sciendo

Time Series Forecasting with LightGBM under Data Scarcity: An Application to Romania's Inland Gas Consumption

Robert-Stefan CONSTANTIN*
Bucharest University of Economic Studies, Bucharest, Romania
**Corresponding author, constantinrobert21@stud.ase.ro*

Adriana AnaMaria DAVIDESCU
Bucharest University of Economic Studies, Bucharest, Romania
adriana.alexandru@csie.ase.ro

Eduard Mihai MANTA
Bucharest University of Economic Studies, Bucharest, Romania
eduard.manta@csie.ase.ro

Abstract: *Developing forecasting models capable of learning from small datasets is increasingly valuable for scenarios with limited computational resources and tight time constraints. In this case study, we processed monthly data on Romania's inland gas consumption—a primary benchmark reflecting the country's industrial growth, level of technological advancement, and reliance on non-renewable energy sources. This research tests the extent to which LightGBM, a gradient boosting framework, can predict seasonal patterns in monthly gas consumption. To aid the machine learning framework in better understanding the series pattern, we applied a first-order differencing to the data. By combining hyperparameter tuning, cross-validation, and tailored feature engineering (including lagged variables and rolling-window statistics), the analysis thoroughly evaluates LightGBM's performance under data-scarce conditions. Model accuracy was assessed using Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE), demonstrating the extent of LightGBM's predictive capacity across multiple horizons despite constrained data settings. These findings offer insights into the feasibility of employing fast-adapting, lightweight machine learning techniques for reduced time series datasets, while minimizing both computational effort and processing time.*

Keywords: gradient boosting framework, energy consumption, LightGBM, machine learning, forecasting.

Introduction

Forecasting energy consumption in rapidly evolving industries has long been a focal point of research, particularly as resources and computational capabilities may be constrained. Robust forecasting models are essential for effective resource allocation, policy-making, and long-term strategic planning. However, the significant challenge of data scarcity remains: many energy-related datasets, especially those collected monthly or quarterly, often provide only a limited number of observations. This limitation can reduce the efficacy of models that rely on large-scale data or extensive series transformations.

Another critical issue we tackle is the reliance on conventional time-series analysis approaches, such as strictly identifying additive or multiplicative seasonality patterns. By removing the necessity to predetermine these relationships or maintain a strict threshold of variables, we mitigate the risk of misclassifying seasonality types and avoid overfitting models to assumed patterns. This approach allows for a more flexible, data-driven extraction of underlying trends and dependencies even under stringent data constraints. In other words, this study explores a

DOI: 10.2478/picbe-2025-0118

streamlined forecasting framework that involves minimal data transformation, employing only a one-time first-order differentiation to assist the machine learning process in understanding the series pattern. By keeping prerequisites to a minimum, we not only reduce the modeling overhead but also allow the algorithm to uncover underlying temporal patterns directly from the data, enabling a general model to study an extensive range of time series.

This study addresses the challenge of predicting monthly inland gas consumption in Romania, an economy where gas usage is a key indicator of industrial progress and reliance on non-renewable resources. Given the data-scarce context, we employ LightGBM due to its inherent advantages. Specifically, LightGBM's leaf-wise tree growth strategy prioritizes splits that maximize loss reduction, leading to enhanced predictive accuracy while mitigating overfitting—a critical advantage with limited data. Furthermore, its feature-bundling and gradient-based one-side sampling techniques improve computational efficiency and further reduce overfitting risks. To isolate and emphasize this algorithm capacity to learn directly from time series data, we intentionally avoid data transformations and exogenous variables. Our primary objective is to determine LightGBM's effectiveness under minimal data conditions. Methodologically, we will constrain hyperparameter optimization to small steps and minimal values (e.g., leaf data, residual bins) to ensure effective learning from limited information. Finally, to further prevent overfitting and ensure feature relevance, we will adopt a higher acceptable error threshold.

The methodology includes a systematic feature engineering strategy, incorporating lagged variables and rolling-window statistics (mean and standard deviation) at different intervals (at 3, 6 and 12 months, allowing the model the understand immediate and long-term effects). This design aims to capture patterns in gas consumption while maintaining a parsimonious and computationally efficient approach. To ensure repeatable and unbiased assessments, hyperparameter tuning and cross-validation are employed, providing a multifaceted view of the model's reliability and accuracy. Model performance is evaluated using Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE), metrics that underscore the feasibility of LightGBM's predictions on the testing horizon.

By illuminating how a lightweight model architecture can operate effectively under data-scarce conditions, this research contributes to ongoing discussions on sustainable energy planning and resource optimization. The findings have broader relevance for industries and regions facing similar constraints, offering a practical blueprint for leveraging machine learning techniques where computational and data resources are limited.

In the following paper, our analysis pipeline will be structured as follows: first, we establish the seasonal characteristics of our data using autocorrelation function (ACF), partial autocorrelation function (PACF), and seasonality plots. Following this initial analysis, we train a LightGBM regression model. In this first iteration, we focus on recursive forecasting, back-testing, and cross-validation with respect to the time index, without applying specific hyperparameter tuning. Specifically, we use the training segment to predict three-month intervals within the validation set, and subsequently evaluate the model's performance on the test split.

Next, we optimize the model's hyperparameters using a Bayesian approach, conducting 1000 iterations to identify the best configuration based on validation set performance. We then re-evaluate the model's predictive accuracy on the test split.

