Lisbon School of Economics & Management
Universidade de Lisboa

Master

Mathematical Finance

Master´s Final Work

Internship Report

Forecasting Electricity Prices in the Portuguese Market: A Machine Learning Approach

Daniel de Jesus Silva Campanário

June-2025

MASTER
MATHEMATICAL FINANCE

MASTER´S FINAL WORK
INTERNSHIP REPORT

FORECASTING ELECTRICITY PRICES IN THE PORTUGUESE MARKET: A MACHINE LEARNING APPROACH

DANIEL DE JESUS SILVA CAMPANÁRIO

SUPERVISION:
DINU GRANACI
PROFESSOR JOÃO AFONSO BASTOS

JUNE-2025

Glossary

EDA – Exploratory Data Analysis

GARCH – Generalized Autoregressive Conditional Heteroskedasticity

GDP – Gross Domestic Product

JEL – Journal of Economic Literature

LightGBM – Light Gradient Boosting Machine

MAE – Mean Absolute Error

ML – Machine Learning

MSE – Mean Squared Error

MFW – Master's Final Work

OLS – Ordinary Least Squares

RMSE – Root Mean Squared Error

RNN – Recurrent Neural Network

LSTM – Long Short-Term Memory

SME – Small and Medium-sized Enterprise

SVR – Support Vector Regression

ABSTRACT, KEYWORDS AND JEL CODES

The Portuguese electricity market, integrated within the MIBEL, exhibits high volatility and structural complexity, primarily driven by the growing share of intermittent renewable energy, unpredictable demand patterns, and meteorological variability. These factors pose significant challenges for SMEs, which often lack the analytical tools required to anticipate price fluctuations and manage energy costs effectively.

This dissertation proposes a machine learning-based forecasting model tailored to predict hourly electricity prices over a seven-day horizon. The model is specifically designed to support industrial SMEs in improving energy planning and mitigating exposure to price risk. A comprehensive dataset was constructed, comprising hourly observations from March 2020 to March 2025 and incorporating 37 variables across five key dimensions: energy production, consumption, market prices, cross-border exchanges, and weather conditions.

The methodological framework combines robust preprocessing techniques, including outlier mitigation, robust normalization, and one-hot encoding, with advanced learning algorithms. LightGBM was selected for its predictive performance and scalability. Hyperparameter tuning was conducted using Bayesian optimization via Optuna.

The final model achieved a MAE below 6 €/MWh, in line with industry standards for short-term forecasting. Results underscore the relevance of meteorological factors and cross-border dynamics in shaping market behaviour.

This study contributes a practical and interpretable tool that enhances SMEs' decision-making capacity, while also demonstrating the effectiveness of machine learning methods in navigating the complexities of modern electricity markets.

KEYWORDS: Electricity Markets; Time Series Forecasting; LightGBM; Feature Engineering; SMEs; Demand-side Risk Mitigation

JEL CODES: C53; Q41; Q47; C52; L94; M21.

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my entire family and friends for their unwavering support, constant encouragement, and unconditional belief in me throughout this journey. Their presence, both in moments of triumph and difficulty, has been essential to my personal and academic growth.

A heartfelt thank you also goes to everyone who accompanied me during this process, whether through guidance, shared experiences, or simple words of motivation. Each contribution, however small it may seem, played a fundamental role in the completion of this work.

I extend my sincere appreciation to Vannaci Prime for the trust placed in me and for the opportunity to develop this project during my internship. The support provided by the company, along with the autonomy and resources made available, were crucial in enabling me to pursue a project of real impact and academic value. I am particularly grateful for the professional environment that stimulated critical thinking and allowed me to apply and deepen the knowledge acquired throughout the master's program.

I am especially thankful to Professor João Bastos for his availability, dedication, and willingness to assist throughout this process. His guidance and readiness to help were essential for the successful development of this work.

Finally, I thank my academic supervisors, professors, and the ISEG community for their valuable insights and guidance, which helped shape the direction and quality of this research.

To all of you, thank you.

1. INTRODUCTION

This report, developed as part of the completion of the Master's Degree in Mathematical Finance at ISEG, results from a four-month internship at Vannaci Prime, within the Finance Department. During this period, an extensive project was conducted focusing specifically on forecasting electricity prices in the Portuguese market. The project was strategically oriented towards addressing the critical needs of Small and Medium-Sized Enterprises operating within the industrial sector, where electricity costs constitute a substantial and often volatile component of overall operational expenses.

In recent years, the Portuguese electricity market, an integral component of the MIBEL, has increasingly gained attention due to its inherent complexity and significant volatility. Several factors contribute to this unpredictability, chief among them being the variability of renewable energy production, fluctuations in energy demand, and the highly variable nature of meteorological conditions. The prevalence and increasing penetration of renewable energy sources, such as wind and solar, whose outputs are intrinsically intermittent and weather-dependent, have significantly altered market dynamics. These renewable sources, despite their environmental benefits, pose substantial forecasting challenges, as their production levels are subject to rapid and unpredictable variations. Consequently, electricity prices have become more sensitive and prone to sudden fluctuations, creating significant risks for energy-dependent industrial SMEs.

A key characteristic of electricity as a commodity is its inability to be stored efficiently at scale. This limitation necessitates a real-time balancing of supply and demand, exacerbating market volatility. Short-term imbalances, even minor ones, can result in significant price spikes, which further complicate the already challenging environment faced by market participants. SMEs, with limited resources and expertise in advanced energy management practices, often find themselves particularly vulnerable to these unpredictable price shifts, which can adversely affect their cost management strategies and profitability.

Recognizing these challenges, the internship aimed to develop a robust yet user-friendly forecasting tool explicitly tailored to the needs of industrial SMEs. The primary goal was to enable these enterprises to make informed decisions regarding their energy consumption, ultimately leading to more efficient cost management and enhanced competitiveness. To ensure practical relevance and utility, the forecasting model was designed to achieve a MAE of less than 6 €/MWh for hourly electricity price predictions over a seven-day forecasting horizon. This benchmark was carefully selected to reflect the industry standard for predictive accuracy, ensuring the model's relevance and usability in practical, real-world scenarios.

To realize this objective, a comprehensive and extensive database was compiled, encompassing hourly data across a five-year period from March 20, 2020, to March 20, 2025. This dataset consisted of 37 variables that represent diverse and critical aspects of the electricity market. These variables included detailed data on energy production segmented by source (renewable and conventional), consumption patterns, real-time and historical meteorological conditions, market prices, and cross-border energy exchanges between Portugal and Spain. The richness and granularity of this dataset provided a robust foundation for developing highly accurate and reliable forecasting models.

The methodological approach employed in this project was centered around advanced machine learning techniques. These methods have increasingly proven their efficacy in capturing complex, non-linear relationships within large and multivariate datasets. In particular, the models developed leveraged algorithms such as LightGBM, known for its efficiency and predictive accuracy, and various neural network architectures, renowned for their capability to model intricate temporal dependencies. During the modeling phase, significant emphasis was placed on carefully fine-tuning and optimizing these models. This included rigorous hyperparameter optimization and validation processes, leveraging techniques such as cross-validation and Bayesian optimization. The objective was to strike a balance between high predictive accuracy and simplicity, ensuring that the final solution would be accessible and comprehensible even to users without specialized technical backgrounds.

An essential component of this project was also the extensive preprocessing of data, which is critical for the performance of any ML-based forecasting solution. Various preprocessing steps were applied, including robust normalization (Robust Scaler) of numerical variables to mitigate the impact of outliers and encoding categorical variables, such as wind direction, using techniques like One-Hot Encoding. These preprocessing measures ensured that the underlying data was well-suited to the modeling tasks, thereby enhancing model stability and performance.

Structurally, this report follows a logical and coherent sequence designed to facilitate a clear and thorough understanding of the research conducted. Initially, the paper contextualizes the research problem and delineates its practical importance, particularly highlighting the specific challenges faced by SMEs in managing energy costs within a volatile market. Subsequently, a comprehensive review of existing literature is presented, establishing a theoretical foundation for electricity price forecasting, with a particular emphasis on machine learning applications and methodologies. Following this theoretical underpinning, the methodological approach is detailed, clearly outlining data collection, preprocessing strategies, and the specific modeling techniques employed. The subsequent section focuses on the empirical results obtained, providing a detailed assessment of the forecasting model's performance against established benchmarks and objectives. Critical insights and performance metrics are discussed, offering a clear evaluation of the tool's predictive efficacy. Finally, the report concludes with a synthesis of the key findings, acknowledges the inherent limitations of the study, and proposes potential avenues for future research. This conclusion underscores the project's value and utility, emphasizing its potential impact on improving energy management practices within industrial SMEs, ultimately enhancing their competitiveness and operational efficiency in the increasingly dynamic energy market.

