POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

# Benchmarking Models for Very Short-Term Forecasting of Imbalance Prices in the Italian Electricity Market

Laurea Magistrale in Energy Engineering - Ingegneria Energetica

**Author:** Valerio Rivadossi

**Advisor:** Prof. Filippo Bovera

**Co-advisor:** Andrea Scrocca

**Academic year:** 2024-2025

## 1. Introduction

To comply with the European 'Fit for 55' decarbonization targets and the carbon neutrality goals set by the European Green Deal, there is an increasingly massive integration of non-programmable renewable sources, such as wind and solar, into the Italian energy mix. This growing penetration has increased forecast uncertainty regarding generation, making grid balancing management more complex. In the Italian context, the Transmission System Operator (TSO), Terna, compensates for deviations between scheduled and actual energy production/consumption (i.e., imbalance volumes) through the European balancing platforms and the Balancing Market (BM), where the main components of the imbalance price are determined, which represents the economic cost to maintain grid equilibrium. The TIDE reform (Testo Integrato del Dispacciamento Elettrico) introduced in January 2025 has revolutionized this mechanism, transitioning from a dual pricing and hourly settlement system to a single pricing and quarter-hourly granularity system. This transformation makes imbalance price forecasting crucial, as it represents a concrete financial risk and opportunity for market operators. Therefore, the development of predictive models capable of supporting real-time decisions is essential to effectively manage systemic imbalances.

A comprehensive review of existing literature reveals that while Day-Ahead Market forecasting has reached maturity, with fundamental benchmarks and state-of-the-art models already established by Lago et al. [1], the literature becomes significantly sparser when shifting towards near real-time balancing markets. As highlighted by O'Connor et al. [2], price formation in these markets is highly stochastic, leading to inconsistent forecasting performance that heavily depend on the specific geographical area and the settlement rules adopted by the Transmission System Operators. Furthermore, they demonstrated that in these near real-time markets, strongly regularized regressions and Machine Learning (ML) algorithms frequently outperform more complex Deep Learning (DL) architectures [3]. Specific to imbalance price point forecasting, Deng et al. [4] explored various models, proposing advanced deep learning frameworks to capture extreme price variations. Conversely, Narajewski's work [5] highlighted the significant difficulty of outperforming simple naive benchmarks in terms of standard point metrics (MAE and RMSE), underscoring the extreme unpredictability of imbalance prices.

To date, the point forecasting of Italian imbalance prices has not been addressed. This study aims to fill this research gap by establishing a comprehensive benchmark for the near real-time point forecasting of imbalance prices in the Italian electricity market. To mirror practical trading operations, the forecasting framework simulates the perspective of a market operator, executing point forecasts for a given quarter-hour delivery period under strict real-time constraints (approximately 10 minutes before delivery, or $t - 10$ min), relying exclusively on public data available at that exact moment.

The novelties and contributions of this work include:

- Comparing five distinct algorithms: a regularized linear model (Lasso Estimated AutoRegressive), three Machine Learning algorithms (Random Forest, eXtreme Gradient Boosting, Support Vector Regression), and one Deep Learning network (Long Short-Term Memory) evaluated against a Naive baseline, a simple model serving as a reference point.
- Quantifying the predictive performance of the proposed models under near real-time constraints to establish a realistic benchmark for Italian imbalance price forecasting.
- Evaluating the practical economic value of the developed models by simulating a trading scenario of voluntary imbalance.

## 2. Imbalance Pricing Mechanism

Following the TIDE reform, the Italian market transitioned to a quarter-hourly single pricing mechanism for imbalance settlement. The macrozonal imbalance price $P^{sb}_{mz}$ is determined based on the macrozonal imbalance $S_{mz}$ and the total downward $(\overline{Q \downarrow^{bil}_{mz}})$ or upward $(\overline{Q \uparrow^{bil}_{mz}})$ activated balancing energy:

$$P^{sb}_{mz} = \begin{cases} P^{sb+}_{mz} & \forall t \mid S_{mz} > 0 \land \overline{Q \downarrow^{bil}_{mz}} \neq 0 \\ P^{sb-}_{mz} & \forall t \mid S_{mz} < 0 \land \overline{Q \uparrow^{bil}_{mz}} \neq 0 \\ P^{AE}_{mz} & \text{otherwise} \end{cases}$$

Under this framework, market operators are economically rewarded for system-helping imbalances that contribute to restoring grid equilibrium, whereas they are financially penalized for system-aggravating imbalances.