With the optimized hyperparameters, we proceed to feature selection, aiming to identify and eliminate redundant features. After establishing our final model, we analyse the impact of our engineered features using SHAP values, as well as feature importance values. Finally, we generate an out-of-sample forecast for 36 steps and a bootstrapped prediction encompassing 200 scenarios.

We generate the bootstrapped forecasts by resampling the residuals of the LightGBM predictions 200 times, and adding those resampled residuals to the original LightGBM forecasts. We then assess the bootstrap estimations using a ridge plot over the specified steps.

The paper is structured as follows: the next section is dedicated to a literature review related to natural gas consumption and forecasting methods. Next section is dedicated to the data and methodology in which the datasets are presented, and methodology aspects are provided. Section 4 is dedicated to the empirical results and discussions followed by the conclusions of the study.

Literature review

Romania's natural gas consumption has long been a focal point for gauging the country's dependency to non-renewable sources, technological advancement, and energy security. Beyond fueling industries and households, inland gas usage signals deeper structural shifts, from modernizing extraction sites to upgrading commercial and residential infrastructures (Neagu et al., 2015). Improved natural gas resources can reduce heating and power costs, stimulate both local and foreign investment in industrial ventures, and mitigate emissions, all while indicating a nation's capacity to modernize its technological framework—especially within energy-intensive sectors. From an industrial perspective, affordable and stable gas supplies exert significant multiplier effects by strengthening manufacturing competitiveness, fostering new industrial ventures, and attracting foreign investors.

Against this backdrop, understanding a country's energy consumption becomes pivotal, serving as a key indicator of both industrial development and technological progress (Odularu & Okonkwo, 2009). In Romania, inland gas consumption reflects technological advancement and underscores the nation's reliance on non-renewable energy sources: elevated consumption typically accompanies a more robust industrial sector, demanding greater energy inputs. An upper-middle-income EU member state, Romania's economy has displayed robust growth since 2000, accelerated further by its 2007 EU integration (Magda, Bozsik, & Meyer, 2019). Ranked as the 54th largest energy consumer worldwide and the second-largest natural gas producer in the European region (second only to the Netherlands), Romania demonstrates the complex interplay between its industrial aspirations and the need to balance economic development with sustainability goals. To effectively analyse and forecast these patterns of gas consumption we turn to a robust machine learning framework: LightGBM.

LightGBM, short for Light Gradient Boosting Machine, is a popular gradient boosting framework known for its efficiency and accuracy. While it often shines with large datasets, questions arise about its effectiveness when dealing with limited data or the complexities of time series data exhibiting seasonality. The effectiveness of machine learning models often depends on the availability of abundant data. However, in many real-world applications, obtaining large datasets can be challenging or expensive.

A core challenge with smaller datasets is accurately estimating information gain for optimal split points in decision trees, a fundamental aspect of LightGBM's underlying Gradient Boosting Decision Tree (GBDT) algorithm. Traditional GBDT algorithms can be computationally expensive when dealing with limited data, as they need to scan all data instances to estimate information gain for all possible split points (Ke, et al., 2017). This can be significantly slower than LightGBM, which can speed up the training process of conventional GBDT by up to over 20 times while achieving almost the same accuracy.

LightGBM tackles this challenge through its innovative Gradient-based One-Side Sampling (GOSS) technique 1. GOSS focuses on data instances with larger gradients, which play a more

significant role in calculating information gain. By prioritizing these instances, GOSS achieves a more accurate gain estimation than uniform random sampling, with the same target sampling rate (Ke, et al., 2017). This leads to improved efficiency without significantly sacrificing accuracy, especially when the value of information gain has a large range. While this algorithm generally excels with larger datasets, studies have shown its potential with smaller datasets as well. For instance, in a study on dementia detection using a relatively small dataset of 1000 patient records, LightGBM achieved the highest accuracy (98%) among six machine learning classifiers (Jahan et al., 2024).

Natively LightGBM is a framework designed for classification and regression tasks, but it can be adapted for time series forecast by incorporating special feature engineering and temporal dependencies. The first step is transforming the time series data into a supervised learning problem. This involves using past values of the time series (lags) as predictors for future values (Amat Rodrigo & Escobar Ortiz, 2023). For example, to predict the value at time t, we can use the values at times t-1, t-2, and so on, as features. This allows LightGBM to learn the temporal dependencies and patterns in the data. Another way to enhance intrinsic prediction accuracy is to introduce rolling statistics, this way we can observe proximal effects or long-term dependencies (Hyndman & Athanasopoulos, 2021). Because of this model ability to deliver great performing models at a low computation cost its popularity has risen fast, his applications extending from medical field to stock (Hartanto et al., 2023) and sales prediction (LightGBM, 2023).

Methodology

This study analyzes Romania's inland gas consumption with the objective of forecasting future consumption using time series modeling techniques. The data used in this analysis was extracted from the Eurostat dataset titled “Supply, transformation and consumption of gas - monthly data” (dataset code: nrg_cb_gasm). This dataset provides monthly observations of inland gas consumption, measured in million cubic meters (MCM). Inland consumption is defined as the total gas consumed within Romania, regardless of whether it was produced locally or imported. The dataset covers a period of 133 months, extending from January 2014 to January 2025.

The raw data was initially inspected for missing values, and none were found, this also applies in the case of possible outliers, no significant outlier which could alter the predictions was found. Regarding data alteration, a first order differencing was applied, as it helped our ML framework better understand the series pattern. Benefits of applying differentiation are also presented in the study of Schmid on benchmarking machine learning against statistical forecasting methods (2025).