2. LITERATURE REVIEW

2.1 The Portuguese Electricity Sector

2.1.1 – Price Formation Mechanism and Structural Volatility

The Portuguese electricity sector is integrated into the Iberian Electricity Market, a supranational market that brings together the electricity systems of Portugal and Spain, promoting efficiency and competitiveness through the free formation of electricity prices. MIBEL was formally established with the objective of harmonizing the operational rules of the energy markets of the Iberian Peninsula, enhancing liquidity, transparency, and security of supply, factors considered essential for the creation of a unified European energy market.

The Portuguese System Price formation in MIBEL is based on the marginal pricing principle, according to which the price of electricity in each hourly period is determined by the cost of the last unit of energy required to meet demand. This mechanism, applied in both the day-ahead and intraday markets, aims to ensure economic efficiency by encouraging producers to offer energy at the lowest possible marginal cost. However, this methodology exhibits high sensitivity to short-term volatility, the intermittency of renewable energy production, and exogenous factors such as weather conditions and the availability of energy resources. These elements introduce additional complexity to the price formation process, making the market more unstable and subject to abrupt fluctuations.

The Portuguese electricity market is segmented into different trading mechanisms, which primarily differ in terms of the time horizon of the transactions conducted. The day-ahead market constitutes the main pricing mechanism, where market participants submit their purchase and sale offers for electricity for the following day. Reference prices are determined through auctions based on the marginal pricing methodology, ensuring efficient price formation based on supply and demand dynamics. Additionally, the intraday market allows participants to adjust their positions closer to the time of energy delivery, enabling the correction of deviations from initial forecasts and reflecting unexpected changes in production or consumption conditions. This market contributes to greater operational flexibility and efficiency within the electricity system.

2.1.2 Challenges in Electricity Price Forecasting

A study by Weron (2014) emphasizes that electricity price forecasting remains highly challenging under volatile market regimes, particularly during structural breaks caused by regulatory shifts or energy crises. The authors show that even advanced multivariate modeling frameworks, which typically outperform simpler models under stable conditions, can struggle to maintain accuracy and generalization when confronted with sudden regime changes, highlighting the inherent fragility of forecasting systems in such scenarios. Forecasting electricity prices is widely recognized as one of the most complex challenges in liberalized markets due to the high levels of volatility and the frequent occurrence of price spikes, which are abrupt and extreme variations over short time intervals. These price spikes are often the result of sudden imbalances between supply and demand and remain particularly difficult to anticipate with precision. Their occurrence is closely linked to structural factors, notably the intermittence of renewable energy sources particularly wind and solar whose generation capacity is highly dependent on meteorological variables marked by substantial uncertainty and low predictability. As highlighted by Zamudio Lopez, Zareipour, and Quashie (2024), such volatility is often exacerbated by forecasting errors in intermittent generation and abrupt shifts in electricity consumption, reinforcing the intrinsic challenges in predicting these events accurately. Other factors include changes in electricity demand, limitations in electricity storage capacity that require almost instantaneous adjustments between supply and demand, thus exacerbating price volatility, restrictions in cross-border interconnection capacity, which limit the system's flexibility in responding to supply or demand shocks and many others.

In addition to these structural factors, electricity price forecasting faces the added complexity of incorporating exogenous high-impact events, often related to geopolitical and macroeconomic dynamics, whose anticipation and statistical modelling are extremely challenging. A paradigmatic example is the 2022 energy crisis, triggered by the invasion of Ukraine and the subsequent restrictions on natural gas supplies to Europe. This context led to an increase in energy prices, with direct impacts on the Iberian market, where electricity prices reached historic highs, exceeding 500 €/MWh at certain times. This episode highlights the limitations of both traditional models and more advanced machine learning and deep learning models, which, despite their effectiveness in identifying complex historical patterns, exhibit weaknesses in forecasting structural disruptions and unexpected external shocks. This limitation is corroborated by Ghelasi and Ziel (2024), who demonstrated that even econometric models enriched with fundamental market information strove to anticipate the magnitude and persistence of the price arises observed during this crisis. Their findings further emphasize that, while such models perform well under normal conditions, their predictive capacity deteriorates significantly in the presence of unprecedented geopolitical shocks. Notably, they also report similar difficulties in capturing market behavior during earlier periods of extreme volatility, such as the 2021 emerges in European gas prices, reinforcing the notion that external and structural disruptions remain a major forecasting challenge for state-of-the-art models.

2.1.3 The importance of Hourly Forecasting for Market Participants

In a context marked by high volatility and structural uncertainty, hourly electricity price forecasting plays a strategic role for market participants. The ability to anticipate price fluctuations enables the formulation of more efficient buying and selling strategies, risk mitigation, and more rational management of operational costs.

As pointed out by Conejo et al. (2005), hourly price forecasts are especially valuable for large consumers and industrial participants, who rely on precise cost anticipation for operational and hedging strategies. From a theoretical perspective, forecasting electricity prices is hindered by the high complexity of the market, which is characterized by strong non-linearities, multivariate interdependencies among economic, meteorological, and technical factors, and the presence of heteroscedasticity, where the variability of prices depends on their own levels.

Additionally, the increasing penetration of intermittent renewable energy sources, combined with the lack of economically viable large-scale storage solutions, exacerbates market price instability. This reality creates a highly uncertain environment in which traditional predictive models are often inadequate to capture the underlying complexities.

2.2 Comparison Between Traditional Models and Machine Learning Approaches

2.2.1 Limitations of classical statistical models

Classical statistical models such as ARIMA, SARIMA and GARCH have long been applied to univariate time-series forecasting, including in the energy sector. However, their practical use in electricity markets is limited by underlying assumptions that rarely hold in this context. These models generally posit that the future value of the variable depends on past observations through structures such as

$$(1) \qquad y_t = c + \sum_{\{i=1\}}^{\{p\}} \phi_i \, y_{\{t-i\}} + \sum_{\{j=1\}}^{\{q\}} \theta_j \varepsilon_{\{t-j\}} + \varepsilon_t$$

where $\varepsilon_t \sim \mathcal{N}(\,0, \sigma^2)$ and $\phi_i, \theta_j \in R$, and typically require the series to be stationary with constant mean and variance. Yet electricity prices are strongly influenced by sudden exogenous shocks—such as the 2022 energy crisis or unexpected grid outages—as well as by abrupt spikes, multiple seasonalities, and stochastic volatility. Moreover, their dynamics depend heavily on external drivers including meteorological conditions, consumption patterns, and regulatory interventions, which univariate frameworks cannot directly accommodate. Even models designed to capture time-varying volatility, such as GARCH, specify a conditional variance following

$$(2) \qquad \sigma_t^2 = \alpha_0 + \sum_{\{i=1\}}^{\{p\}} \alpha_i \varepsilon_{\{t-i\}}^2 + \sum_{\{j=1\}}^{\{q\}} \beta_j \sigma_{\{t-j\}}^2$$

and face difficulties when confronted with the structural breaks, large jumps, and multivariate interactions common in electricity markets. Consequently, while these approaches can serve as useful baselines, they are generally inadequate to fully describe the complexity and rapidly changing dynamics observed in liberalized power markets (Weron, 2014).

2.2.2 Advantages of Machine Learning Approaches

According to Taieb and Hyndman (2021), the growing emphasis on explainable machine learning models in energy systems has led to the increasing adoption of interpretable algorithms such as tree-based methods over deep learning architectures, particularly in industrial and regulatory contexts.

This represent a substantial evolution in relation to traditional statistical models, as they allow classic assumptions to be relaxed and complex, non-linear relationships to be modelled. Instead of starting from a pre-defined functional form, ML models learn directly from the data, automatically adjusting to the structure of the observed information.

ML models deal with functional relationships of the type:

$$(3) \qquad \hat{y}_t = f(X_t; \theta)$$

where f is a non-parametric function learnt from the data, $X_t \in R^d$ represents a vector of explanatory variables (internal and exogenous), and θ are the parameters (or weights) to be optimized based on an error criterion such as MAE or MSE.

