UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Industrial

Deterministic forecasting of short-term imbalance in the Hungarian electric power system using Maching Learning

Trabajo Fin de Grado

Grado Universitario en Ingeniería Industrial-Grau Universitari en Enginyeria Industrial

AUTOR: Guillén Marzo, Óscar

Tutores: Tormos Juan, María Pilar; Markovics, Dávid

CURSO ACADÉMICO: 2023/2024

Department of Energy Engineering
Faculty of Mechanical Engineering
Budapest University of Technology and Economics

Deterministic Forecasting of Short-term Imbalance in the Hungarian Electric Power System Using Machine Learning

Degree in Industrial Technology Engineering

Author: Óscar Guillén Marzo
Tutor: Dávid Markovics
Course: 2023-2024

ABSTRACT

The increasing development and integration of renewable energy sources into the electrical system present significant challenges in ensuring a constant balance between energy generation and demand. Unlike conventional energy sources, renewable energies, such as solar and wind power, are inherently intermittent and not always available. This variability in energy generation complicates the task of maintaining the stability and reliability of the electrical system, and therefore, short-term imbalance forecasting becomes crucial to achieve this goal.

This thesis project focuses on the application of Machine Learning techniques in the electrical system of Hungary, with the main objective of improving the ability to forecast these imbalances. Although various algorithms will be explored, the research will focus on LightGBM, explaining its characteristics in detail and evaluating its performance using appropriate metrics. The results obtained from this study will provide valuable information for grid operators and regulatory entities, allowing them to improve efficiency in managing the electrical system imbalance, resulting in reduced operational costs and optimized energy planning.

Keywords: renewable energy; electrical system; energy balance; Machine Learning; Hungary; LightGBM; grid operators; forecasting methods.

ACKNOWLEDGMENTS

First of all, I would like to thank my family, especially my parents, who support me in everything I do and gave me the opportunity to complete my studies in Hungary. Secondly, I am grateful for the indications and advice of Jose Javier López Sánchez (UPV professor), for the selection of the topic of my thesis. Thanks to him I have discovered that I would like to continue researching Machine Learning and data analysis. Last but not least, I would like to acknowledge the explanations and guidelines that my thesis tutor Dávid Markovics has provided me. He introduced me to the energy market and Machine Learning, explaining to me some key concepts. Moreover, he was always willing to help and answer any questions I had.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
ARIMA	AutoRegressive Integrated Moving Average
aFRR	automatic Frequency Restoration Reserve
BRP	Balance Responsible Party
CatBoost	Categorical Boosting
EFB	Exclusive Feature Bundling
FCR	Frequency Containment Reserve
GA	Genetic Algorithm
GBM	Gradient Boosting Machine
GOSS	Gradient-based One-Side Sampling
IGCC	International Grid Control Cooperation
MAE	Mean Absolute Error
MARI	Manually Activated Reserves Initiative
MAVIR	Hungarian Independent Transmission Operator Company Ltd.
mFRR	manual Frequency Restoration Reserve
MWh	Megawatt-hour
nMAE	normalized Mean Absolute Error
PICASSO	Platform for the International Coordination of Automated Frequency Restoration and Stable System Operation
RMSE	Root Mean Square Error
TPE	Tree-structured Parzen Estimator
TSCV	Time Series Cross-Validation
TSO	Transmission System Operator
XGBoost	eXtreme Gradient Boosting

CHAPTER 1

INTRODUCTION

Electricity is the cornerstone of our modern society, powering homes, businesses, and factories, and achieving more importance as technology develops. Nevertheless, since electric energy can't be stored through a simple method, it must be generated and consumed at the same time and, therefore, it requires a complex system to manage it.

The main objective of the electric power system is to match the quantity of supplied energy with the demanded one, ensuring that way the stability and reliability of the grid and maintaining it within the frequency limits. However, this is not an easy task.

Over the last decades, several initiatives have moved many countries to modernise their electricity grids, incorporating renewable energy sources. This behaviour is crucial to take steps to fight against climate change and will help to meet the objectives of the Paris Agreement [1] and the EU Renewable Energy Directive [2], which aim to increase the renewable energy in the system. But on the other hand, the incorporation of this kind of sources implies a challenge to grid operators. Unlike traditional energy sources such as coal, petroleum, or natural gas, renewables are of variable nature and most of the time unpredictable (especially solar and wind power), complicating the task of maintaining the balance on the grid.

These imbalances can lead to several impacts, including grid instability, power outages, increased electricity prices, and inefficiency in energy planning. For that reason, short-term forecasting has emerged as an excellent tool to deal with this problem and provide so valuable information.