## 3. Dataset and Feature Selection

To respect near real-time operational constraints, the predictive models exclusively utilize exogenous variables available up to 10 minutes before the delivery period ($t - 10$ min). The selected features encompass the primary economic and physical drivers of the imbalance market:

- **Preliminary imbalance data:** unconsolidated real-time estimates of the imbalance price ($P^{sb}_{mz,prel}$) and volume ($S_{mz,prel}$).
- **Spot market prices:** the Day-Ahead Market zonal clearing price ($P^{MGP}_{z}$), the volume-weighted average price of Intraday auctions ($P^{mean}_{MI-A}$), and the last coupling price from continuous trading ($P^{last}_{XBID}$).
- **Physical system variables:** the aggregated forecast error of solar and wind generation ($Err^{sw}_{gen}$), alongside the total load ($L^{f}_{tot}$) and generation ($G^{f}_{tot}$) forecasts.

Furthermore, the feature set is enriched with time-related variables (dummy or cyclical) and rolling statistics of $P^{sb}_{mz,prel}$. Depending on the model architecture, input data may also undergo variance stabilization and normalization transformations. The analysis focuses exclusively on the North macrozone. To simulate a realistic operational environment, the experimental pipeline is structured into sequential phases. First, the entire dataset, spanning from 1 April 2022 to 30 September 2025, is chronologically divided. The period up to 30 September 2024 is dedicated to training and validation, while the subsequent year serves as an out-of-sample test set. During this initial data preparation phase, multiple historical time lags are selected as predictors for each exogenous feature. Subsequently, a rigorous hyperparameter tuning procedure is applied: depending on the specific architecture, this optimization is conducted either through the Least Angle Regression algorithm combined with the Akaike Information Criterion, or via the Tree-structured Parzen Estimator utilizing the Optuna framework across one thousand trials. Finally, the training and evaluation strategy is defined based on computational constraints. LEAR, RF, XGBoost, and SVR are retrained daily using a 365-day rolling window to continuously learn from the most recent data, whereas LSTM is evaluated using a static hold-out strategy, being trained only once on the initial dataset.

# 4. Results

To establish a reference for performance evaluation, a Naive model is introduced as a baseline; it utilizes the preliminary imbalance price observed 60 minutes before delivery ($t - 60$ min) as the forecast for the quarter-hour $t$.

## 4.1. LEAR Preliminary Analysis

A preliminary analysis on the LEAR model investigated the impact of specific features and time lags by pairing the autoregressive proxy ($P^{sb}_{mz,prel}$) with individual exogenous variables. Tables 1 and 2 show that using a single best-performing lag helps mitigate large price deviations, improving the RMSE, but generally worsens average accuracy compared to the Naive baseline. Conversely, utilizing a deeper sequence of historical data (up to $t - 60$ min) consistently improves all predictive metrics, allowing the LEAR model to fully outperform the benchmark.

The results also indicate a saturation effect regarding the informational value of historical depth: no appreciable improvements are observed when extending the lag window from 60 to 120 minutes, and including lags up to the previous day negatively impacts performance.