Among the most relevant machine learning approaches for time series forecasting, tree-based models such as Gradient Boosting Machines are especially effective due to their capacity to model complex, high-order interactions without requiring these relationships to be explicitly defined. These models are well suited to heterogeneous datasets, as they accommodate exogenous variables with varying frequencies and scales, including meteorological indicators, market prices, and operational metrics. In addition, deep learning architectures such as Recurrent Neural Networks and Long Short-Term Memory networks offer powerful mechanisms for capturing temporal dependencies and dynamic structures within sequential data. A key advantage of both approaches lies in their ability to adapt to evolving data patterns by retraining with updated observations, thereby enhancing their responsiveness to structural changes in highly volatile environments such as electricity markets.

The robustness of machine learning models is highly dependent on the ability to prevent overfitting, which compromises the generalizability of the model to unseen data. Several regularization and optimization techniques are commonly employed to address this issue. One widely used approach is L1/L2 regularization, which penalizes the complexity of the model by minimizing the loss function:

$$(4) \qquad \mathcal{L}(\theta) = \sum_t (y_t - \hat{y}_t)^2 + \lambda \|\theta\|_p$$

Where the value of p determines the type of regularization applied, such that when $p = 1$ the formulation corresponds to the LASSO, while p = 2 leads to Ridge regression.

Cross-validation is another essential method, whereby the dataset is divided into multiple folds to evaluate the model’s ability to generalize beyond the training data. In addition, early stopping is often applied during training to halt the learning process when the validation error begins to increase, thereby mitigating overfitting. Finally, the selection of hyperparameters can be automated through Bayesian optimization techniques, such as those implemented with Optuna, which efficiently search the hyperparameter space based on empirical performance during model validation

Scalability is one of the main advantages of machine learning models such as LightGBM, which are designed to operate efficiently in environments with high dimensionality and data volume. These algorithms allow for reduced training times even in databases with millions of observations, using optimized discretized histogram structures to reduce computational complexity while performing automatic variable selection based on metrics such as information gain. In the field of electricity price forecasting, these capabilities prove particularly suitable, given the multivariate and heterogeneous nature of the data involved including meteorological, operational and market variables as well as the presence of seasonal patterns, non-linear relationships and non-stationary behavior. The high predictive accuracy that these approaches make possible is essential to support decision-making processes related to risk management, production planning and the formulation of energy procurement strategies by the sector's agents. This perspective is strongly supported by Rubattu, Maroni, and Corani (2023), who demonstrate the effectiveness of LightGBM in electricity load and peak forecasting tasks, highlighting its robustness in handling high-dimensional inputs, fast training capabilities, and suitability for multivariate time series with complex temporal dependencies, such as those found in the energy sector. These insights naturally lead to a more focused discussion on the algorithm itself, its core mechanisms, and the reasons why it emerges as an effective and pragmatic choice in the context of electricity price forecasting.

2.2.3 LightGBM as an Effective Intermediate Solution

Among the various machine learning algorithms, the LightGBM has stood out as an efficient and robust solution for forecasting tasks in environments with high dimensionality and temporal granularity, such as the electricity market. Developed by Microsoft, this model is based on decision trees and uses a leaf-wise growth strategy, as well as the use of discretized histograms for computational optimization. These features give it high training speed, even on datasets with millions of observations, and efficient use of memory.

Empirical comparisons such as those by Deng et al. (2023) show that LightGBM consistently offers an excellent balance between accuracy, interpretability, and computational efficiency in electricity price forecasting, making it a strong candidate for industrial deployment. LightGBM has shown superior performance compared to linear models and conventional tree-based algorithms such as Random Forest (Breiman, 2001), particularly in metrics such as mean absolute error and root mean square error. In addition to its predictive power, it allows the importance of variables to be quantified, which is an asset from an interpretative point of view, as it makes it easier to identify the factors with the greatest influence on price behavior. In addition, it is more computationally efficient than deep neural networks in situations where computational resources are limited or data volumes are high, without significantly compromising forecast accuracy. Another relevant aspect is its adaptability to sparse and heterogeneous data, such as that which characterizes the electricity sector, where meteorological, operational and market variables coexist. This versatility makes LightGBM particularly suitable for applications that require a compromise between performance, interpretability and computational feasibility, and it is therefore widely adopted in industrial and institutional settings.

2.2.4 Limitations and Challenges of Machine Learning Approaches

Despite the obvious advantages associated with approaches based on machine learning and deep learning, it is necessary to recognize the limitations inherent in these methodologies. As highlighted by Lago et al. (2021), machine learning models for electricity price forecasting are particularly sensitive to input data quality and outlier handling, which significantly influence model robustness and error variance under stress scenarios. The selection and fine-tuning of hyperparameters is another critical point, requiring rigorous validation methodologies and often the use of intensive computational optimization techniques.

One of the most recurrent criticisms relates to the low interpretability of the models, which are often classified as black boxes. This opacity can be an obstacle to their acceptance by decision-makers and stakeholders in contexts where transparency in decision-making is essential, such as in the regulation of the electricity sector or in risk management in corporate environments.

In addition, these models do not explicitly incorporate structural knowledge about the underlying system, relying exclusively on statistical inference from historical data. In scenarios of marked structural change such as those that occur as a result of energy crises, regulatory changes or disruptive technological transformations this limitation can compromise the ability to generalize and therefore the reliability of forecasts (Nowotarski & Weron, 2018) .

3.Pre-processing and Exploratory Analysis Methodology

3.1 Data acquisition

The construction of a robust and predictive dataset is a cornerstone of any data-driven modelling process. In this dissertation, the initial dataset was compiled by aggregating hourly observations from multiple data sources, covering the Portuguese electricity system from March 20, 2020, to March 20, 2025. The primary objective of this database is to support the development of accurate short-term electricity price forecasting models by ensuring the inclusion of exogenous variables with significant explanatory power over market dynamics.

The dataset comprises 32 numerical features and 5 categorical features, which are detailed in the tables IV and V, distributed across five conceptual dimensions: temporal structure, generation mix, market prices, cross-border exchanges, and load variables. These dimensions encapsulate both endogenous and exogenous determinants of electricity price formation in the MIBEL, with a particular focus on the Portuguese bidding zone.

Each observation is timestamped at an hourly frequency, ensuring temporal alignment across all variables. The temporal component includes not only time indexes (hour, day, month, year) but also engineered features such as sine and cosine transformations to capture cyclic seasonality (e.g., daily and yearly periodicity), indicators for weekends and holidays. These sine and cosine transformations are specifically designed to encode the inherent cyclical structure of time-related patterns, enabling the model to recognize and learn from recurring seasonal behaviors without introducing artificial discontinuities. For instance, although hours 23 and 0 are consecutive in real time, treating them as raw numerical values would misleadingly imply a large gap. By projecting time variables onto a unit circle using sine and cosine functions, this discontinuity is eliminated, preserving both proximity and periodicity in the data representation. This approach is particularly effective in capturing daily and annual seasonality in electricity prices. To further reflect temporal structure, the dataset also includes categorical indicators for weekends and holidays, which help account for structural breaks in price trajectories driven by calendar-related effects. These variables are essential to model periodic behaviors and anticipate structural breaks in price trajectories due to calendar effects.

The generation mix component comprises actual aggregated production values, disaggregated by technology. This includes both renewable sources and conventional sources. The inclusion of these variables enables the model to capture the variability in supply conditions, especially considering the intermittency of renewables and their impact on the marginal price setting mechanism.

In addition, market-related features include the system marginal price for both the Portuguese and Spanish zones, enabling the study of market coupling effects. The interconnection balance, represented by imports and exports between Iberian regions, serves as a proxy for regional supply-demand imbalances and transmission constraints.

Finally, the load variables represent the actual electricity consumption and forecasted demand levels, providing insights into consumption patterns and load forecasting errors, which are crucial in explaining short-term price volatility.

The next sections describe the steps taken to preprocess this dataset, addressing issues such as missing data, outliers, multicollinearity, and the scaling and encoding procedures applied to numerical and categorical variables, respectively. These transformations are critical to ensure the stability and generalizability of machine learning models trained on this dataset.