The primary objective of this research is to develop a predictive model using Machine Learning techniques to improve the accuracy of short-term imbalance forecasts in the Hungarian electrical system. While various algorithms will be evaluated, this research focuses particularly on LightGBM, a gradient boosting algorithm known for its speed and performance with large datasets.

The study aims to uncover the complex relationships influencing energy imbalances by using historical data from MAVIR (the Hungarian TSO), such as historical imbalances, power generation differentiating among different sources (with an emphasis on renewables), energy demand and its forecasted values, energy prices in the day-ahead market, and the imported-exported balance of energy, among others.

The proposed LighGBM model provides a deterministic forecast to predict the average of the next quarter-hour imbalance. Within its configuration, we can highlight the Time Series Cross-Validation (TSCV) method, which ensures that the model is trained in the correct chronological order, and the randomized search, to explore a huge amount of hyperparameter

combinations and find the one that works the best. Moreover, evaluation metrics such as the nMAE or the variance ratio will be used to assess its accuracy.

To sum up, this research will provide meaningful information to TSOs and BRPs about the power system imbalances, that can be used to optimize the energy planning. And, at the same time, it will prove the potential of Machine Learning techniques to obtain information in complex research fields.

The algorithm performs a deterministic forecast to predict the average of the next quarter-hour imbalance. Using 22 months of historical data from MAVIR to train and test the model. The objective of the study is to prove that GBM techniques can provide valuable information that can be used by TSOs and BRPs.

The structure of this thesis is as follows:

- Chapter 2 Theoretical Framework: Explains in detail all the necessary concepts of energy systems to fully understand this work.
- Chapter 3 Literature Review: Examines previous research on imbalance forecasting and the application of Machine Learning in energy system imbalances.
- Chapter 4 Methodology: Describes the data analysis process and preprocessing steps, also explaining Machine Learning in general and LightGBM in particular.
- Chapter 5 Experimental Setup: Details the design of experiments such as the hyperparameter tunning and the evaluation metrics.
- Chapter 6 Results: Presents the comparative performance of different models and discusses the findings.
- Chapter 7 Conclusion and further steps: Summarizes the key findings, contributions, and suggests directions for future research.

CHAPTER 2

THEORETICAL FRAMEWORK

2.1 Power System Imbalances Effects

Electrical power system mismatches might cause fluctuations in system parameters such as voltage and frequency (Figure 1), which could cause adverse impacts. These effects include:

Grid instability and unreliability: That could lead to power outages, disrupting both residential and commercial activities. Producing that way financial losses for business due to the machinery damage and decrease of productivity. In addition, this consistent power supply could be especially critical to hospitals and emergency services.

Higher electricity prices: Managing expensive resources such as backup power plants or energy storage systems can affect electricity prices for consumers. Moreover, imbalances can also lead to volatile energy prices.

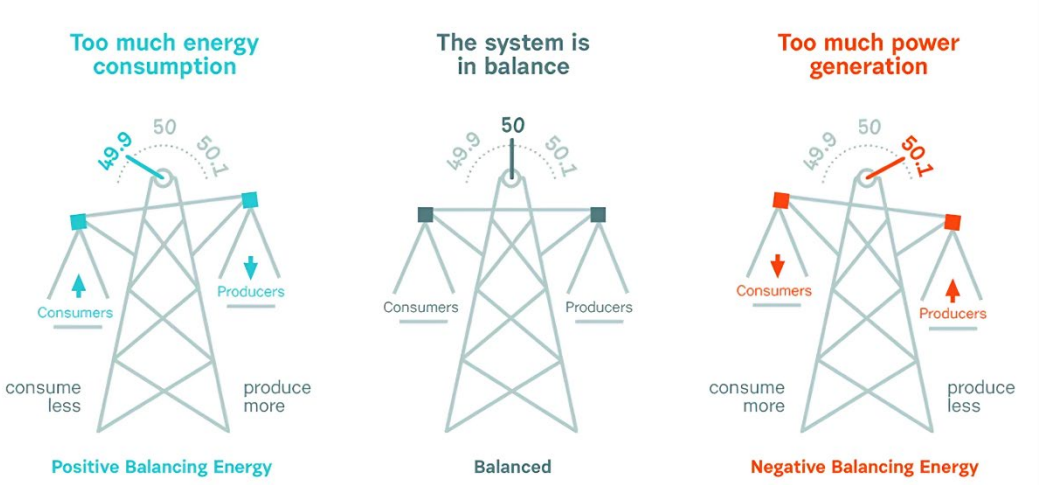

Figure 1: How the frequency varies with the energy imbalances [1]

On the other hand, if these grid fluctuations are correctly managed, good outcomes can be achieved such as:

Energy Efficiency and Sustainability: Effective management of power system imbalances facilitates the integration of renewable energy sources, reducing reliance on fossil fuels and contributing to sustainability goals.

