SHORT-TERM LOAD CHARACTERIZATION AND BASELINE FORECASTING FOR A STEEL PLANT ENERGY DATASET – PART I: EXPLORATORY ANALYSIS AND SIMPLE MODELS

Bogdan Diaconu, *University "Constantin Brâncuși" of Tg-Jiu, ROMANIA*
Lucica Anghelescu, *University "Constantin Brâncuși" of Tg-Jiu, ROMANIA*
Mihai Cruceru, *University "Constantin Brâncuși" of Tg-Jiu, ROMANIA*

ABSTRACT: Publicly available industrial energy datasets are increasingly used to benchmark machine learning methods for load prediction. However, many studies focus directly on complex algorithms and report high accuracy without first analysing the temporal structure of the process or establishing strong, transparent baselines under realistic validation. This paper presents Part I of a two-part study on the "Steel Industry Energy Consumption" dataset, a one-year, 15-minute resolution time series from a steel manufacturing plant. We first perform a compact but systematic exploratory analysis of the dataset, examining daily and weekly load patterns, the distribution of energy consumption across declared operating regimes, and the relationship between active energy, reactive energy and associated CO_2 emissions. On top of this characterization, we construct several short-term forecasting baselines for one-step-ahead prediction of active energy usage: a mean model, naive and seasonal naive benchmarks, and a linear regression model using only calendar and categorical variables (week status, day of week and load type). Models are trained on the first ten months of 2018 and evaluated on the last two months to preserve temporal ordering. Results show a highly structured temporal behaviour, with pronounced production windows on weekdays and reduced operation at weekends. Load type labels correspond to distinct operating regimes, though with some overlap. The linear regression baseline clearly outperforms the naive and seasonal naive models, achieving a mean absolute error of about 2.35 kWh and a root-mean-square error of about 3.50 kWh on the test set. The study demonstrates that simple models, when evaluated with time-aware splits, already provide strong performance and interpretable insight, and therefore form a necessary reference point for more sophisticated machine learning approaches that will be analyzed in Part II.

Key-Words: industrial energy consumption; steel industry; load forecasting; time series; baseline models; exploratory data analysis

1.INTRODUCTION

Electric steelmaking is one of the most energy-intensive industrial activities, and electricity costs represent a major share of operating expenses in modern steel plants. Accurate short-term forecasting of electrical load is therefore important for production scheduling, demand-side management and participation in electricity markets. In parallel, decarbonization policies and carbon pricing increase the interest in reliable, data-driven models that explain and predict both energy usage and associated emissions. In the last few years, several open datasets from industrial facilities have appeared on public platforms such as Kaggle [1]. They provide a convenient benchmark for testing machine learning models and for teaching energy analytics. One such dataset is the "Steel Industry Energy Consumption" dataset [1], which covers a full year of operation in a steel plant at a 15-minute resolution and includes active and reactive energy, power factor, derived CO_2 emissions and categorical indicators of operating regime. Much of the published work that uses this dataset focuses on comparing algorithms—random forests, gradient boosting, artificial neural networks, recurrent neural networks—under various feature engineering choices. While these studies are valuable [2-6], they often devote limited space to a detailed exploratory analysis of the dataset and to the construction of simple, transparent baseline models. As a result, reported performance numbers may be difficult to interpret, especially when random train–test splits are used without regard to the time-ordered nature of the data. The goal of this paper is to take a step back and answer two basic questions. First, what does this dataset actually look like in terms of temporal patterns and operating regimes? Second, how well can we already predict short-term load using very simple models and a realistic temporal split between training and test data? Addressing these questions is important both for practitioners who might use the dataset and for researchers who wish to assess the added value of more sophisticated methods.

This contribution is Part I of a two-paper series. Here we focus on exploratory data analysis and baseline modelling. Part II will build on these results to investigate more advanced machine learning models, time-aware validation strategies and model interpretability techniques.

2. DATASET AND PRE-PROCESSING

The Steel Industry Energy Consumption [1] dataset comprises 35,040 records collected during 2018 with a sampling interval of 15 minutes. After parsing the time stamp and sorting chronologically, the dataset spans from 1 January 2018 00:00 to 31 December 2018 23:45.