3.2 Exploratory Data Analysis

Before proceeding to the model-building stage, a comprehensive Exploratory Data Analysis was conducted in order to characterise the statistical properties of the dataset, understand variable behaviour, identify outliers, and assess potential multicollinearity and feature redundancy. This phase plays a crucial role in guiding the selection and transformation of variables for predictive modelling.

The univariate analysis began with the visual inspection of boxplots for all continuous variables, which revealed a non-negligible presence of outliers, particularly in features related to energy generation and meteorological factors. Noteworthy examples include precipitation in Peso da Régua (19.5% outliers), exports from Portugal to Spain (18.9%), and hydroelectric consumption in pumped-storage units (10.4%). Other variables, such as solar radiation, solar generation, and Iberian marginal prices, exhibited moderate levels of extreme values ranging from 5% to 10%. This empirical evidence underscored the need for robust pre-processing techniques to mitigate their influence on downstream models.

Histogram analysis confirmed that most variables significantly deviate from the normal distribution, exhibiting skewness, long tails and multimodal behaviour. These characteristics are particularly pronounced in solar radiation, precipitation and export flows, which justifies the adoption of non-parametric methods and robust scaling techniques, discussed in the following section.

To investigate the presence of linear dependencies among the features, a Pearson correlation analysis was performed. The resulting correlation matrix, shown in Figure 3.1, reveals high pairwise correlations within thematic groups, especially between temperature readings across cities, between load forecasts and actual load, and among cross-border energy flow variables.

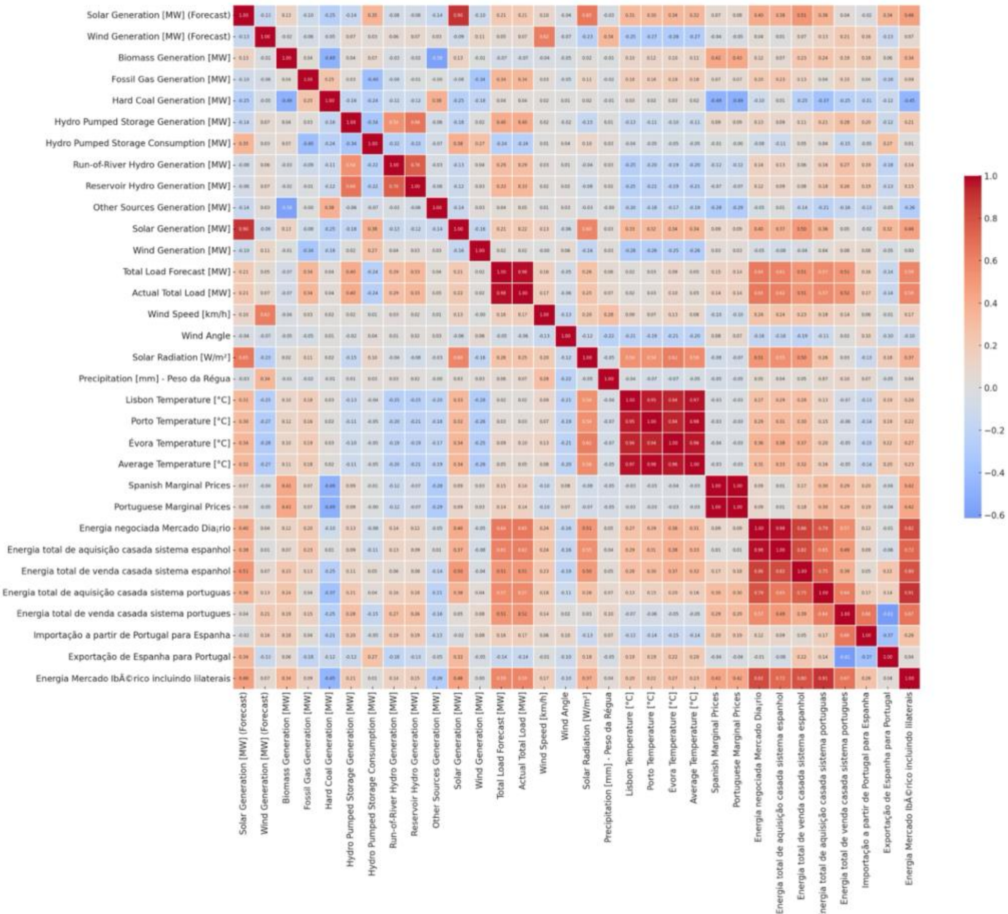

Figure 1-Pearson Correlation Matrix between the main explanatory variables

The visualisation provided by this matrix served as a critical decision-making tool in the feature selection process. Variables exhibiting high collinearity were removed to reduce redundancy and mitigate the risk of overfitting. For example, temperature series for Évora and Porto, as well as the aggregated average temperature, were excluded in favour of the Lisbon temperature, which retained strong explanatory power with lower redundancy.

In the domain of load variables, the Day-ahead Total Load Forecast was excluded due to its near-perfect correlation with the Actual Total Load, which is preferred as it reflects realised demand rather than anticipated values. Similarly, the marginal price in the Spanish market, although intuitively relevant, was discarded to avoid potential data leakage, as it may contain information temporally aligned with the Portuguese marginal price, our model's target variable.

Redundant variables in the category of energy trade and commercial flows, such as daily traded energy volumes, total purchases and sales and exports from Spain to Portugal, were also removed. These variables either overlapped informationally with others or measured the same underlying dynamic from different angles.

Lastly, the weekend indicator was omitted given its informational redundancy with the day-of-week variable, which is more granular and flexible for modelling weekly seasonality. The dimensionality reduction and feature filtering procedures carried out in this section ensured the parsimony and stability of the dataset while preserving relevant explanatory information for the subsequent modelling stage.

3.3 Data Pre-processing

Following the selection of the most relevant variables, the dataset underwent a rigorous pre-processing stage to ensure its structural adequacy for the application of machine learning algorithms. This phase was essential for eliminating sources of distortion, standardizing variable formats, and optimizing the model's ability to learn from historical patterns.

A primary focus of this stage was the normalization of numerical variables, which aimed to mitigate the influence of scale disparities and the presence of extreme values (outliers). As revealed in the exploratory analysis, several numerical features exhibited high skewness and a significant proportion of outliers. Consequently, the Robust Scaler was adopted as the normalization method of choice:

$$(5) \qquad x_t scaled = \frac{x - \mathrm{Median}(x)}{\mathrm{IQR}(x)} \qquad\qquad (6) \qquad \mathrm{IQR}(x) = Q_3(x) - Q_1(x)$$

Unlike conventional transformations such as the StandardScaler, which standardizes data by centering it around the mean and scaling it by the standard deviation, and tends to perform optimally when the data distribution is approximately symmetric or Gaussian, or the MinMaxScaler, which rescales features to a fixed range (typically [0, 1]), compressing variance in the presence of outliers, the Robust Scaler operates based on the median and the interquartile range (IQR).

This makes the Robust Scaler less sensitive to skewness and outliers, preserving the internal structure of the data in the presence of heavy tails or extreme values.All numerical features, with the exception of the target variable, were standardized using the Robust Scaler. The target variable, representing the marginal price of the Portuguese electricity system, was deliberately maintained in its original scale to preserve its economic interpretability. This decision ensures that the model's output remains directly comparable to real-world price values, facilitating interpretation and communication with non-technical stakeholders.

In addition to the normalization of continuous variables, categorical variable encoding was performed. In particular, the wind direction variable, originally recorded as a nominal feature with labels such as N, NE, SW, etc., required transformation to be usable in numerical modelling frameworks. Given the lack of ordinal relationship among these categories, the One-Hot Encoding technique was employed. This approach created a set of binary indicator variables, each corresponding to a specific wind direction, and marked their presence or absence for each hourly observation.

This encoding method preserves the nominal nature of the variable and avoids the introduction of artificial numerical relationships that would result from inappropriate ordinal encoding. Furthermore, One-Hot Encoding is highly compatible with tree-based models such as LightGBM, as it allows the algorithm to assess the importance of each category independently, thereby enhancing interpretability and model performance.

Overall, the pre-processing pipeline implemented in this stage established a clean, structured, and standardized dataset, ready to be ingested by machine learning models with minimal risk of bias or instability. This ensured a strong foundation for the subsequent modelling and evaluation phases.

