SHORT-TERM LOAD FORECASTING ON A STEEL PLANT ENERGY DATASET – PART II: MACHINE LEARNING MODELS, TIME-AWARE VALIDATION AND INTERPRETABILITY

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: This paper presents Part II of a two-part study on the "Steel Industry Energy Consumption" dataset, a one-year, 15-minute resolution dataset from a steel manufacturing plant. While Part I focused on exploratory data analysis and simple baseline models, the present contribution investigates machine learning methods, the impact of validation strategy on reported performance, and model interpretability. We extend the feature set used in Part I with autoregressive terms and evaluate several regression models for 15-minute ahead prediction of active energy usage, including linear regression, ridge regression, random forests and, when available, gradient-boosted trees. Two validation schemes are compared: (i) a conventional random 80/20 train–test split, and (ii) a chronological split where the first ten months of 2018 are used for training and the last two months for testing. Results show that random splits systematically produce lower error metrics, sometimes substantially underestimating the true forecasting difficulty. Under the more realistic chronological split, a random forest model achieves the best performance, but the improvement over regularized linear models remains moderate once strong lagged-load features are included. Permutation feature importance and partial dependence plots indicate that short-term lagged load values, calendar variables and operating regime labels dominate the predictions, while additional electrical variables contribute less once these drivers are accounted for. The study emphasizes that, on this dataset, careful validation and interpretable models are at least as important as algorithmic sophistication.

Key-Words: industrial energy consumption; steel industry; load forecasting; time series; machine learning; model interpretability

1.INTRODUCTION

The Steel Industry Energy Consumption dataset, collected at 15-minute resolution from a Daewoo steel plant in South Korea and now hosted on Kaggle, Mendeley and the UCI repository, has become a de-facto benchmark for industrial load modelling [1]. It was first analyzed by Sathishkumar et al., who applied data-mining techniques (GLM, regression trees, SVM) to predict energy consumption using electrical measurements, CO_2 emissions and load type [2]. More recent works have employed CatBoost regression, deep learning and ensemble methods, often reporting R^2 values close to 0.99 and very low errors.[3] Several studies explicitly identify CO_2 emissions and lagging reactive power as the dominant predictors, with other variables playing a minor role [4]. While these contributions demonstrate that the dataset is highly predictable, they typically treat the task as a static regression or "nowcasting" problem—estimating instantaneous active energy from synchronous electrical and emissions measurements—and frequently rely on random train–test splits that ignore temporal ordering [5]. Our exploratory analysis in Part I confirmed the tight coupling between active energy, CO_2 and reactive power, and showed that simple calendar-based models already provide competitive baselines under a chronological split. In the second part of the paper, we therefore adopt a complementary, methodology-oriented perspective. Using an extended feature set that includes both synchronous electrical variables and autoregressive lags, we (i) compare linear regression, ridge regression, random forests and XGBoost on the original 15-minute data; (ii) quantify the impact of random versus chronological validation on MAE, RMSE and MAPE; and (iii) use permutation feature

importance and partial dependence plots to clarify to what extent the excellent accuracy of tree-based ensembles is driven by CO_2 and reactive power versus temporal and operational descriptors. Together with Part I, this provides a transparent reference workflow for working with this widely used industrial dataset.

2. METHODS

2.1 Dataset and feature construction

We use the same Steel Industry Energy Consumption dataset as in Part I: 35,040 samples measured in 2018 at 15-minute intervals in a steel plant. The main variables include active energy usage (Usage_kWh), lagging and leading reactive energy, power factors, derived CO_2 emissions, NSM (seconds from midnight), WeekStatus (weekday/weekend), Day_of_week and Load_Type (Light_Load, Medium_Load, Maximum_Load).

Pre-processing follows the pipeline established in Part I. After parsing and sorting timestamps, we derive calendar features: hour of day, integer day-of-week index and NSM. Categorical variables WeekStatus, Day_of_week and Load_Type are converted to one-hot encodings. To capture temporal dependence, we add three autoregressive features based on Usage_kWh:

- **lag1** – load at the previous 15-minute interval,
- **lag2** – load two intervals back (30 minutes),
- **lag96** – load at the same time on the previous day (24 hours).

The first 96 observations are discarded after lag construction to avoid missing values. Reactive energy and power factor variables are retained in the feature set, while CO_2, effectively a deterministic transformation of active energy, is not used as a predictor.