The dataset was subsequently divided into training, validation, and test sets, strictly maintaining the temporal order essential for time series analysis. Training set consists of 72 observations spanning from January 2014 to December 2019. Validation set includes 33 observations covering the period from January 2020 to August 2022 and was used for hyperparameter tuning of the forecasting models. The test set comprises 30 observations from September 2022 to January 2025 and was reserved for the final evaluation of model performance and back-testing. Percentage wise training set consisted of 54.13% of the total data, while validation and test accounted to 24.81% and 22.55%. Those specific percentages also account to 6 years of monthly data (72 months) in the training set and almost 2 and a half (33 and 30 months) years for each one of validation and testing. The training set comprises only about half of the total data due to the limited dataset, a common approach when working with such data constraints.

However, the training sample will have an evolving size, as we will apply rolling forecast windows of three steps each and update the training set with the actual values after each forecast.

The entire data analysis workflow, encompassing data retrieval, model development, and visualization, was carried out using Python Version 3.12. Several key libraries were employed: scikit-learn for general machine learning tasks, lightgbm for applying our gradient boosting based model, pandas and numpy for data manipulation, skforecast specifically for time series forecasting and plotly for creating interactive visualizations, statsmodels. Model performance was assessed using three metrics: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE). MAE measures the average magnitude of the forecast errors, expressed in million cubic meters, the same units as the original data. While MAPE expresses the error as a percentage, RMSE brings more contrast on penalisation of errors with a higher magnitude.

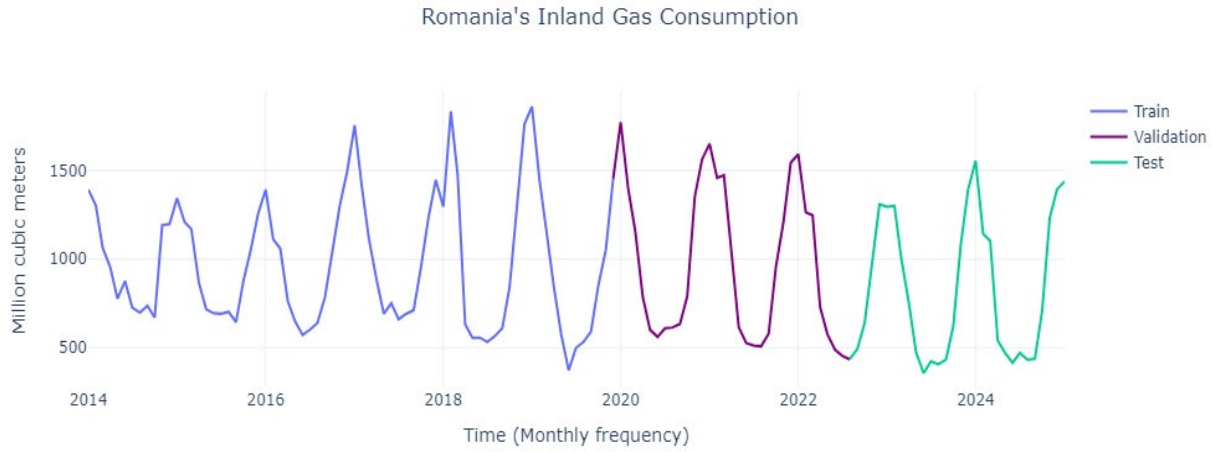

Figure 1. Splits for train test validation data sets

Source: Author's own creation.

The methodology employed in this study centers around the implementation of the LightGBM framework and SHAP (Shapley Additive Explanations). LightGBM is a gradient boosting decision tree algorithm, notable for its computational efficiency and robustness, especially under limited data conditions. Mathematically, LightGBM constructs additive models in a forward stage-wise fashion. Given a training dataset, LightGBM predicts the output $\hat{y}_i$ by minimizing a predefined loss function L:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$$

where $h_m(x)$ denotes the decision trees, γ_m the weight of each tree and M indicates the total number of trees. Each iteration involves minimizing a loss function such as MAE or MSE defined as follows:

$$\gamma_m = argmin_\gamma \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + \gamma h_m(x_i))$$

where N is the number of data points and L represents the loss function. Additionally, LightGBM leverages Gradient-based One-Side Sampling (GOSS) to manage computational demands by

preferentially sampling data points with higher gradients, thus effectively handling computational and data limitations.

SHapley Additive explanations (SHAP) (Ekanayake et al., 2022) provide an advanced mathematical framework to quantify feature contributions to model predictions based on cooperative game theory. SHAP assigns each feature an importance value derived from Shapley values, calculated as:

$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [f(S \cup j) - f(S)]$$

Where, ϕ_j represents the contribution of feature j, S is a subset of features excluding feature j and $f(S)$ denotes the expected prediction from the subset **S**.

In the end an implementation of Recursive Feature Elimination with Cross-Validation (RFECV), an iterative approach for feature selection aimed at identifying the optimal subset of features by minimizing predictive error is employed. This procedure mathematically evaluates each subset (X_k) through:

$$J(X_k) = \frac{1}{K} \sum_{i=1}^{K} CV_{score}(X - X_k)$$

where: $J(X_k)$ is the average cross-validation error across K iterations.

This mathematical foundation ensures rigorous model development, interpretability, and robustness, specifically targeting the challenges of forecasting Romania's inland gas consumption under conditions of data scarcity.

Also to ensure the model rightly captures effects of seasonality, Fourier transformations were applied to monthly values for both sine and cosine models and introduced as exogenous variable alongside with dummy variables representing crisis events, specifically COVID-19, European energy crisis and Ukraine-Russia conflict. Adding these trigonometric functions as exogenous variables will aid the model in better understanding the seasonality.