4. Testing and validation of predictive models

4.1 Benchmarking Models with Default Parameters

In order to establish a rigorous comparative basis for evaluating the predictive performance of the models, an initial empirical analysis was carried out with different machine learning algorithms, using only their default settings and a fixed number of estimators (100 trees or iterations, depending on the model structure). The main aim of this phase was to quickly gauge the relative performance of each approach without carrying out any kind of hyperparameter optimization, thus acting as an initial filter for selecting the models to be explored further in the following phases of the work.

The data set was divided up over time to realistically simulate the real-time forecasting process. Data from March 2020 to December 2023 was used to train the models, while the year 2024 was reserved for model validation and the year 2025 was used entirely for out-of-sample evaluation (test set). This separation guarantees not only the integrity of the training process, but also a robust assessment of the algorithms' ability to generalize in future periods.

As an initial reference, a naïve model (Naïve Forecasting) was implemented which assumes that the marginal hourly price of electricity is h is equal to the value observed at the same time in the previous week:

$$(5) \qquad \hat{y}_h = y_{h-168}$$

This extremely simple approach, based solely on weekly seasonality, reflects a valid baseline for electricity markets, where patterns of hourly and weekly repetition are evident. This model showed a mean absolute error of around 23 €/MWh, which served as a lower benchmark for comparison with the more sophisticated models.

The next phase consisted of applying six models based on decision trees, namely a single Decision Tree, Random Forest, Extra Trees, LightGBM and XGBoost, all trained with 100 estimators and without tuning the respective hyperparameters. It should be noted that, in this exploratory stage, the performance analysis was exclusively guided by the mean absolute error criterion, without explicit consideration of complexity metrics, regularization or signs of overfitting. The aim was therefore to identify the algorithms with the greatest predictive capacity under standard conditions, before implementing optimization strategies. The models were evaluated on the three subsets of data defined - training, validation and testing - and the results obtained are summarized in table I.

TABLE I-FORECASTING PERFORMANCE OF TREE-BASED MODELS WITH STANDARD HYPERPARAMETERS

MODEL	**MAE (€/MWH)**
DECISION TREE	14.39
RANDOM FOREST	9.66
EXTRA TREES	9.14
LIGHTGBM	9.38
XGBOOST	9.43

The results show that the Extra Trees model had the best overall performance, with the lowest mean absolute error in all the data subsets, especially the test set. Gradient boosting models, such as LightGBM and XGBoost, followed closely, with slightly higher errors than Extra Trees but outperforming the Random Forest model.

This benchmarking phase made it possible to identify the algorithms with the greatest potential for generalization and predictive capacity in non-optimized conditions, justifying the selection of the Extra Trees, LightGBM and XGBoost models for subsequent calibration, optimization of hyperparameters and evaluation in more demanding contexts, including periods of stress or market transition.

4.2 Model Optimization with Optuna and Evaluation by Cross-Validation

After identifying the three models with the best predictive performance in standard configurations namely Extra Trees, XGBoost and LightGBM. A second experimental phase was carried out with the aim of reducing the mean absolute error (MAE) through hyperparameter optimization and cross-validation techniques. The main purpose of this stage was to systematically explore the space of possible configurations for each algorithm, seeking to maximize its generalization capacity.

Optimization was carried out using the Optuna library, which implements an approach based on Bayesian optimization, using sequential sampling and learning based on the history of previous evaluations. To ensure a robust evaluation and reduce the statistical variance of the results, a cross-validation strategy with k = 5 folds was adopted in all tuning processes.

The results obtained are shown in table II. It can be seen that the Extra Trees model, after optimization, achieved an MAE of only 8.85 €/MWh in the test set, while XGBoost and LightGBM achieved 8.18 €/MWh and 8.12 €/MWh respectively. However, it should be noted that the average errors in the training sets were substantially lower, with values such as 0.0000191 in the case of Extra Trees, which is indicative of an overfitting problem.

TABLE II-FORECASTING PERFORMANCE AFTER HYPERPARAMETER OPTIMIZATION WHIT OPTUNA

MODEL	**MAE TRAIN (€/MWH)**	**MAE VAL (€/MWH)**	**MAE TEST (€/MWH)**
EXTRA TREES	0.0000191	9.09	8.85
XGBOOST	1.57	8.38	8.18
LIGHTGBM	2.12	8.33	8.12

The discrepancy between performance in the training and validation/test sets, particularly in the Extra Trees model, raises concerns about the generalization capacity of the optimized models. This phenomenon can be attributed to the high number of trees and the depth allowed during optimization, which leads the models to overfit the training data, capturing noise instead of relevant structural patterns.

Despite this, the overall results represent a significant improvement over the models in standard configuration, with reductions of more than €1/MWh in the test errors for the optimized models. This phase confirms the usefulness of hyperparameter calibration in improving predictive performance, while highlighting the need for more rigorous regularization and validation mechanisms, especially in highly complex stochastic environments such as electricity markets.

The following section provides a more detailed analysis of the predictive behavior of these models in critical and regular periods, as well as assessing the stability of the forecasts over the hourly horizon.

4.3 Reducing Predictive Error and Controlling Overfitting

Following the optimization stage using the Optuna algorithm, although there was a significant reduction in the mean absolute error in out-of-sample data, the results showed clear signs of overfitting, particularly in the performance of the Extra Trees model, whose error in the training set was close to zero. This asymmetry prompted the development of a new methodological stage focused on mitigating overfitting and improving the generalization of the models by incorporating new explanatory variables with a temporal structure and removing redundant variables.

The first intervention consisted of introducing lags for the dependent variable, i.e. creating lag variables for the marginal price of electricity in Portugal. Lags 1, 3, 5, 7, 12 and 24 were considered in order to capture the temporal autocorrelation underlying the behavior of the time series. This approach aims to allow the models to learn dynamic relationships between the current value and recent history, which is essential in a market with high weekly regularity and short-term dependency.

At the same time, the composition of the set of variables was reassessed, based on an analysis of the relative importance assigned by a simple decision tree model. This analysis revealed the systematic irrelevance of the variables associated with wind direction, previously coded using one-hot encoding, whose marginal contribution to explaining the dependent variable was negligible. It was therefore decided to exclude them, promoting model parsimony and reducing the risk of introducing statistical noise.

As an additional boost to predictive capacity, variables with a rolling window structure were introduced, calculated from the moving averages of the marginal price and the standard deviations of the total load observed, considering windows of 24, 72 and 168 hours. This component aimed to provide the model with recent statistical memory, which is essential for capturing local fluctuation patterns, trend smoothing and episodes of instability.

Given that the models continued to be excessively sensitive to extreme values, a process of truncating the target variable (clipping) was adopted, limiting its values to the interval defined by the 5th and 95th percentiles of the empirical distribution. The main aim of this measure was to reduce the disproportionate impact of atypical observations on the loss function used during training, without compromising the structure of the sample.

The last intervention consisted of generating interaction variables between seasonal components (derived from trigonometric functions applied to the calendar) and critical energy variables. The aim was to capture multiplicative effects between the annual cycle and variables such as solar radiation or energy traded on the Iberian market, allowing the model to adjust the sensitivity of certain predictors to the temporal context in which they are inserted.

The sequential application of these transformations was reflected in a substantial improvement in performance indicators. As can be seen in tabel III, the LightGBM model now has an MAE of less than 6.40 €/MWh in the test set, making it the most effective of the models evaluated. XGBoost also showed consistent behavior, with an error of less than 8 €/MWh, while Extra Trees maintained values of more than 9 €/MWh, still reflecting some degree of overadjustment.

TABLE III-FINAL FORECASTING PERFORMANCE AFTER OVERFITTING MITIGATION

MODEL	MAE TRAIN(€/MWH)	MAE VAL(€/MWH)	MAE TEST(€/MWH)
EXTRA TREES	1.80	9.79	9.97
LIGHTGBM	3.96	6.12	6.39
XGBOOST	3.12	7.75	7.81

The methodological evolution adopted in this chapter leads us to conclude that the combination of time lags, rolling variables, clipping of the dependent variable and seasonal interactions contributes decisively to the robustness of electricity marginal price forecasting models, mitigating over-adjustment and promoting predictive stability over time.