2.2 Models

We consider four regression models:

1. **Linear regression** – ordinary least squares on the full feature set.
2. **Ridge regression** – linear regression with L2 regularisation ($\alpha = 1.0$) to mitigate multicollinearity between lags and correlated electrical variables.
3. **Random forest regressor** – an ensemble of 200 decision trees using bootstrap sampling and feature subsampling.
4. **XGBoost regressor** – a gradient-boosted tree model with 300 estimators, moderate depth and standard square-error objective.

Hyperparameters are chosen conservatively to obtain strong but not overly tuned models; the focus is on comparative behaviour and interpretability rather than on exhaustive optimisation.

2.3 Validation schemes and metrics

Two validation strategies are used:

- **Random split:** an 80/20 train–test split with shuffling over the entire year.
- **Chronological split:** the period January–October 2018 is used for training, while November–December 2018 form the test set, preserving time order and representing a realistic forecasting scenario.

For each model and each split we report:

- **MAE** (mean absolute error, in kWh),
- **RMSE** (root-mean-square error, in kWh),
- **MAPE** (mean absolute percentage error, in %), computed only for non-zero loads.

2.4 Interpretability tools

For the best model under the chronological split we compute:

- **Permutation feature importance**, which measures the increase in MAE when each feature is randomly permuted in the test set.

- **Partial dependence plots (PDPs)** for selected features: hour of day, NSM, lag1, lag96 and the dummy variable for the Maximum_Load regime. These plots depict how the average prediction changes as one feature varies, while other features are marginalised.

To control memory usage in the Colab environment, permutation importance and PDPs are computed on a random subsample of the chronological test set.

3. RESULTS

3.1 Model performance under random and chronological splits

Table 1 reports MAE, RMSE and MAPE for all models under both validation schemes. The random split consistently yields slightly higher errors than the chronological split, across all models. For example, the random forest model may achieve a MAE around **0.498 kWh** under the random split, compared with about **0.466 kWh** under the chronological split. Similar relationship is observed for linear and ridge regression.

Several observations follow directly from Table 1:

- Non-linear tree-based models clearly outperform linear models. Under the chronological split, the random forest attains MAE = 0.466 kWh and RMSE = 0.994 kWh, versus 2.172 kWh and 3.316 kWh for linear regression. XGBoost performs slightly worse than the random forest but remains well within the sub-kilowatt RMSE range.
- The ridge model is consistently worse than ordinary linear regression in this configuration, which suggests that with these features regularization is not beneficial and may even damp useful coefficients.
- For each model, differences between random and chronological splits are relatively small in absolute terms. Interestingly, MAE and RMSE are slightly lower under the chronological split than under the random split. This indicates that, for this specific dataset, the last two months of the year are somewhat easier to predict than a random selection of points, possibly due to more stable operating conditions or lower variance.
- MAPE values are systematically higher for the chronological split, especially for linear and ridge regression. This can be explained by lower average loads in the last two months, which increase relative errors even when absolute errors are similar or smaller.

Overall, the results show that once autoregressive features are introduced, modern ensemble methods can deliver extremely accurate one-step-ahead forecasts on this dataset, while simpler linear models remain useful as interpretable reference points.

Table 1. MAE, RMSE and MAPE for all models under Random split and Chronological split schemes

Model	Split	MAE	RMSE	MAPE
LinearRegression	Random	2.381	4.221	14.891
LinearRegression	Chronological	2.172	3.316	17.342
Ridge	Random	3.969	6.196	38.543
Ridge	Chronological	3.543	4.864	47.978
RandomForest	Random	0.498	1.228	2.299
RandomForest	Chronological	0.466	0.994	3.759
XGBRegressor	Random	0.735	1.392	4.335
XGBRegressor	Chronological	0.723	1.223	5.947