Table 1: LEAR performance with $P^{sb}_{mz,prel}$ and single exogenous variables using the single best-performing lag.

| Feature | MAE | RMSE | sMAPE | $R^2$ |
|---|---|---|---|---|
| Naive Baseline | 54.16 | 86.50 | 52.12 | -0.05 |
| $P^{sb}_{mz,prel}$, $P^{MGP}_{z}$ | 56.18 | 77.14 | 54.58 | 0.16 |
| $P^{sb}_{mz,prel}$, $P^{mean}_{MI-A}$ | 55.36 | 76.54 | 53.91 | 0.17 |
| $P^{sb}_{mz,prel}$, $P^{last}_{XBID}$ | 54.55 | 75.98 | 53.41 | 0.18 |
| $P^{sb}_{mz,prel}$, $S_{mz,prel}$ | 57.89 | 80.27 | 57.22 | 0.09 |
| $P^{sb}_{mz,prel}$, $Err^{sw}_{gen}$ | 58.83 | 81.16 | 57.15 | 0.07 |
| $P^{sb}_{mz,prel}$, $L^{f}_{tot}$ | 58.44 | 80.69 | 57.22 | 0.08 |
| $P^{sb}_{mz,prel}$, $G^{f}_{tot}$ | 58.18 | 80.43 | 57.53 | 0.09 |
| $P^{sb}_{mz,prel}$ alone | 58.42 | 81.08 | 56.84 | 0.07 |

*MAE, RMSE expressed in €/MWh, sMAPE in %.*

Table 2: LEAR performance with $P^{sb}_{mz,prel}$ and single exogenous variables using all available lags up to $t - 60$ min.

| Feature | MAE | RMSE | sMAPE | $R^2$ |
|---|---|---|---|---|
| Naive Baseline | 54.16 | 86.50 | 52.12 | -0.05 |
| $P^{sb}_{mz,prel}$, $P^{MGP}_{z}$ | 51.38 | 72.74 | 50.52 | 0.25 |
| $P^{sb}_{mz,prel}$, $P^{mean}_{MI-A}$ | 51.32 | 73.08 | 50.40 | 0.25 |
| $P^{sb}_{mz,prel}$, $P^{last}_{XBID}$ | 50.83 | 72.94 | 50.11 | 0.25 |
| $P^{sb}_{mz,prel}$, $S_{mz,prel}$ | 51.98 | 75.26 | 51.80 | 0.20 |
| $P^{sb}_{mz,prel}$, $Err^{sw}_{gen}$ | 53.28 | 75.80 | 52.26 | 0.19 |
| $P^{sb}_{mz,prel}$, $L^{f}_{tot}$ | 52.73 | 75.39 | 52.15 | 0.20 |
| $P^{sb}_{mz,prel}$, $G^{f}_{tot}$ | 53.41 | 75.86 | 52.53 | 0.19 |
| $P^{sb}_{mz,prel}$ alone | 52.93 | 75.81 | 52.01 | 0.19 |

*MAE, RMSE expressed in €/MWh, sMAPE in %.*

Based on this first assessment, a forward feature selection approach was employed to determine the optimal number and set of exogenous variables. Starting from $P^{sb}_{mz,prel}$, the most promising variables were incrementally added to the model while maintaining the $t - 60$ min lag setup. The optimal result is achieved by combining $P^{sb}_{mz,prel}$ with the three market price features $P^{last}_{XBID}$, $P^{mean}_{MI-A}$, $P^{MGP}_{z}$ and the preliminary imbalance volume $S_{mz,prel}$.

## 4.2. Model Performance

Table 3 compares all implemented models using the $t - 60$ min lag configuration, chosen to balance computational efficiency and accuracy while ensuring a consistent comparison. While the results for LEAR reflect its previously identified optimal five-variable setup, the other architectures utilize the complete set of exogenous features. XGBoost achieves the best MAE (43.77 €/MWh), corresponding to a 19.18% improvement over the Naive Baseline and indicating the highest average accuracy. RF records the best RMSE (65.91 €/MWh), improving by 23.80%, which demonstrates superior robustness against price spikes. SVR achieves the second-best MAE and sMAPE, while LSTM and LEAR record the lowest improvements. Overall, all models outperform the baseline. However, the absolute metric values highlight the extreme difficulty of the forecasting task: the best $R^2$ (0.39, RF) indicates that most of the price variance remains unexplained, and the best sMAPE (43.57%, XGBoost) indicates a high relative error, confirming that forecasts on average deviate by a large percentage from the true price magnitudes. All models successfully capture overall market trends and achieve lower errors when prices remain low and stable, but consistently struggle with sudden price spikes.