4.4 Final Model: Exclusion of Exogenous Shocks and Temporal Enrichment

Considering the limitations identified in the predictive capacity in periods strongly influenced by extraordinary external factors not explicitly captured by the original set of explanatory variables, it was decided to restrict the time period of analysis to cover only the years 2023, 2024 and 2025. This restriction aims to minimize the distorting impact of external events such as the COVID-19 pandemic in 2020, which led to a drastic reduction in energy demand and, consequently, marginal prices. In addition, the subsequent energy crisis, triggered in 2021 and exacerbated in 2022 by geopolitical instability stemming from the war in Ukraine, caused a sharp and abnormal increase in prices, not reflected in the conventional variables. The exclusion of these years thus allows for a more robust and representative modeling of the structural and seasonal patterns of the market in more stable periods.

Given its proven effectiveness, the analysis focused exclusively on the LightGBM model. At this stage, several additional explanatory variables were incorporated into the model, with the explicit aim of improving predictive capacity by enriching the temporal structure and interaction between critical market variables.

Initially, the target variable, corresponding to the hourly marginal price of the Portuguese system with a time lag of 168 hours (one week), ensuring a forecast aligned with the required time horizon. At the same time, auxiliary variables were added with additional short-term lags (167 and 166 hours earlier), allowing the model to capture immediate sequential changes in the marginal price and recent trends.

In order to incorporate explicit seasonality into the model, cyclically based time variables were created using trigonometric sine and cosine functions applied to the months and days. These cyclical transformations allow for a continuous and adequate representation of the monthly and daily periodicity, avoiding distortions introduced by discrete categorical variables.

Next, several interaction variables were implemented through the cross-product between these cyclical components and key temporally shifted variables (168 hours), such as solar radiation, energy traded on the Iberian market and total observed load, to capture temporal effects conditioned by key energy variables.

To better capture short-term local dynamics, moving statistics (average, minimum, maximum and standard deviation) were calculated over 24-hour windows, both for the dependent variable and for time-shifted solar radiation. In addition, a relative normalization of solar radiation was introduced through its ratio to the recent maximum value in a 24-hour window, allowing the model to interpret relative variations in solar production.

In order to enhance the robustness of the model and mitigate the risk of overfitting, an early stopping mechanism was incorporated during the training process. This technique monitors the performance on a validation set and halts training if no improvement in the validation loss is observed over a predefined number of iterations. In this study, a patience value of 50 rounds was adopted, not arbitrary, but based on preliminary experiments which indicated that this threshold provided a balanced trade-off between training duration and generalization performance. By preventing unnecessary training beyond the point of diminishing returns, early stopping contributions to the selection of a model configuration that generalizes more effectively to unseen data.

Derived variables were also calculated, such as trends and recent absolute and percentage variations in the marginal price and total load, and quadratic terms and interactions between these variables were created to explicitly capture complex non-linear relationships.

Finally, using rigorous temporal validation - in which the model was trained exclusively with data from the year 2023, validated with data from 2024 and tested with data from 2025 - the predictive performance resulting from these extensive modifications was gauged. The results obtained show a substantial improvement in the out-of-sample prediction capacity, with the LightGBM model achieving a mean absolute error of 4.52 €/MWh in the training set, 5.54 €/MWh in the validation set and 5.29 €/MWh in the test set.

Given the initial objective, which stipulated a maximum MAE limit of 6 €/MWh in the test set, these final results show that the modifications and optimizations applied to the LightGBM model fully met the operational requirements, thus establishing this model as the final solution recommended for practical application in the business context.

5. Conclusion

The analysis of the results obtained for the hourly forecast of marginal electricity prices in the Portuguese market revealed a consistent and robust performance of the final model developed, fully in line with the initial objectives proposed in this study. The evaluation was structured around the Mean Absolute Error indicator, specifically aimed at providing a clear and easily interpretable metric for the quality of the forecasts made by the model.

For a detailed analysis, two representative weeks of performance were identified: the "best week", corresponding to the period in which the model showed the lowest mean absolute error, and the "worst week", reflecting the period with the highest mean absolute error, albeit low in absolute terms. This selection provided a comprehensive view of the model's predictive capacity in different contexts and market conditions.

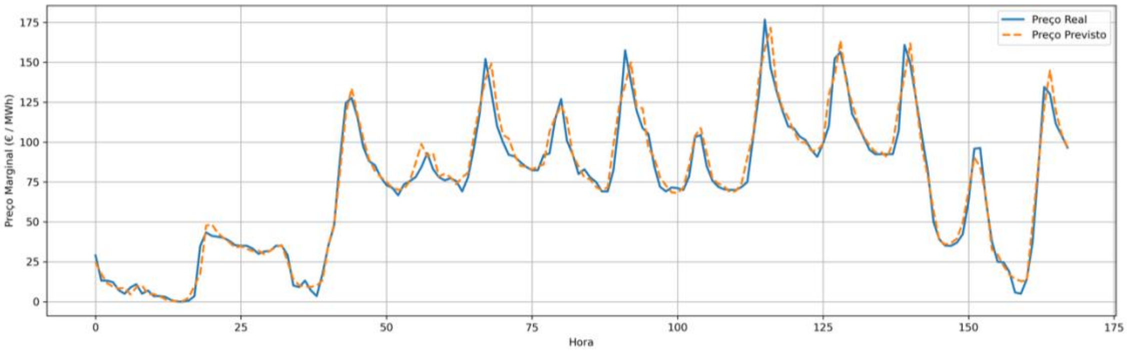

Figure 2- Forecast vs actual prices for the best-performing week

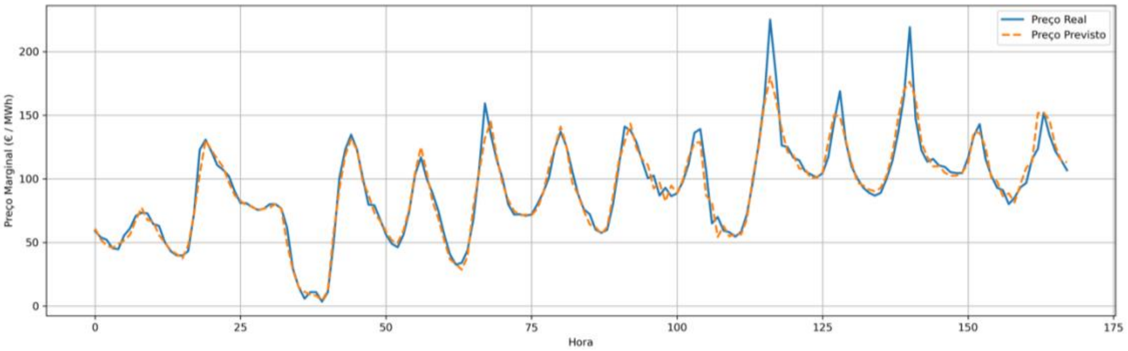

Figure 3- Forecast vs actual prices for the most challenging week

In the "best week", the model achieved a remarkable MAE of just 4.86 €/MWh, indicating high accuracy in predicting marginal prices. The visual analysis of the forecasting results presented in Figure 1 illustrates a notable alignment between actual and predicted electricity prices. While such alignment may superficially resemble patterns achievable by simpler seasonal models, the examined model distinguishes itself through its capacity to accurately anticipate sudden price spikes, abrupt declines, and irregular fluctuations inherent in electricity markets. Importantly, the model does not solely reproduce cyclical patterns but effectively captures nuanced variations triggered by rapid changes in demand, renewable energy supply fluctuations, and meteorological influences. These characteristics demonstrate the model's robustness and practical utility, suggesting its applicability for strategic decision-making and risk management in contexts characterized by inherent price volatility and uncertainty.

In the "most challenging week", despite being identified as the one with the lowest relative performance, the MAE remained extremely competitive, reaching 5.26 €/MWh, as shown in Figure 2. This particular week is characterized by greater volatility and some irregularities in the usual market patterns. However, it is noteworthy that, even in these challenging conditions, the model managed to preserve significant accuracy. When examining the actual and forecast price curves graphically, it can be seen that the model has maintained considerable consistency, with occasional divergences being limited to brief periods, typically associated with external events or unexpected spikes in the market.

This consistency of performance, even in periods considered more difficult, reinforces the robustness of the model, which is especially relevant for its practical application in business contexts where minimizing forecast error can directly translate into substantial financial benefits for small and medium-sized industrial companies, the target audience for this solution.