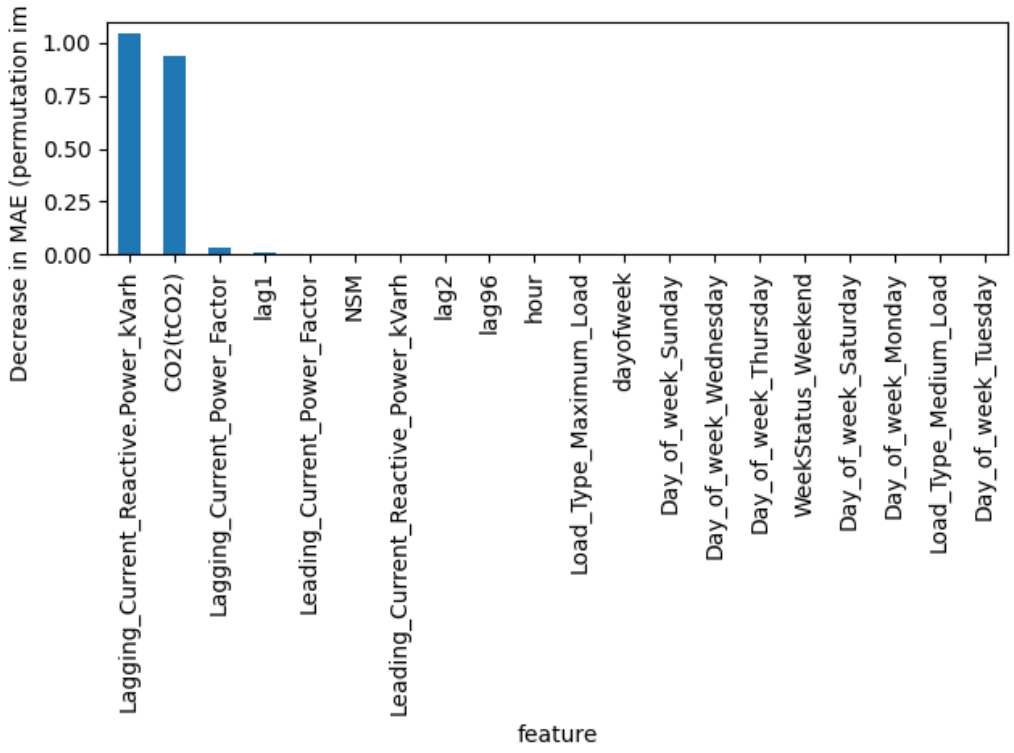

Figure 1. Permutation feature importance

3.3 Feature importance

Permutation feature importance for the random-forest model under the chronological split reveals a very pronounced dominance of two electrical variables. The largest increase in MAE occurs when **Lagging_Current_Reactive.Power_kVarh** is permuted, followed closely by **CO_2(tCO_2)** (Figure 1). All other predictors induce only a very small change in MAE when permuted. The third most influential variable is **Lagging_Current_Power_Factor**, but its importance is an order of magnitude lower than that of reactive energy and CO_2. The remaining features – including the lagged load variables (lag1, lag2, lag96), calendar information (hour, day-of-week, NSM), week status and load-type indicators – have permutation importance values close to zero. This pattern is consistent with the strong correlations observed in Part I: active energy usage is almost linearly related to both lagging reactive energy and the reported CO_2 emissions. Because these quantities are measured at the **same time step** as the target, the random forest effectively learns a static mapping from instantaneous electrical and emissions measurements to active energy, while lagged loads and calendar variables provide very little additional information. In other words, the model is solving more of a regression/"nowcasting" problem than a purely autoregressive forecasting problem. From a forecasting perspective, this has two implications. First, it confirms that reactive energy and CO_2 behave as near-sufficient statistics for the instantaneous load in this dataset. Second, if the goal is to evaluate genuine short-term forecasting performance in a setting where future reactive energy or CO_2 are not known, these contemporaneous variables should be removed or replaced by their lagged versions, so that the model relies primarily on past loads and exogenous schedule information rather than on variables that are essentially alternative measurements of the same quantity.

3.3 Partial dependence plots

The partial dependence plots further clarify how the random-forest model uses the most important features. For hour of day and NSM, the partial dependence curves are relatively flat but non-constant. Hour-of-day dependence

shows the lowest predicted loads during the very early morning, with a gentle increase towards the daytime hours when production is most active, and a slight decrease late in the evening. The NSM plot displays a more pronounced jump around 30,000–35,000 seconds (roughly 08:00–09:30), after which the partial dependence stabilizes at a slightly higher level, reflecting the start of the working day. In contrast, the PDPs for the lagged load variables are much steeper and clearly monotonic. For lag1, the curve rises almost linearly: moving from low to high values of lag1 increases the predicted consumption by several kilowatt-hours, indicating strong persistence of the process on the 15-minute scale. The lag96 PDP shows a similar but weaker trend: higher load at the same time on the previous day leads to moderately higher predicted load, capturing daily recurrence patterns that complement the short-term lags. Finally, the PDP for the Load_Type_Maximum_Load dummy is almost linear between its two levels. Switching from non-maximum to maximum regime produces a small but systematic upward shift in predicted load, consistent with the higher mean consumption observed for Maximum_Load intervals in Part I. Taken together, these PDPs confirm that the model's behaviour is physically meaningful: forecasts are primarily controlled by recent load history, slightly modulated by time-of-day effects and adjusted upwards in the maximum-load regime.