Results and discussions

This section focuses on the main findings of the LightGBM model. First, to understand the temporal dynamics of gas consumption, we assessed the time series for inherent seasonality. A monthly seasonality plot presented in Figure 2, constructed by overlaying yearly consumption patterns, effectively visualizes a pronounced and consistent annual cycle. This cyclicality, characterized by high peaks in winter seasons and lows in summer indicates a strong and predictable seasonal influence on gas consumption, as this variable is mainly influenced by temperature and shows the dependency on gas consumption. This preliminary analysis suggests that any effective forecasting model must account for this significant seasonal component to accurately capture the underlying consumption patterns.

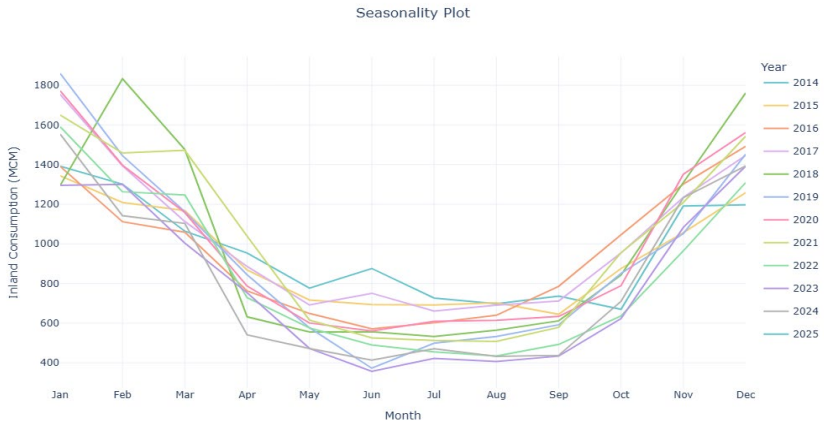

Figure 2. Seasonality plot for monthly inland gas consumption data

Source: Author's own creation.

Inland gas consumption consistently peaks in December and January, with maximum values generally ranging from around 1400 to 1800 MCM. Conversely, consumption reaches its lowest point during the summer months, typically between June and August, with values fluctuating between approximately 400 and 800 MCM.

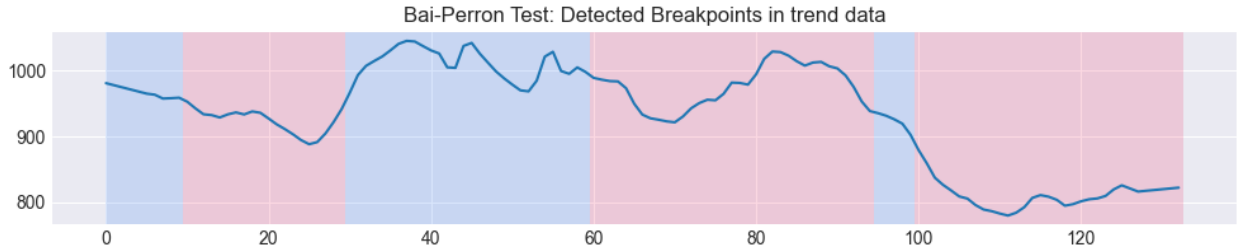

Figure 3. Bai-Perron test visualisation on trend data

Source: Author's own creation.

After decomposing the seasonality of our data to visualize the underlying trend, which exhibited periods of stability, growth, and decline, a Bai-Perron analysis was conducted. This analysis identified clear structural breaks in the trend at approximately time points 10-30 (a period of decline), 60-95 (a more volatile period), and 100-133 (a period of lower values), indicating several regime changes over the observed period. Notably, the trend visualization also reveals a significant reduction in gas consumption in the later years of the observed period. This decrease particularly evident after the last structural break around point 100, could indicate a transition toward a more electricity-focused industry in Romania. Separately, a Zivot-Andrews unit root test also applied on trend series, which allows for a single endogenous structural break, suggested a potential break point around the 13th observation. However, the test results indicated that we could not reject the null hypothesis of a unit root (P-value of 0.5169), suggesting the series remains non-stationary even when accounting for this potential break.

The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots (Figure 4) also provide valuable insights into the temporal dependencies within the gas consumption data. The ACF displays a pronounced sinusoidal pattern with slowly decaying oscillations. The peaks in the ACF occur at lags that are multiples of 12 (6, 12, 18, 24, 36, etc.),

strongly indicating a 12-month (annual) seasonality. The PACF, in contrast, shows a significant spike at the first lags, having the last out of confidence interval at lag 12.

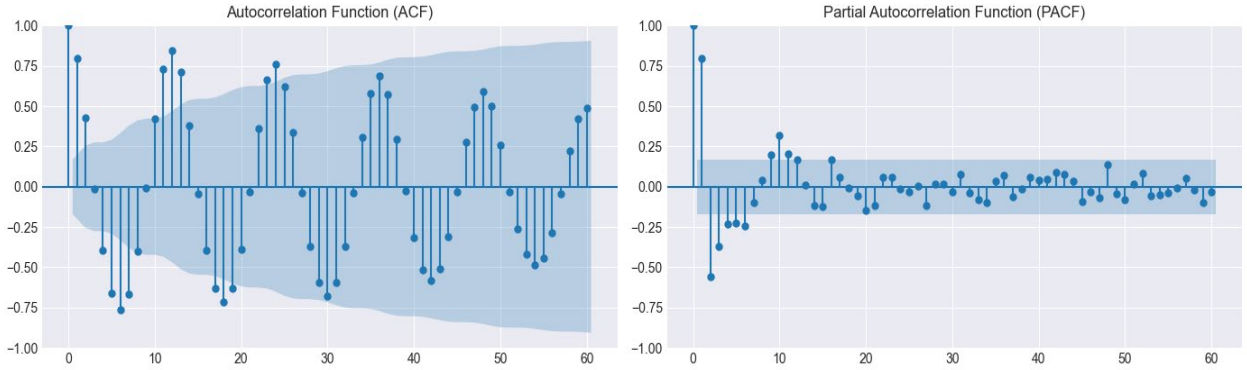

Figure 4. ACF and PACF plots of inland gas consumption

Source: Author's own creation.