The main variables considered in this study are:

- **Usage_kWh** – active electrical energy in kWh for each 15-minute interval (target variable);
- **Lagging_Current_Reactive.Power_kVarh** and **Leading_Current_Reactive_Power_kVarh** – lagging and leading reactive energy;
- **CO2(tCO2)** – equivalent CO_2 emissions, presumably obtained by applying an emissions factor to the active energy;
- **Lagging_Current_Power_Factor** and **Leading_Current_Power_Factor**;
- **NSM** – number of seconds from midnight (0–86,400);
- **WeekStatus** – categorical flag distinguishing weekdays from weekends;
- **Day_of_week** – categorical day name;
- **Load_Type** – categorical label describing the operating regime: Light_Load, Medium_Load or Maximum_Load.

For the analyses below, we derive additional calendar features from the timestamp: hour of day and integer day-of-week index. No missing values were identified, so no imputation was carried out. To mimic a realistic forecasting scenario and avoid information leakage, we adopt a strictly chronological train–test split. The first ten months (January–October 2018) are used as the training set, and the last two months (November–December 2018) constitute the test set on which all performance metrics are computed.

3.EXPLORATORY DATA ANALYSIS

3.1 Daily and weekly load patterns

To understand the temporal structure of the load, we first examine average daily profiles. For each combination of hour of day and week status (weekday or weekend), we compute mean energy usage. The resulting profiles represented in Figure 1, show a very clear pattern. On weekdays, there is a low "base load" during the night and early morning. Around 08:00–09:00 the load rises sharply, reaching a high plateau that extends through most of the working day. After 20:00 the load decreases again towards evening levels. Weekends display a similar shape but at significantly lower magnitude across all hours, suggesting reduced production activity and possible maintenance. This behavior indicates that much of the variation in short-term load is driven by the production schedule, which is strongly aligned with the calendar. Even without considering more detailed process variables, time-of-day and the weekday/weekend label already provide a strong explanatory signal for the load.

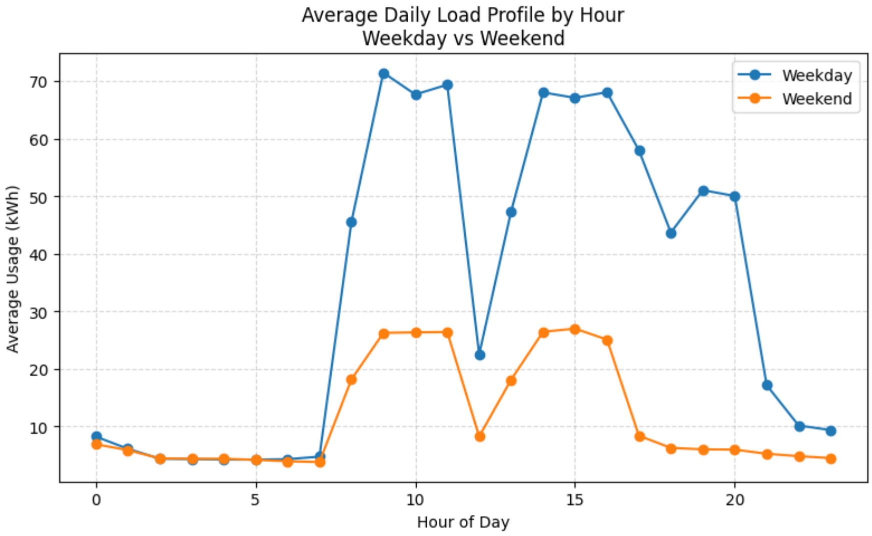

Figure 1. Load pattern by regular work day and weekend

3.2 Operating regimes: distribution by load type

The dataset includes a categorical variable, Load_Type, that indicates Light, Medium or Maximum load. We analyse the distribution of Usage_kWh within each category. Light_Load intervals cluster tightly around low energy values, typically between 3 and 6 kWh, with a narrow spread. Medium_Load data cover a much broader range, extending from a few kWh up to around 60 kWh, while Maximum_Load intervals have higher median values and reach well above 100 kWh. This confirms that Load_Type corresponds to distinct operating regimes in the plant: low-load standby or auxiliary operation, normal production and high-load periods. At the same time, the ranges of Medium and Maximum load overlap to some degree, which hints at varying equipment combinations or process steps within each regime. From a modelling standpoint, Load_Type is therefore a meaningful explanatory variable for energy usage and should be included in any baseline.