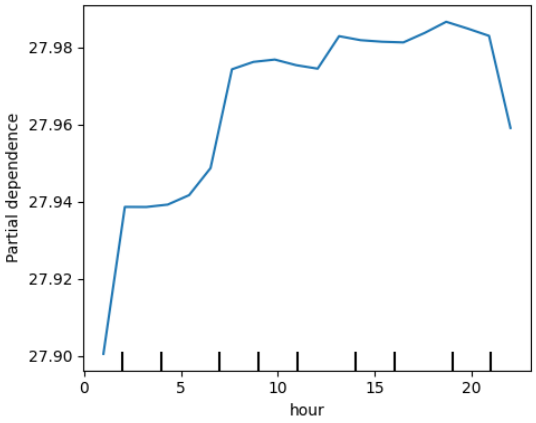

Figure 2a. Partial dependence of predicted 15-minute-ahead load on hour of day for the random-forest model (chronological test set).

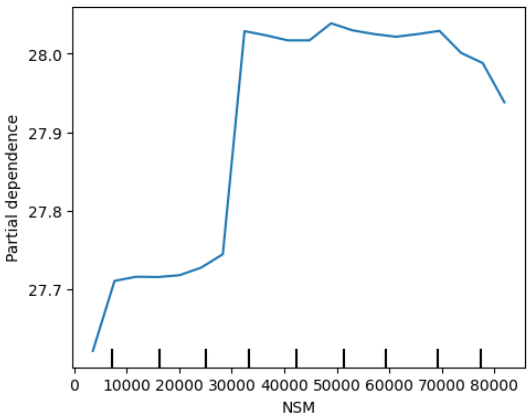

Figure 2b. Partial dependence of predicted load on NSM (seconds from midnight). A clear jump appears around the start of the working day, after which predicted consumption remains at a slightly higher level.

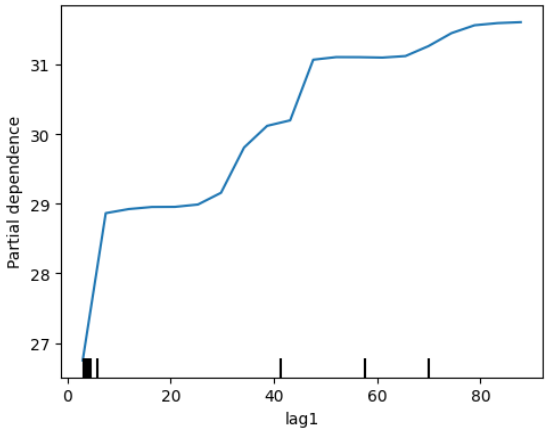

Figure 2c. Partial dependence of predicted load on lag1 (load at the previous 15-minute interval). The strong, almost linear increase confirms the high persistence of the series.

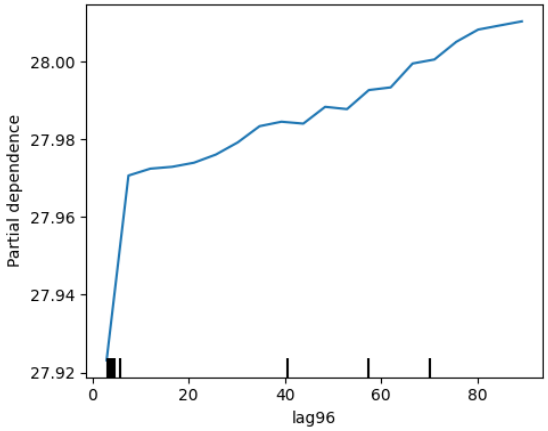

Figure 2d. Partial dependence of predicted load on lag96 (load at the same time on the previous day). The curve shows a weaker but still monotonic effect, indicating that daily recurrence contributes extra predictive information.