After the presentation of the previous effects, it's clear that the significance of these imbalances cannot be overstated under any circumstance and must be correctly managed.

2.2 Energy Reserves and Cooperations

The rectification of these fluctuations needs to be quantified to take the proper solution. This process involves a combination of real-time monitoring, forecasting techniques, and mathematical modelling. Then, [3] in order to manage these mismatches, TSOs (Transmission System Operators) will activate different types of energy reserves, which are responsible for maintaining the grid frequency within acceptable limits and can be classified into three types:

Frequency Containment Reserve (FCR) is an example of a primary reserve. It acts as the first measure to compensate for deviations since can stabilize the frequency within seconds. However, it is only used for the initial stabilization of the grid.

Some common FCR providers are hydropower plants and battery generators. Which are continuously monitoring grid frequency to reset the balance as quickly as possible. That immediacy has as a cost a higher price than other reserve types.

Automatic Frequency Restoration Reserve (aFRR) is a type of secondary reserve. It's activated by TSOs based on pre-defined frequency limits and restores the balance of the grid. Its time response is slower than the FCR, responding within minutes (5 or 7.5 minutes).

These reserves are typically provided by pumped-storage power plants or gas turbines, which can be quickly activated and used for a longer duration. Besides that, they are less expensive than the primary ones.

Manual Frequency Restoration Reserve (mFRR) belongs to the tertiary reserves category. These reserves must be activated within 12.5 and 15 minutes. As its name indicates they are always triggered manually by TSOs, who are the ones responsible for informing the operator when the device must be turned on. Since they have the lowest time-response they are the least expensive reserve.

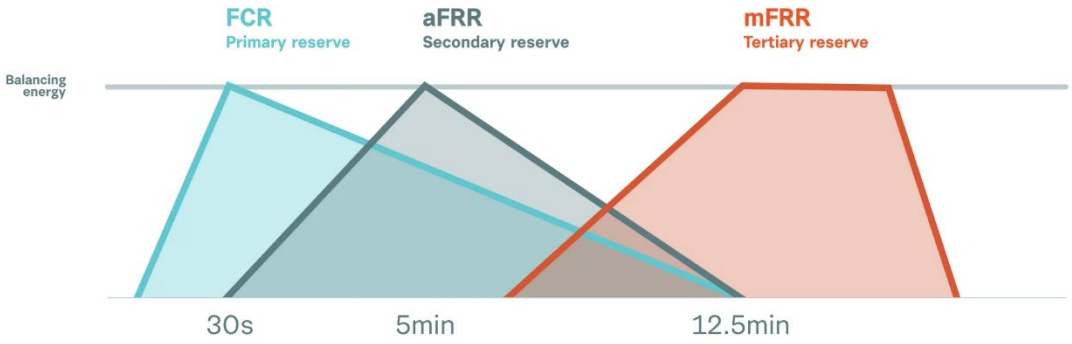

Figure 2: Activation time on the different reserves. [4]

Furthermore, the stability and reliability of power systems could be reinforced by the collaboration among grid operators across different regions. This procedure receives the name of **FRR cooperations**.

These alliances ensure the efficient utilization of frequency restoration reserves among the interconnected power systems by sharing reserves and coordinating response actions to restore the grid frequency.

Several initiatives, such as MARI (Manually Activated Reserves Initiative) [5], PICASSO (Platform for the International Coordination of Automated Frequency Restoration and Stable System Operation) [6] and IGCC (International Grid Control Cooperation) [7] coordinate several European grid operators. They aim to establish common standards and practices for frequency restoration reserve management.

To better illustrate the concept of FRR cooperations, IGCC is going to be explained as an example:

The International Grid Control Cooperation has twenty-one operational TSOs across Europe (including Hungary) [7] which are physically connected to perform the imbalance netting process.