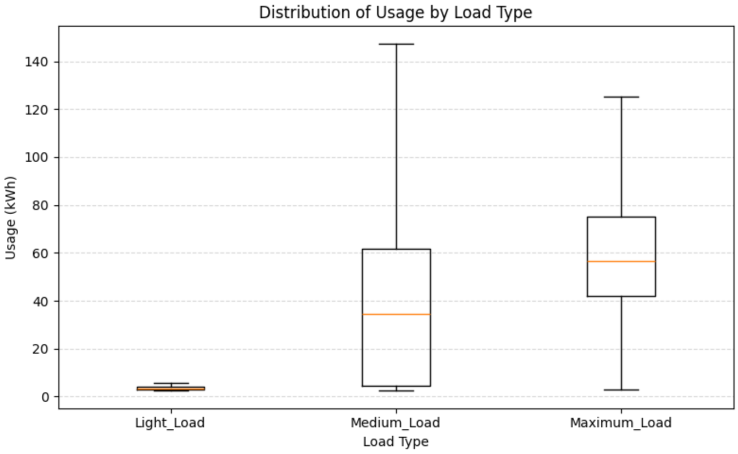

Figure 2. Usage distribution by load type

3.3 Relationship between energy usage, CO_2 and reactive energy

We next examine how active energy usage relates to CO_2 emissions and reactive energy. A scatter plot of Usage_kWh versus CO_2 (tCO_2) reveals an almost piecewise linear relationship: data points fall on a small number of distinct lines, each corresponding to a fixed ratio of CO_2

per kWh. This behaviour is consistent with emissions being calculated from energy consumption via an emission factor. It also means that, in this dataset, CO_2 does not provide additional independent information once Usage_kWh is known. For forecasting electricity usage, CO_2 can therefore be safely omitted or treated as a redundant variable. The relationship between Usage_kWh and Lagging_Current_Reactive.Power_kVarh is also strongly positive, but with more dispersion.

Higher active energy usage tends to be accompanied by higher lagging reactive energy, reflecting the inductive nature of industrial loads. The spread around the trend suggests differences in power factor associated with different equipment or process stages. Unlike CO_2, reactive energy and power factor variables hold physical information about the plant's operating state and may thus be valuable predictors or indicators of efficiency.

