


Figure 11: Total prediction of the model

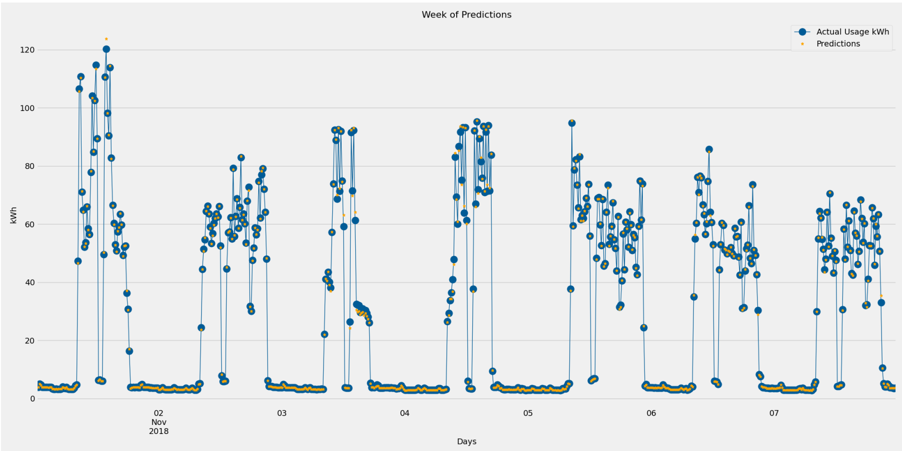


Figure 12: Randomly chosen week prediction

# 5 Conclusion

In colclusion, during this project a predictive model using Gradient Boosting algorithm (XGBoost) was developed to forecast energy consumption in steel industry. It can also be said that with an $R^2$ of 0.9996, a high precision is achieved. Overall, The analysis shows different patterns of daily and hourly energy consumption. Also, other findings such as power efficiency, and load types' impact on $CO_2$ emissions, are beneficial to facilitate energy management and sustainability in the industry.