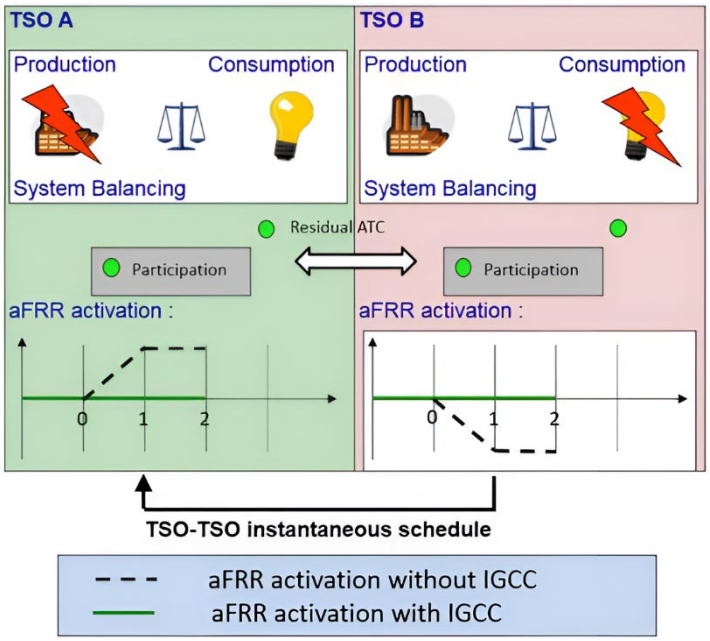

Figure 3: Imbalance netting in IGCC [7]

The imbalance netting process consists of continuous communication among TSOs in order to avoid different TSOs utilizing frequency restoration reserves (FRR) in opposite directions. Their objective is to use their respective frequency restoration control errors, compensating each other their grid imbalances and optimising the amount of FRR used.

The employment of these reserves is not free. it has a particular operational cost. This expense is determined by the energy bought on the markets, particularly on the day-ahead and intraday markets.

2.3 Day-ahead and Intraday Markets

[8] Given the absence of a system to easily store large quantities of electricity and considering the need to balance the electricity grid, energy must be traded daily.

To execute these electricity transactions, two kinds of markets are employed:

Day-ahead market: In this market, participants submit their bids and offers for electricity for the following day, based on their anticipated energy needs and generation capacity. In addition, are the bids and offers the ones that establish the market-clearing price, as shown in Figure 4. That way, market participants optimize their operations and protect against

volatility. However, real-time deviations from these forecasts are inevitable, necessitating the use of the intraday market, which will be explained below.

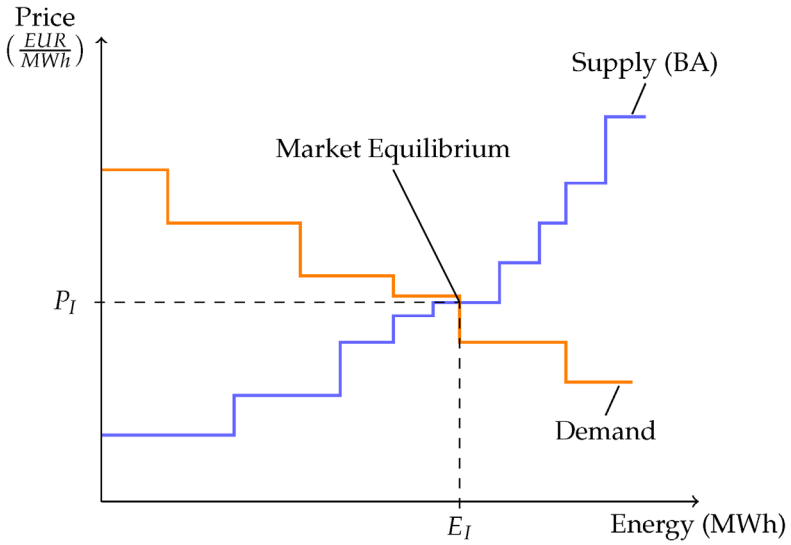

Figure 4: Intersection of the supply (blue) and demand (orange) [9]

Figure 4 shows all the bids ordered from higher to lower in colour orange and all the offers ordered from lower to higher coloured in blue. The intersection between them indicates the price of energy in MWh for the following day.

Intraday market: In contrast with the previous market, this is a short-term one; it allows trading closer to real-time. Its target is to correct any unexpected variations in demand or supply that may arise along the day, which are not covered by the energy bought on the day-ahead market.

These two markets are necessary for grid stability. Their combination makes it possible to plan the energy bought and correct it if there's a mismatch between supply and demand.

2.4 TSOs and BRPs Roles

As will be explained in the data analysis section, there exists a tendency toward negative imbalances in the data. Far from being a coincidence, this trend is logically explicable. To do it, it's essential to comprehend the role of the BRPs (Balance Responsible Parties), which are typically entities such as power generators, suppliers, retailers, or large consumers of electricity.

Electricity suppliers and consumers set contractual agreements specifying the quantity of electricity to be bought and sold. Nevertheless, as has been explained before, the actual amounts differ from the scheduled production or consumption values and such variations must be managed to maintain grid balance. In this process, grid operators apply a balancing responsibility system, which makes all market participants responsible for the imbalances they generate.