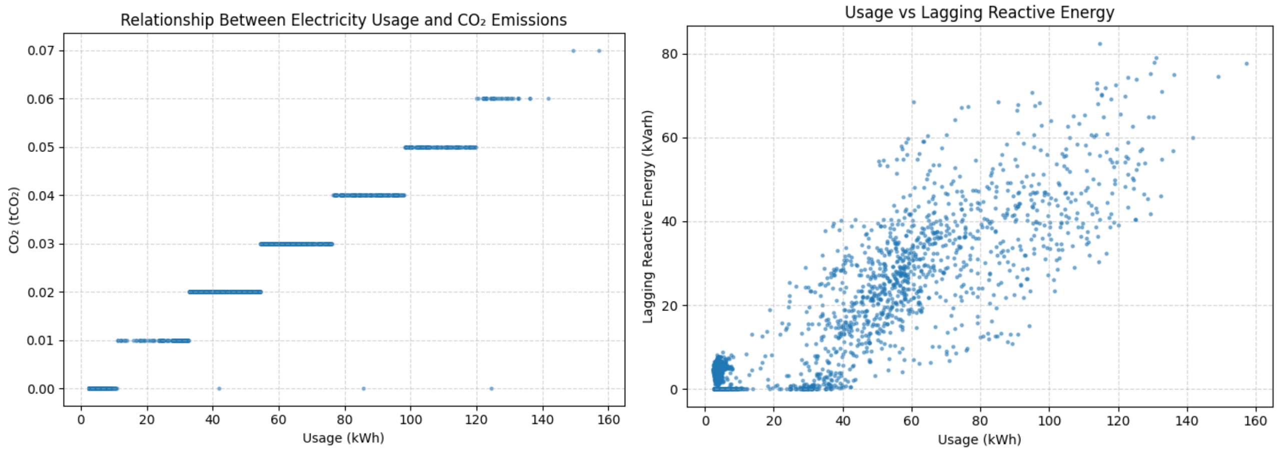

Figure 3. CO_2 emissions and Lagging Reactive Energy versus Usage (left and right, respectively)

4. Baseline Forecasting Models

4.1 Problem formulation and evaluation metrics

We consider one-step-ahead forecasting of energy usage at the 15-minute horizon. At each time step t, given information up to time t, the task is to predict Usage_kWh at time $t + 1$. Model performance is evaluated on the November–December 2018 test set using three standard metrics:

- **Mean Absolute Error (MAE)** – average absolute difference between predicted and actual values;
- **Root-Mean-Square Error (RMSE)** – square root of the mean squared error, emphasizing larger deviations;
- **Mean Absolute Percentage Error (MAPE)** – average absolute error as a percentage of the actual value, computed only for non-zero loads.

4.2 Baseline models

Four baseline models are implemented:

1. **Mean model**
   The predictor always outputs the mean load observed in the training set. It is an intentionally weak baseline.
2. **Naive model (last value)**
   The forecast for the next interval equals the last observed value: $\hat{y}_{t+1} = y_t$. This is a standard benchmark in time series analysis.
3. **Seasonal naïve model (same time yesterday)**
   The forecast uses the value at the same time one day earlier: $\hat{y}_{t+1} = y_{t+96}$, where 96 time steps correspond to 24 hours. This captures purely daily seasonality.
4. **Linear regression with calendar and categorical features**
   A multiple linear regression model is built using only exogenous features derived from the timestamp and categorical variables: hour of day, integer day-of-week, and one-hot encoded WeekStatus, Day_of_week and

Load_Type. No autoregressive terms (lagged Usage_kWh) are included in this Part I baseline, to keep the model transparent and easy to interpret. The model is fitted on the training period and then applied to the test period without re-estimation.

Table 1. Performance metrics of the models considered

Model	MAE	RMSE	MAPE (%)
Mean (train average)	27.575	31.281	419.52
Naïve (y_t-1)	5.197	11.845	20.93
Seasonal Naive (y_t-96)	14.363	26.662	133.501
Linear Regression (calendar + categorical)	2.351	3.495	20.174

4.3 Results

Table 1 summarizes the performance of the four baselines on the test set.

- Mean model: MAE ≈ 27.6 kWh, RMSE ≈ 31.3 kWh, MAPE ≈ 420 %.
- Naive model (last value): MAE ≈ 5.20 kWh, RMSE ≈ 11.8 kWh, MAPE ≈ 20.9 %.
- Seasonal naive model (same time yesterday): MAE ≈ 14.4 kWh, RMSE ≈ 26.7 kWh, MAPE ≈ 133.5 %.
- Linear regression (calendar + categorical): MAE ≈ 2.35 kWh, RMSE ≈ 3.50 kWh, MAPE ≈ 20.2 %.

As expected, the mean model performs very poorly and serves only as a lower bound. The naive last-value model already provides a relatively strong baseline: predicting the next 15-minute load by simply carrying forward the current value yields a MAE of about 5.2 kWh. Interestingly, the seasonal naive model based on daily periodicity performs worse, indicating that short-term fluctuations and intra-day dynamics are more informative than the pure daily pattern for this plant. The linear regression baseline, despite its simplicity and lack of autoregressive terms, clearly outperforms the naive models in terms of MAE and RMSE. With an MAE around 2.35 kWh and RMSE around 3.5 kWh, it captures much of the systematic variation in the load using only clock time, weekday/weekend status and operating regime labels. The MAPE of the linear model is similar to that of the naive model; this is partly due to periods of very low load, where even small absolute errors translate into large percentage errors.