Table 3: Comparison of all models using lags up to $t - 60$ min. Best values are in bold.

| Model | MAE | RMSE | sMAPE | R² |
|---|---|---|---|---|
| Naive | 54.16 | 86.50 | 52.12 | -0.05 |
| LEAR | 48.93 | 72.05 | 48.80 | 0.27 |
| RF | 47.99 | **65.91** | 47.60 | **0.39** |
| XGBoost | **43.77** | 67.63 | **43.57** | 0.35 |
| SVR | 45.92 | 69.23 | 45.46 | 0.32 |
| LSTM | 48.66 | 70.32 | 48.12 | 0.30 |

*MAE, RMSE expressed in €/MWh, sMAPE in %.*

Furthermore, the SHAP (SHapley Additive exPlanations) framework was employed to measure feature importance. Table 4 reports these values for each model, expressed as relative percentages, calculated by normalizing the mean absolute SHAP value of each feature against the total impact of all predictors. RF and XGBoost exhibit intuitive patterns, relying predominantly on $P^{sb}_{mz,prel}$ and $S_{mz,prel}$. Conversely, SVR and LSTM assign substantially higher importance to aggregated physical variables ($L^{f}_{tot}$ and $G^{f}_{tot}$), with the preliminary imbalance data being almost negligible for SVR. LEAR model was evaluated specifically in its optimal feature configuration, which explains the absence of values for certain variables.

Table 4: Feature importance (%) for the evaluated models. Top-3 features per model are in bold.

| Feature | LEAR | RF | XGB | SVR | LSTM |
|---|---|---|---|---|---|
| $P^{sb}_{mz,prel}$ | **27.0** | **33.4** | **38.0** | 1.8 | **14.2** |
| $P^{MGP}_{z}$ | **18.3** | 6.5 | 9.1 | 10.2 | 1.9 |
| $P^{mean}_{MI\text{-}A}$ | **32.1** | **12.1** | **11.7** | 4.5 | 3.0 |
| $P^{last}_{XBID}$ | 6.3 | 5.4 | 4.8 | 0.7 | 1.5 |
| $S_{mz,prel}$ | 14.4 | **33.0** | **22.1** | 7.1 | 8.5 |
| $Err^{sw}_{gen}$ | – | 0.4 | 1.0 | **9.6** | 1.2 |
| $L^{f}_{tot}$ | – | 1.3 | 2.3 | **41.9** | **11.6** |
| $G^{f}_{tot}$ | – | 5.2 | 6.5 | **23.6** | **57.2** |
| Rolling | – | 1.6 | 2.3 | 0.0 | 0.4 |
| Time | 2.0 | 1.1 | 2.2 | 0.7 | 0.6 |

## 5. Economic Assessment

Beyond statistical point-forecasting metrics, an economic assessment was conducted to quantify the practical financial impact of the models' predictions through a simulated trading strategy. The study evaluates a scenario where a market operator, with 1 MW of flexible capacity, intentionally deviates from the commercial schedule previously sold on the Day-Ahead Market (DAM) at the zonal price ($P^{MGP}_{z}$). The decision-making logic is driven by the relationship between the forecasted imbalance price ($\hat{P}^{sb}_{mz}$) and the DAM price, categorized into two main strategic cases:

1. **Forecasted Negative Macrozonal Imbalance ($\hat{P}^{sb}_{mz} > P^{MGP}_{z}$)**: the operator anticipates a system energy deficit and decides to inject 1 MW of additional power into the grid compared to the baseline. Profit is generated if the actual realized imbalance price $P^{sb}_{mz}$ is higher than the DAM price, as the marginal megawatt is sold at a premium. However, a financial loss occurs if the actual price is lower than the DAM price, meaning the directional prediction was incorrect. The actual profit for the quarter-hour is formulated as:

$$\text{Actual Profit} = 1\text{ MW} \times 0.25\text{ h} \times \\ \times (P^{sb}_{mz} - P^{MGP}_{z})$$