As it has been previously noted that first-order differentiation tends to be advantageous for time series displaying complex patterns, a subsequent examination of the ACF and PACF is therefore conducted. Although further differentiation could theoretically address this, as in the case of seasonal differencing, we are concerned that it might lead to an excessive loss of valuable information from our data. Therefore, we have decided to proceed with training our model at this stage. Our rationale is that the processed ACF and PACF indicates that the dominant seasonal effects, clearly visible in the initial data, have been largely mitigated after the first two lags.

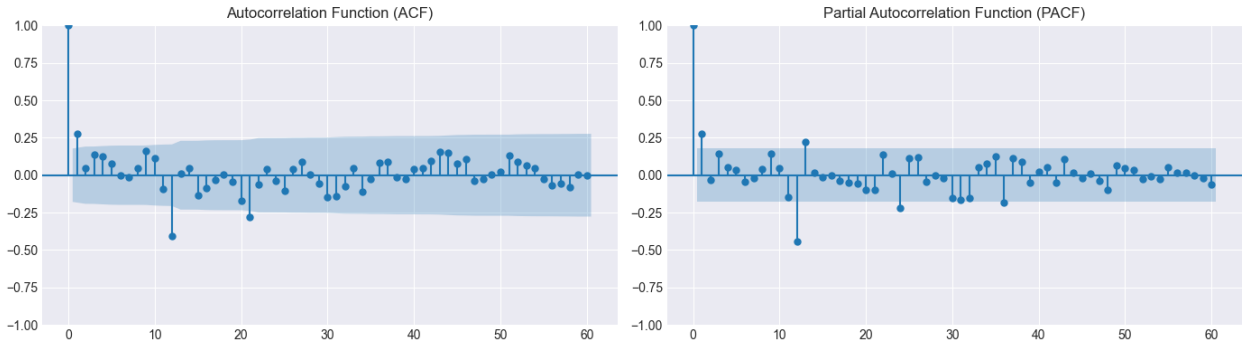

Figure 5. ACF and PACF plots of differentiated inland gas consumption

Source: Author's own creation.

After acknowledging the seasonality pattern, we deploy our first iteration of the model. To capture these patterns, we included the first 12 lags as features, along with rolling averages and standard deviations (windows of 3, 6, and 12 months) to represent evolving trends and volatility. Hyperparameters were kept with default function values on the first try, but we applied a back-testing LightGBM forecaster with a cross-validation of 3 steps each time and refitting on training set after every iteration. Thus, the first model trained against the validation test resulted in a MAE of 168, MAPE of 20% and RMSE of 0.78. We also fit this model on test, where we obtain a better result on metrics error, showing how much relevance brings more information to the model, respectively a MAE of 146, MAPE equal to 17.7% and RMSE of 0.72. One more time we evaluate the model but this time with a back-testing process directly onto test sample, obtaining error metrics of 93 for MAE, 14.4% for MAPE and 0.44 to RMSE. Before we go onto the next step of the

pipeline, a model is evaluated again test set but this time with the addition of exogenous variables, resulting the following metrics: MAE 92.88, MAPE 13.17%, RMSE 45.10.

To see if better parameters for the model are available, we deploy hyperparameter selection phase over 1000 model estimation. Setting for back-testing and cross-validation were kept the same during the process (more details about the top 3 models estimated can be seen in Table 1). Here the best model obtained an MAE of 52, accounting for a MAPE of 15% and RMSE of 0.62. Now we can evaluate this model against the test set.

Table 1. Best performing models of inland gas consumption forecasting

Model	lags	MAE	MAPE	RMSE	Estimator	Max depth	Min data in leaf	Learning rate	Lambda
I	1,2,3,4,5	120.01	0.15	0.62	155	50	20	0.05	0.66
II	1,2,3,4,5	129.41	0.15	0.62	185	55	20	0.03	0.68
III	1,2,3	135.45	0.16	0.64	170	20	20	0.05	0.70

Source: Authors' own creation.

In the figure 6 we can see how this model performed against test observed variables, where we obtained a MAE of 91.96 (MAPE of 11.28%) and RMSE of 0.44. Upon evaluating the LightGBM predictions on the unseen test data, the refined model demonstrates another improvement in accuracy, with predictions closely aligning with the actual 'Inland Consumption' values. The refined version successfully replicates the underlying sinusoidal pattern of the seasonality within the test set. As visually confirmed in the plot, both the predicted (red) and actual (black) time series exhibit a similar cyclical rhythm, with peaks and troughs temporally aligned, indicating that the model continues to effectively capture the inherent seasonal dynamics driving the time series. Areas where the model tends to either overestimate or underestimate are the peaks of gas consumption, more specifically in the colder months.