5. DISCUSSION

The exploratory analysis and baseline results lead to several observations that are relevant both for practical use of the dataset and for future methodological work. First, the steel plant's operation is strongly structured in time. Clear production windows exist on weekdays, and weekends operate at significantly lower levels. Any forecasting model that ignores calendar information is bound to miss a large part of the signal. Conversely, even very simple models that rely only on such information, as demonstrated by the linear regression baseline, can achieve surprisingly good performance. Second, the provided Load_Type labels correspond to distinct operating regimes with well-separated typical energy usage ranges. Their inclusion in the regression model likely explains a significant share of the improvement over purely time-based benchmarks. In practice, this underlines the importance of combining process knowledge (here, operating regime) with time-series structure. Third, the analysis of CO_2 emissions reveals that they are effectively a deterministic transformation of active energy usage. While this is not surprising, it serves as a reminder that including such variables without scrutiny can inflate the apparent dimensionality of the feature space without adding genuine information. In contrast, reactive energy and power factor variables carry physically meaningful information about the nature of the loads and may be particularly relevant for power quality or efficiency studies in future work.

Finally, the relatively strong performance of the naive last-value model emphasizes the need for rigorous baseline comparisons. Any complex machine learning model applied to this dataset should be evaluated against both the naive benchmark and the simple linear regression with calendar and categorical features, using a time-ordered split similar to the one adopted here. Improvements that do not clearly exceed these baselines may not justify the added complexity.

CONCLUSIONS

This paper presented a first, deliberately simple step in analysing the Steel Industry Energy Consumption dataset. We provided a concise overview of the data, explored daily and weekly load patterns, examined the role of operating regimes and investigated the relationships between active energy, reactive energy and derived CO_2 emissions. Based on this understanding, we developed and evaluated four short-term forecasting baselines for 15-minute ahead prediction of energy usage: mean, naive, seasonal naive and a linear regression using only calendar and categorical features. Using a realistic training–test split that respects the time ordering of the data, the linear regression model achieved a mean absolute error of approximately 2.35 kWh and a root-mean-square error of 3.5 kWh, clearly outperforming naïve alternatives. These results show that even very simple models, when combined with appropriate validation, can provide strong and interpretable performance on this industrial dataset. They also provide a reference level against which more advanced methods must be compared. In the companion Part II paper, we plan to extend this work in three directions: (i) incorporating autoregressive terms and lagged process variables into the feature set; (ii) evaluating tree-based and neural models under time-aware cross-validation schemes; and (iii) applying model-agnostic interpretability tools to quantify the relative importance of calendar, regime and electrical variables. Together, the two parts aim to provide both a rigorous benchmark for data-driven energy analytics in industrial settings.

REFERENCES

1. https://www.kaggle.com/datasets/csafrit2/steel-industry-energy consumption?resource=download
2. Sathishkumar V E, Changsun Shin, Youngyun Cho, 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, 2021.
3. Sathishkumar V E, Myeongbae Lee, Jonghyun Lim, Yubin Kim, Changsun Shin, Jangwoo Park, Yongyun Cho, An Energy Consumption Prediction Model for Smart Factory using Data Mining Algorithms KIPS Transactions on Software and Data Engineering, Vol. 9, no. 5, pp. 153-160, 2020. Transactions on Software and Data Engineering, Vol. 9, no. 5, pp. 153-160, 2020.
4. Sathishkumar V E, Jonghyun Lim, Myeongbae Lee, Yongyun Cho, Jangwoo Park, Changsun Shin, and Yongyun Cho, Industry Energy Consumption Prediction Using Data Mining Techniques, International Journal of Energy Information and Communications, Vol. 11, no. 1, pp. 7-14, 2020.
5. Zhang, Y. et al. Load forecasting for iron and steel industry based on hybrid mechanism- and data-driven. Energy 328, 2025. https://doi.org/10.1016/j.energy.2025.136592
6. Tang, L. et al. Synergistic air pollution and CO_2 emission reduction in China's iron and steel industry based on real-time monitoring data. Environmental Pollutiuon 385, 2025. https://doi.org/10.1016/j.envpol.2025.127053