2. **Forecasted Positive Macrozonal Imbalance ($\hat{P}^{sb}_{mz} < P^{MGP}_{z}$)**: the operator anticipates a system energy surplus and opts to inject 1 MW less than the contracted baseline. Profit is generated if the actual realized imbalance price $P^{sb}_{mz}$ is lower than the DAM price, as the energy deficit is settled at a discount compared to the price at which it was already sold. Conversely, a financial loss occurs if the actual imbalance price exceeds the DAM price, meaning the directional prediction was incorrect. The realized profit is formulated as:

$$\text{Actual Profit} = 1\text{ MW} \times 0.25\text{ h} \times \\ \times (P^{MGP}_{z} - P^{sb}_{mz})$$

Therefore, the realized economic outcome depends entirely on the actual price $P^{sb}_{mz}$. The directional accuracy of the models (i.e., correctly predicting whether $P^{sb}_{mz} \gtrless P^{MGP}_{z}$) ranges from

73.34% (Naive) to 76.81% (XGBoost). Table 5 summarizes the annual financial results.

Table 5: Annual economic results of the simulated trading strategy for each model.

| Model | Revenues [€] | Losses [€] | Profit [€] |
|---|---|---|---|
| Naive | 412,981 | −165,612 | 247,369 |
| LEAR | 417,419 | −161,746 | 255,674 |
| **RF** | **443,307** | **−135,858** | **307,449** |
| XGBoost | 432,103 | −147,061 | 285,042 |
| SVR | 427,115 | −152,049 | 275,066 |
| LSTM | 424,469 | −154,669 | 269,828 |

Despite having the lowest overall directional accuracy (73.94%), RF generates the highest profit (€307,449, +24.28% over Naive). This counter-intuitive finding is explained by the asymmetric nature of the imbalance price spread ($P_{mz}^{sb} - P_z^{MGP}$): positive spreads occur only 39.10% of the time but average 91.98 €/MWh, whereas negative spreads are more frequent (58.68%) but average only −51.68 €/MWh in absolute terms. Figures 1– 3 decompose accuracy, revenues, and losses across ten price spread deciles. As shown in Figure 1, while RF generally underperforms in deciles associated with negative or near-zero spreads, its accuracy sharply increases when the spread turns positive. Consequently, RF captures the most lucrative opportunities (Figure 2) while minimizing severe penalties during high-spread intervals (Figure 3). This demonstrates that predicting high-value events with precision is far more profitable than maintaining uniformly high accuracy across low-value conditions.