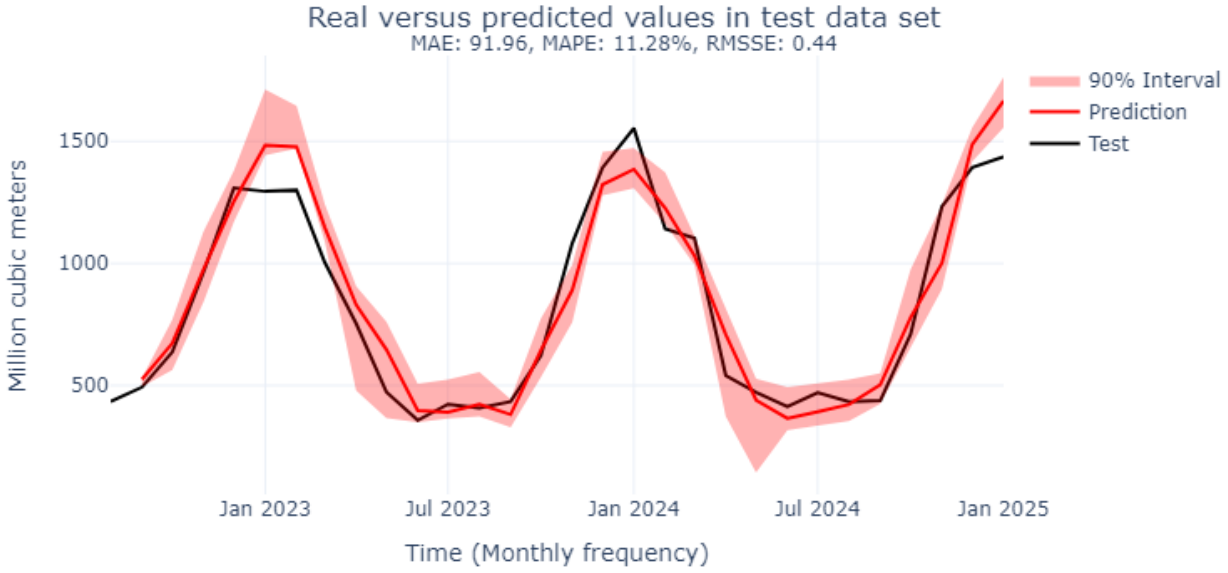

Figure 6. Real versus predicted values in test data set of inland gas consumption

Source: Author's own creation.

For feature selection, we employed recursive feature elimination with cross-validation. Specifically, we implemented 24-fold cross-validation on a 70% subset of the original dataset,

iteratively removing one feature at a time. This process identified all the lags employed (until lag 12), all 6 of the rolling features and only sine and cosine waves as valuable features, indicating that crisis events bring little to no relevance to the series.

Benchmarking this refined model against the test set resulted in a Mean Absolute Error (MAE) of 82.58, Mean Absolute Percentage Error (MAPE) of 10.13% and Root Mean Squared Error of 0.42. Although the improvement is modest, it represented progress toward a more refined model.

To understand the impact of these features on our predictions, we utilized SHAP (Shapley Additive Explanations) plot illustrated in Figure 7. SHAP values, derived from game theory, quantify the contribution of each feature to the model's output. Below, we observe that both trigonometric functions strongly impact predicted values, as well as lags 7 and 12. Furthermore, the concentration of Shapley values around zero as we descend the graph indicates decreasing feature influence. The wider the spread of a feature's Shapley values, the greater its impact on predictions. For instance, low values in lag 12 tend to strongly decrease predictions, while high values inflate them. Clustering of SHAP values vertically indicates that certain features exhibit redundancy in their predictive contribution. This is reflected in the limited influence of the 6-month rolling mean on our forecasts. Moreover, SHAP value visualization provides a diagnostic tool for potential model overfitting. For example, a broad distribution or central agglomeration of SHAP values along all features would suggest that the model is not reliably identifying feature importance, raising concerns about the robustness of future predictions.

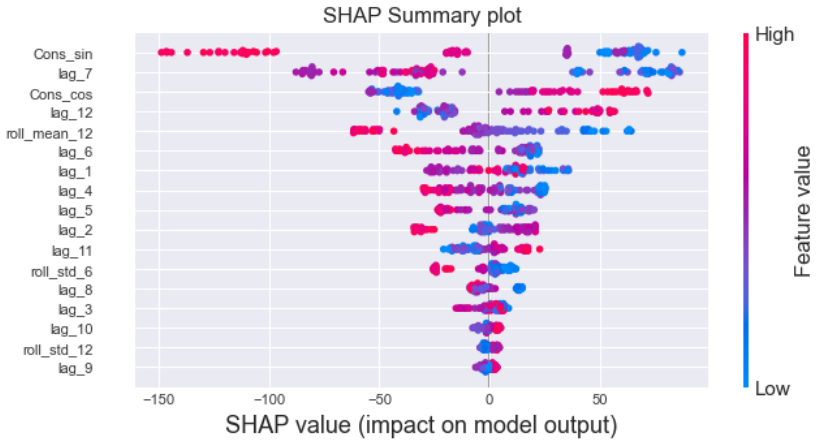

Figure 7. SHAP Summary plot of inland gas consumption

Source: Author's own creation.

Alternatively, we can examine feature importance to determine the most influential variables in node splitting for decision-making, as presented in Table 2. This analysis reveals that the 12-month rolling mean exhibited the highest importance with a value of 67, followed by the sine wave component with a value of 63, and the first lag with 55. The prominence of the 12-month rolling mean, along with both trigonometric functions as highly important features underscores the strong seasonal dependence inherent in our estimations and the annual effect of average consumption on predicted variables.