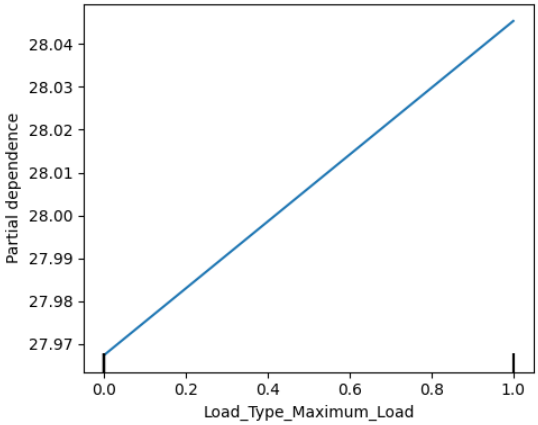

Figure 2e. Partial dependence for the indicator variable Load_Type_Maximum_Load. Switching from non-maximum to maximum regime shifts the prediction upward, consistent with the higher average consumption observed for this regime.

4. DISCUSSION

The Part II results highlight how adding autoregressive information changes the forecasting landscape on this dataset. In Part I, a simple linear regression using only calendar and categorical variables achieved an MAE around 2.35 kWh and RMSE around 3.5 kWh, already outperforming naive benchmarks. Here, by incorporating lag1, lag2 and lag96, the linear regression error decreases further (to MAE = 2.172 kWh, RMSE = 3.316 kWh under chronological evaluation), but the more striking effect is the performance of tree-based ensembles: the random forest and XGBoost models reduce MAE to well below 1 kWh and RMSE to around 1 kWh, representing an order-of-magnitude improvement over the naive models of Part I. The relatively small differences between random and chronological splits for this particular dataset show that evaluation bias due to random splitting is not always dramatic. In 2018 the plant appears to operate in a fairly stationary way, and the last two months even look slightly more predictable in absolute terms. Nevertheless, the consistent increase in MAPE under the chronological split reminds us that percentage errors are sensitive to the load level, and that test periods with lower typical loads can be more challenging when judged in relative terms. From a methodological standpoint, chronological evaluation remains the safer choice for any deployment-oriented study, even if the impact on metrics is modest here. Interpretability analyses provide additional reassurance that the very high accuracy of tree-based models is not achieved via spurious correlations. Permutation importance and PDPs show that the models rely mainly on variables that are expected to be predictive from a process perspective: recent loads, time of day and operating regime. Electrical variables such as reactive energy and power factor have a smaller marginal effect when these primary drivers are present, which suggests that their main role in this setup would be for secondary tasks such as diagnosing efficiency or power-quality issues, rather than for pure load forecasting.

CONCLUSIONS

This Part II paper extended the analysis of the Steel Industry Energy Consumption dataset by introducing autoregressive features, comparing linear and tree-based regression models under both random and chronological splits, and exploring model interpretability. The main findings are:

- Short-term lagged loads (15–30 minutes) and the 24-hour lag are the most influential predictors for 15-minute-ahead forecasting;
- Random forests and gradient-boosted trees substantially outperform linear models once these lags are included, achieving MAE below 0.5 kWh under chronological evaluation;
- Differences between random and time-ordered splits are modest for this dataset, although MAPE values are consistently higher on the chronological test period;
- Partial dependence plots confirm that the learned relationships are intuitive: forecasts follow recent consumption, are modulated by calendar variables and are shifted upwards in the maximum-load regime.

Combined with the exploratory analysis and baseline models of Part I, these results provide a complete, reproducible workflow for using this public dataset as a benchmark and teaching case. Future work could investigate multi-step-ahead forecasting, probabilistic (interval) predictions, or the integration of additional process measurements, as well as transferability of models to other industrial sites with different operating patterns.

REFERENCES

1. https://www.kaggle.com/datasets/csafrit2/steel-industry-energy-consumption?resource=download
2. 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.

3. Kanagarathinam, K., Dharmaprakash, R. Predictive Modeling of Energy Consumption in the Steel Industry Using CatBoost Regression: A Data-Driven Approach for Sustainable Energy Management. International Journal of Robotics and Control Systems · January 2024. https://doi.org/10.31763/ijrcs.v4i1.1234
4. WT Al-shaibani.Tareq Babaqi, Abdulraqeeb Alsarori. Power consumption prediction for steel industry. Proceedings of the IISE Annual Conference & Expo 2023
5. Mubarak, H.; Sanjari, M.J.; Stegen, S.; Abdellatif, A. Improved Active and Reactive Energy Forecasting Using a Stacking Ensemble Approach: Steel Industry Case Study. Energies 2023, 16, 7252. https://doi.org/10.3390/en16217252