The use of the LightGBM model, optimized using advanced hyperparameter optimization techniques (using Optuna), proved to be a highly effective methodological decision. This model demonstrated an exceptional ability to deal with the high dimensionality and heterogeneity of the data set employed, which included meteorological variables, energy production by various sources, international interconnections, seasonal patterns and historical consumption. The success of this model can be attributed to its ability to automatically identify complex and non-linear relationships between these variables, allowing for a more refined forecast adjusted to real market conditions.

The pre-processing approach adopted, which involved robust normalization of numerical variables (Robust Scaler) and effective coding of categorical variables (One-Hot Encoding), also played a key role in the performance achieved. This methodology ensured adequate data preparation, reducing the impact of outliers and preserving the essential structural properties of the original data.

It is important to note that the development of the model considered the need for transparency and interpretability, critical aspects in industrial and regulatory applications. Although the traditional advantages of machine learning models are often accompanied by limitations in terms of interpretability, the use of decision tree-based methods, such as LightGBM, partially mitigates this challenge. By analyzing the importance of variables, obtained directly from the hierarchical structure of the trees, it is possible to identify which features are most decisive for price forecasting. This process provides clear visibility into the variables that most influence electricity price fluctuations, thus allowing for a better understanding of the critical factors underlying the dynamics of the energy market. In addition, the results presented here highlight the model's ability to generalize to different market conditions, while remaining robust to cyclical variations. This quality is particularly crucial in electricity markets characterized by high volatility, which often face significant exogenous shocks, such as regulatory changes or unexpected energy crises.

Finally, certain limitations have been identified that present opportunities for future improvements. For instance, the integration of hybrid models that combine machine learning approaches with econometric techniques or forecasting methods based on recurrent neural networks could be explored to achieve incremental enhancements in predictive performance. Moreover, extending the temporal window of the dataset, as well as incorporating advanced probabilistic forecasting techniques—such as conformal prediction or dynamic confidence intervals could represent significant methodological advancements in future studies.

It can thus be concluded that the model proposed in this master’s thesis successfully met the initially defined objectives, offering a robust and reliable predictive solution capable of delivering tangible benefits for the energy management of Portuguese industrial SMEs. This contributes to more informed, optimized, and economically advantageous decision-making within the energy market.

References

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Conejo, A. J., Contreras, J., Espínola, R., & Plazas, M. A. (2005). Forecasting electricity prices for a day-ahead pool-based electric energy market. *International Journal of Forecasting, 21*(3), 435–462. https://doi.org/10.1016/j.ijforecast.2004.12.005

Deng, S., Su, J., Zhu, Y., Yu, Y., & Xiao, C. (2024). Forecasting carbon price trends based on an interpretable light gradient boosting machine and Bayesian optimization. *Expert Systems with Applications, 242*, 122502. https://doi.org/10.1016/j.eswa.2023.122502

Ghelasi, P., & Ziel, F. (2024). From day-ahead to mid and long-term horizons with econometric electricity price forecasting models. Renewable and Sustainable Energy Reviews, 217, 115684. https://doi.org/10.1016/j.rser.2025.115684

Lago, J., Marcjasz, G., De Schutter, B., & Weron, R. (2021). Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy, 293*, 116983. https://doi.org/10.1016/j.apenergy.2021.116983

Nowotarski, J., & Weron, R. (2018). Recent advances in electricity price forecasting: A review of probabilistic forecasting. European Journal of Operational Research, 261(2), 533–547. https://doi.org/10.1016/j.rser.2017.05.234

Rubattu, N., Maroni, G., & Corani, G. (2023). Electricity load and peak forecasting: Feature engineering, probabilistic LightGBM and temporal hierarchies. In Lecture Notes in Computer Science (Vol. 14271, pp. 303–318). Springer. https://doi.org/10.1007/978-3-031-49896-1_18

Taieb, S. B., & Hyndman, R. J. (2021). Explainable machine learning for time series forecasting. Expert Systems with Applications, 182, 115179. https://doi.org/10.1016/j.eswa.2021.115179

Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting, 30*(4), 1030–1081. https://doi.org/10.1016/j.ijforecast.2014.08.008

Zamudio López, M., Zareipour, H., & Quashie, M. (2024). Forecasting the occurrence of electricity price spikes: A statistical-economic investigation study. *Forecasting, 6*(1), 7. https://doi.org/10.3390/forecast6010007