Table 2. Feature importance of inland gas consumption forecasting. Top 10 features

Feature	Importance
Roll mean 12	67
Sine wave	63
Lag 1	55
Cosine wave	49
Lag 7	44
Lag 12	34
Lag 2	33
Lag 3	33
Lag 11	33
Lag 4	32

Source: Authors' own creation.

In both cases of SHAP variables analysis and feature importance we obtain similar insights on how seasonality is the main driving factor of our predictions.

With a stable model established, we can now generate out-of-sample predictions. We have chosen to make 36 monthly predictions, spanning three years from February 2025 to February 2028 as observed in Figure 8. LightGBM effectively captured seasonality effects, predicting a clear continuation of the cyclical pattern observed in the historical data over the next three years. Our model also manages to estimate certain irregularities that may occur in predictions, and as seen it forecasts a relatively stable overall level of gas consumption throughout the prediction period, with the primary variations driven by the seasonal cycle. Given the robust performance of the model on the unseen test data (as indicated by MAE, MAPE, and RMSE), we have a reasonable degree of confidence in these out-of-sample forecasts.

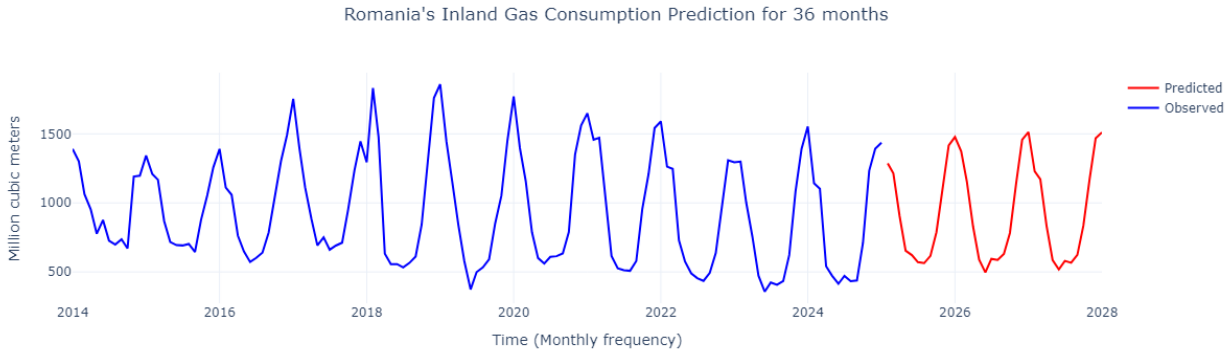

Figure 8. Out of sample forecast of inland gas consumption

Source: Author's own creation.

To enhance the robustness of these out-of-sample predictions and to quantify the associated uncertainty, a bootstrapping approach was employed. Specifically, 200 bootstrapping iterations were conducted for each of the 36 forecast steps, spanning three years from February 2025 to February 2028. This methodology provides a probabilistic perspective on the forecast by estimating the distribution of potential outcomes at each time point. As observed in Figure 9, the variability of the bootstrapped forecast distributions exhibits a continuous increase beyond the initial 12 months. This phenomenon is consistent with the inherent nature of time series forecasting, where the uncertainty surrounding predictions typically expands with an increasing forecast horizon due

to the accumulation of prediction errors and the greater potential for deviations from historical patterns.

While the bootstrap resampling inherently expands the forecast variability, the distinct seasonal pattern from the historical data is consistently preserved. This demonstrates the model's effective capture and extrapolation of monthly seasonality. Probabilistically, a higher number of bootstrap samples will lead to a more precise estimate of the forecast's convergence distribution at each time step.