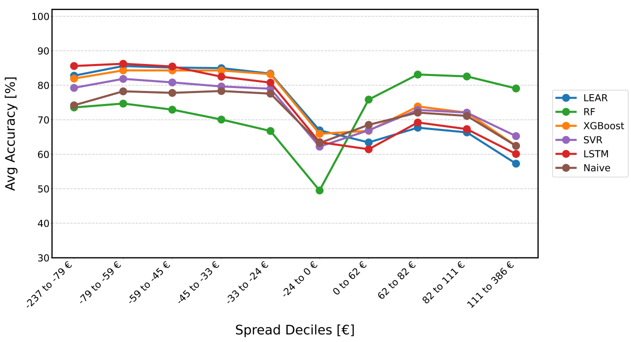


Figure 1: Average directional accuracy

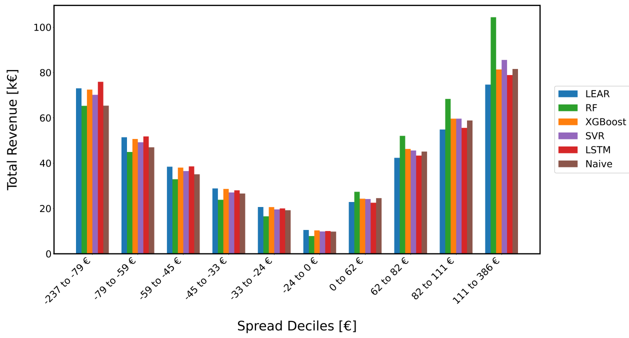


Figure 2: Total gross revenues

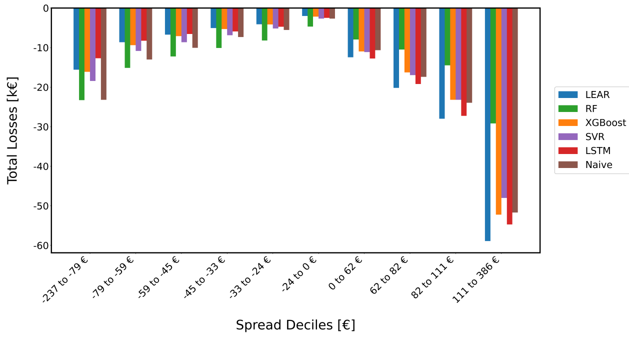


Figure 3: Total financial losses

## 6. Conclusions

This work established a comparative benchmark for the near real-time point forecasting of imbalance prices in the Italian electricity market. Five different predictive models (LEAR, RF, XGBoost, SVR, LSTM) were implemented and evaluated under strict operational constraints ($t - 10$ min), assessing both statistical accuracy and practical economic value through a simulated trading scenario.

The empirical results demonstrate that Machine Learning ensemble methods are the most effective architectures for this task. XGBoost achieved the best typical predictions with a 19.18% improvement in MAE over the Naive baseline, while Random Forest (RF) recorded the lowest RMSE (23.80% improvement), proving more robust against price spikes. Nevertheless, the absolute metric values (maximum $R^2$ of 0.39 and sMAPE over 43%) highlight the extreme stochasticity of the imbalance market, which severely limits the reliability of exact point forecasts.

Despite these statistical limitations, the economic assessment revealed that all models consistently predicted the direction of the system

imbalance. By simulating a trading strategy based on voluntary imbalance, RF generated the highest profit (+24.28% over the baseline) due to its superior performance in capturing high-value price spikes. This confirms that for a market operator, the strategic value of a forecast is dictated by its precision during high-stakes events rather than overall point accuracy in stable conditions. However, it must be noted that intentional imbalance practices are strictly prohibited by current regulations, making these results a measure of predictive value rather than a suggested commercial strategy.
To address current limitations, future work should explore several key areas. First, the transition to probabilistic forecasting would allow to better quantify market uncertainty. Second, new models explicitly designed to handle extreme price spikes could be tested. Furthermore, the optimization process could be refined by directly incorporating economic objectives into the training phase, rather than relying solely on statistical errors. Extending the forecasting horizon beyond the next quarter-hour would also provide market participants with a broader temporal perspective for strategic planning. Finally, given the recent introduction of the TIDE reform and the increasing integration of renewables, price formation dynamics will likely evolve as market participants adapt their behaviors. Therefore, continuous model updating will be essential to ensure ongoing predictive reliability.

## References

[1] Jesus Lago, Grzegorz Marcjasz, Bart De Schutter, and Rafał Weron. "Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark". *Applied Energy*, 293, 2021.

[2] Ciaran O'Connor, Mohamed Bahloul, Steven Prestwich, and Andrea Visentin. "A Review of Electricity Price Forecasting Models in the Day-Ahead, Intra-Day, and Balancing Markets". *Energies*, 18(12), 2025.

[3] Ciaran O'Connor, Joseph Collins, Steven Prestwich, and Andrea Visentin. "Electricity Price Forecasting in the Irish Balancing Market". *Energy Strategy Reviews*, 54, 2024.

[4] Sinan Deng, John Inekwe, Vladimir Smirnov, Andrew Wait, and Chao Wang. "Seasonality in deep learning forecasts of electricity imbalance prices". *Energy Economics*, 137, 2024.

[5] Michał Narajewski. "Probabilistic Forecasting of German Electricity Imbalance Prices". *Energies*, 15(14), 2022.