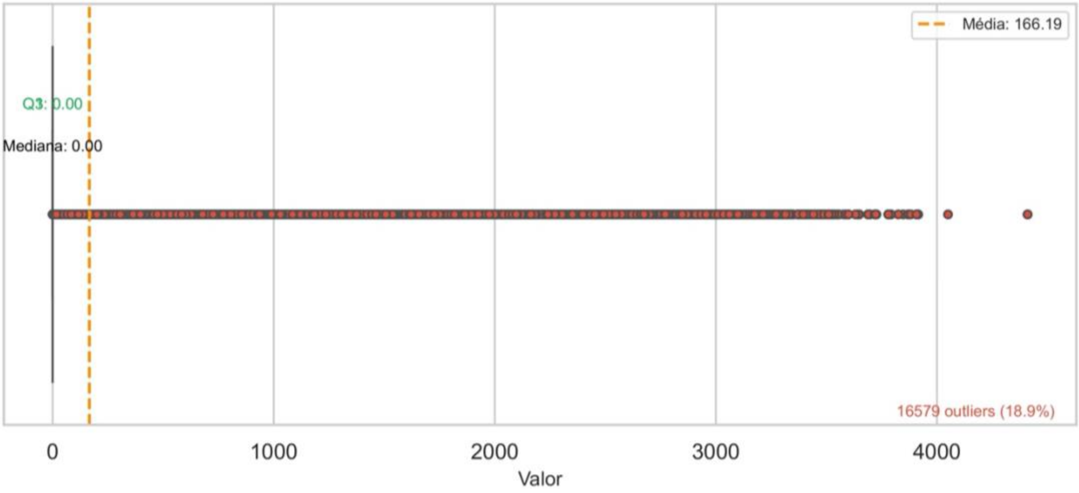

Figure 4 - Boxplot of Electricity Imports from Portugal to Spain

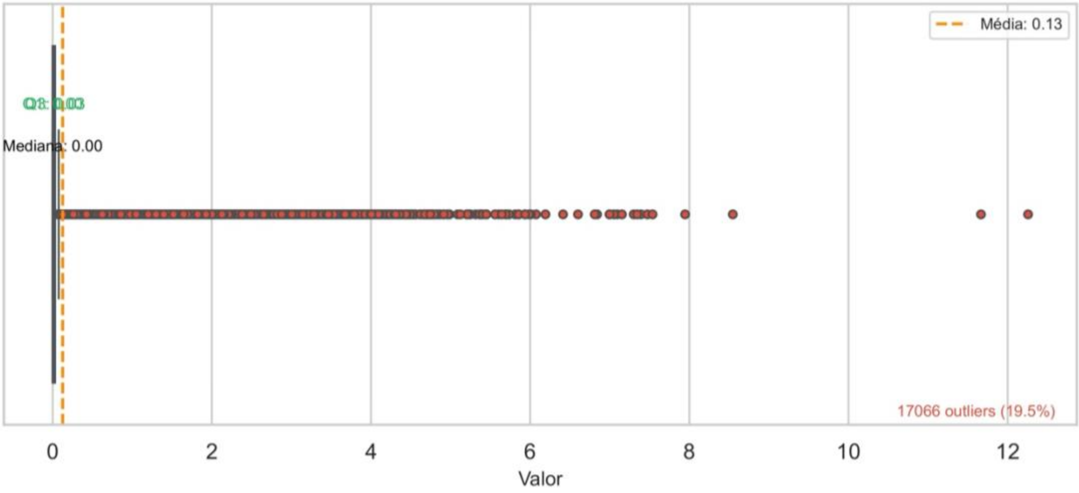

Figure 5-Boxplot of Precipitation (mm) – Peso da Régua

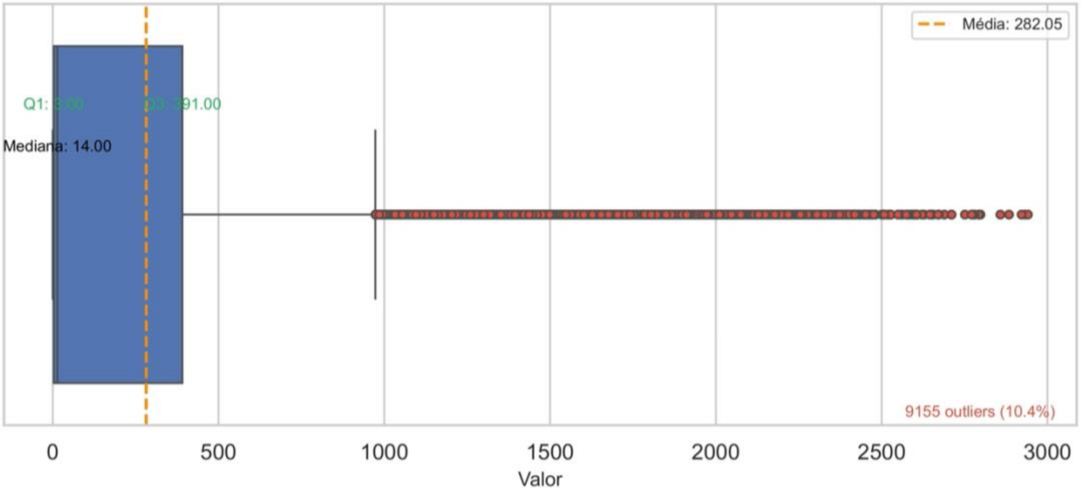

Figure 6-Boxplot of Actual Consumption in Hydro Pumped Storage (MW)

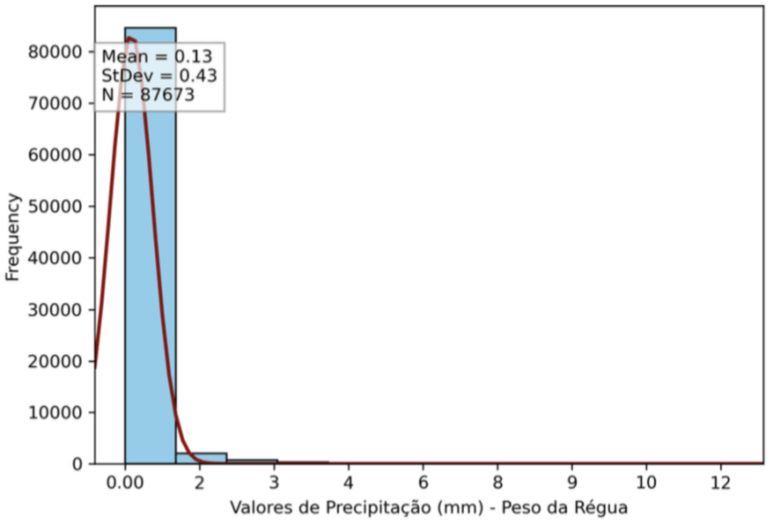

Figure 7-Histogram of Precipitation (mm) – Peso da Régua

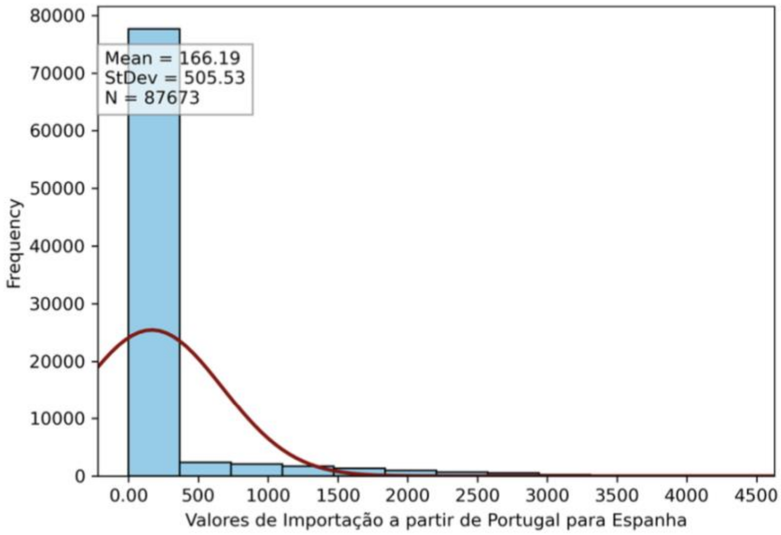

Figure 8-Histogram of Electricity Exports from Spain to Portugal

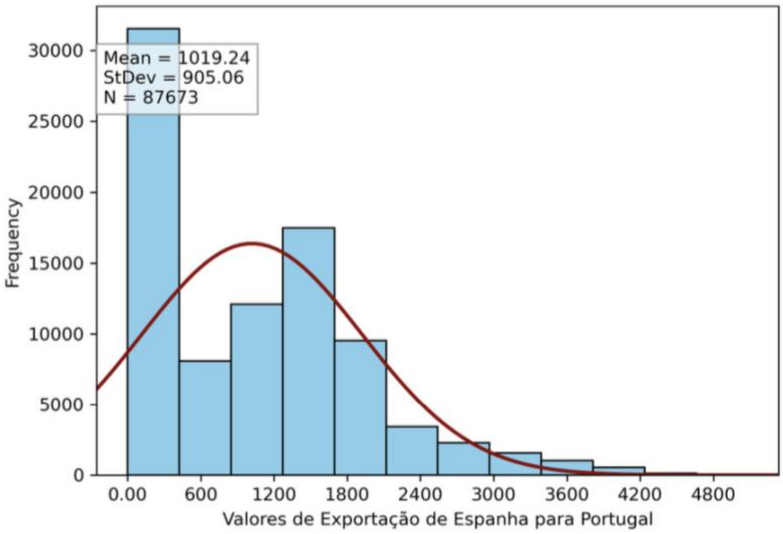

Figure 9-Histogram of Electricity Imports from Portugal to Spain

TABLE IV- LIST OF NUMERICAL VARIABLES USED ON ORIGINAL DATASET

NUMERICAL VARIABLES
SOLAR GENERATION [MW] - DAY AHEAD (PORTUGAL)
WIND ONSHORE GENERATION [MW] - DAY AHEAD (PORTUGAL)
BIOMASS - ACTUAL AGGREGATED [MW]
FOSSIL GAS - ACTUAL AGGREGATED [MW]
FOSSIL HARD COAL - ACTUAL AGGREGATED [MW]
HYDRO PUMPED STORAGE - ACTUAL AGGREGATED [MW]
HYDRO PUMPED STORAGE - ACTUAL CONSUMPTION [MW]
HYDRO RUN-OF-RIVER AND POUNDAGE - ACTUAL AGGREGATED [MW]
HYDRO WATER RESERVOIR - ACTUAL AGGREGATED [MW]
OTHER SOURCES - ACTUAL AGGREGATED [MW]
SOLAR - ACTUAL AGGREGATED [MW]
WIND ONSHORE - ACTUAL AGGREGATED [MW]
TOTAL LOAD FORECAST [MW] - DAY AHEAD (PORTUGAL)
ACTUAL TOTAL LOAD [MW]
WIND SPEED (KM/H)
WIND DIRECTION (°)
SOLAR RADIATION (W/M^2)
PRECIPITATION (MM) - PESO DA RÉGUA
LISBON TEMPERATURE (°C)
PORTO TEMPERATURE (°C)
ÉVORA TEMPERATURE (°C)
AVERAGE TEMPERATURE (°C) - LISBON, PORTO, ÉVORA
MARGINAL PRICES - SPANISH SYSTEM
MARGINAL PRICES - PORTUGUESE SYSTEM
ENERGY TRADED - DAY-AHEAD MARKET

TOTAL MATCHED ENERGY PURCHASED - SPANISH SYSTEM
TOTAL MATCHED ENERGY SOLD - SPANISH SYSTEM
TOTAL MATCHED ENERGY PURCHASED - PORTUGUESE SYSTEM
TOTAL MATCHED ENERGY SOLD - PORTUGUESE SYSTEM
IMPORTS FROM PORTUGAL TO SPAIN
EXPORTS FROM SPAIN TO PORTUGAL
IBERIAN MARKET ENERGY (INCLUDING BILATERAL TRADES)

TABLE V-LIST OF CATEGORICAL VARIABLES USED ON ORIGINAL DATASET

CATEGORICAL VARIABLES
HOUR
DIRECTION OF THE WIND
DAY OF THE WEEK
WEEKDAY OR WEEKEND
HOLIDAY INDICATOR