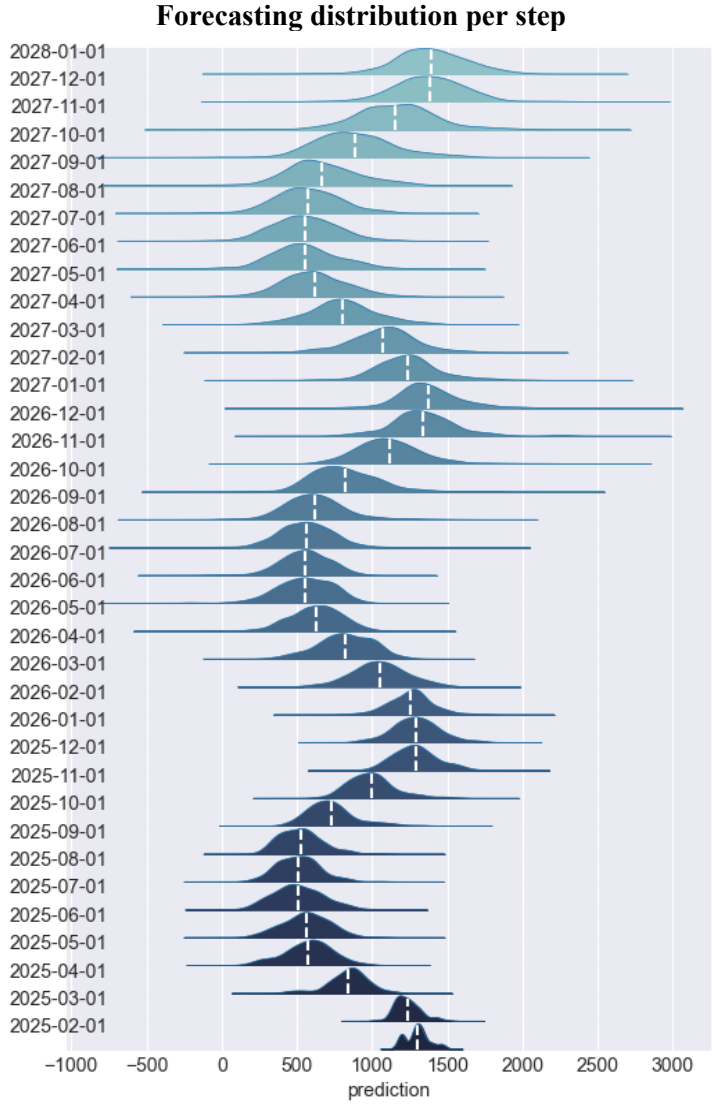

Figure 9. Ridge plot for bootstrapped out-of-sample forecast of inland gas consumption

Source: Author's own creation.

Conclusion

In this study we use machine learning models such as LightGBM to forecast monthly consumption of domestic gas with data scarcity, concentrating specifically in context of Romania. The model successfully captured complex temporal dependencies by mainly utilising engineered features like trigonometric transformations rolling windows and lags. Employing parameter optimization and validation, LightGBM managed to reflect seasonal fluctuation inside gas usage — confirming that

it can be used as a forecasting tool even under lightly staffed or untrained conditions for instance with only three staff members compared to twenty in model workstations.

The model delivered a good predictive accuracy, successfully reproducing the seasonal dynamics of the gas consumption data. The recursive feature elimination and SHAP analysis underscored the significance of each lagged variable, sine and cosine wave and rolling-windows, highlighting their importance in the predictive performance of the model. Furthermore, the methodological approach provides insights for stakeholders in energy and other sectors facing similar challenges, showing how efficient machine learning frameworks can support strategic planning and operational decision-making under data scarcity.

However, this study exhibits some limitations. Firstly, the reliance solely on univariate time series data inherently restricts the model's ability to account for external factors, such as economic fluctuations, technological advancements, and weather variability, which can potentially influence gas consumption especially with seasonal patterns. Secondly, the limited data size constrains the model's ability to generalize beyond short-term forecasting horizons. The longer-term accuracy and stability in this case might be affected. Additionally, although efforts were made to mitigate overfitting through cross-validation and hyperparameter tuning, residual risks of overfitting persist due to the small size of the dataset.

Future research directions should explore the integration of exogenous variables, such as temperature variations, economic indicators, or policy changes in other machine learning models to provide a better forecasting precision. Additionally, extending this approach to other energy markets and comparing performance across different machine learning models could offer deeper insights into the efficiency of various forecasting methodologies under conditions of data scarcity. Further investigation into ensemble or hybrid models combining LightGBM with other statistical techniques may also provide valuable insights and strategies to better mitigate the identified limitations and potentially improve the predictive power of the models and provide robusteness.

Acknowledgements

The research study has been supported by the EU's Next Generation EU instrument through the National Recovery and Resilience Plan of Romania - Pillar III-C9-I8, managed by the Ministry of Research, Innovation and Digitalization, within the project entitled „Non – Gaussian self – similar processes: Enhancing mathematical tools and financial models for capturing complex market dynamics", contract no. 760243 /28.12.2023, code CF 194/31.07.2023.

References

Amat Rodrigo, J., & Escobar Ortiz, J. (2023). *skforecast.* doi:10.5281/zenodo.8382788

Ekanayake, I., Meddage, D., & Rathnayake, U. (2022). A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Studies in Construction Materials, 16*. doi:10.1016/j.cscm.2022.e01059

Hartanto, A. D., Kholik, Y. N., & Pristyanto, Y. (2023). Stock Price Time Series Data Forecasting Using the Light Gradient Boosting Machine (LightGBM) Model. *International Journal on Informatics Visualization, 4*(7). doi:10.30630/joiv.7.4.01740

Hyndman, R., & Athanasopoulos, G. (2021). *Forecasting: principles and practice* (ed. III).

Jahan, F., Shifat, S. M., Anannya, F. Z., & Mostafa, S. &. (2024). Short Paper: Dementia Patient Health, Prescriptions ML Dataset: LightGBM Classification of XAI-based LIME and

SHAP for Dementia Detection. *International Conference on Networking, Systems, and Security*, (pg. 197-202).

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., . . . Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*(30).

LightGBM, I. S. (2023). Tong Zhou. doi:10.48550/arXiv.2305.17201

Magda, R., Bozsik, N., & Meyer, N. (2019). An Evaluation of Gross Inland Energy Consumption of Six Central European Countries. *Journal of Eastern European and Central Asian Research, 2*(6), 270-281.

Neagu, C., Bulearcă, M., Sima, C., & Mărguşa, D. (2015). A SWOT analysis of Romanian Extractive Industry and Re-Industrialization Requirements of This Industry. *Procedia Economics and Finance*(22), 287–295. doi: 10.1016/S2212-5671(15)00288-9

Odularu, G. O., & Okonkwo, C. (2009). Does energy consumption contribute to economic performance? Empirical evidence from Nigeria. *Journal of Economics and International Finance, 1*(2), 044-058.

Schmid, L., Roidl, M., Kirchheim, A., & Pauly, M. (2025). Comparing Statistical and Machine Learning Methods for Time Series Forecasting in Data-Driven Logistics—A Simulation Study. Entropy, 27(1), 25. https://doi.org/10.3390/e27010025