The TSO imposes an imbalance charge on BRPs to cover the costs of balancing the system using power reserves. This mechanism incentivizes the market to reduce imbalances and transfer the financial risk to the BRPs.

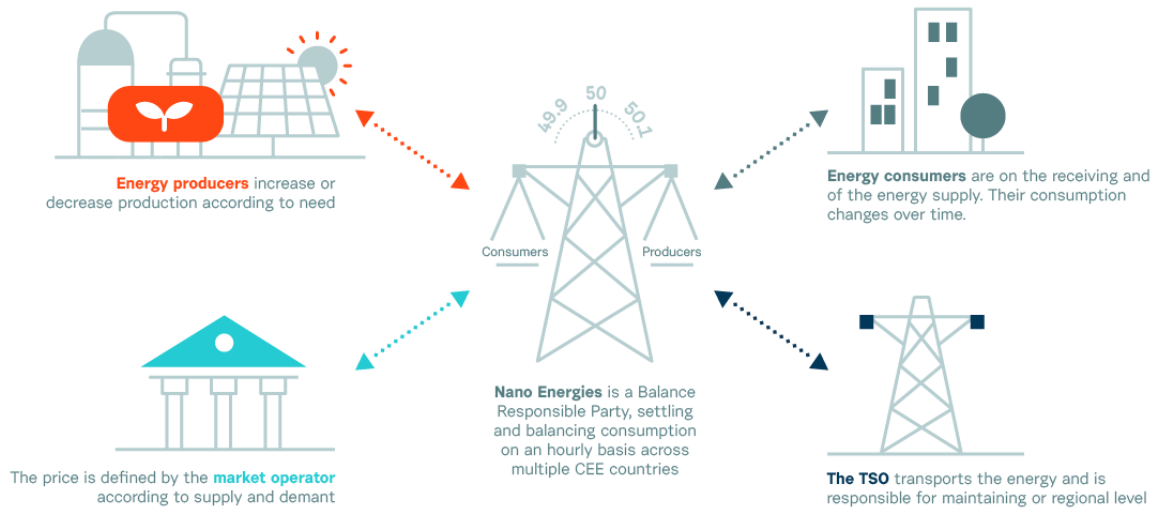

Figure 5:Example of how the BRPs work. [10]

As it has been just explained, BRPs are penalized for the difference between power generated and consumed. However, the amount of the penalization depends on the quantity differed and if the deviation is upwards or downwards (positive or negative imbalance).

Although in both cases an external entity is needed to achieve balance, positive variations have a greater charge than negative ones. The explanation lies in the fact that for correcting positive imbalances, power must be generated, using fuel (natural gas, oil, or coal) and machinery costs (motor oil, soft water, maintenance…). So, the price of producing power is, a priori, higher than the price of not producing it.

For that reason, BRPs are completely rational when they try to avoid causing upward control, causing, that way, a negative preference in the imbalance data.

2.5 Short-Term Imbalance Forecast on Power Systems

As the imbalances of the electric power system can fluctuate within relatively short periods (from minutes to several hours), it will be very useful for grid operators to predict potential changes and anticipate them to avoid too large frequency variations. That's just what the short-term imbalance forecast can provide.

With this tool, operators are provided with information that helps them anticipate fluctuations in the grid by managing the activity of power plants and optimizing the utilization of energy storage systems to ensure that supply matches demand in real-time. Furthermore, due to the growth of renewable energy sources and their variable nature, this kind of forecast has a key role in integrating it into the grid.

CHAPTER 3

LITERATURE REVIEW

In their study, Garcia and Kirschen [11] expose the negative aspects of simplistic methods such as ARIMA (Auto-Regressive Integrated Moving Average) and propose the incorporation of ANN (Artificial Neural Network) to uncover the non-linear and irregular patterns within the data, enabling precise forecasting of daily imbalance medians. The combination of those two models led to more accurate predictions than conventional forecasting methods.

Kratochvil [12] used ARIMA as a model, in which five sections were assigned to the system imbalance in a way that the problem was reduced to a classification. We can outstand from this study that the autocorrelation of the system imbalance shows that the data earlier than two hours was not too useful to do short-term forecasts.

In his master thesis [13], Contreras used GA-optimized Random Forest to make hourly imbalance predictions to build a bidding program that minimizes the imbalance cost. This research has shown that advanced modelling techniques are completely feasible for this purpose, their results are enough valuable to justify their deployment.

In their study [14], Salem et al. used quantile regression forests to predict imbalances in the power system. This method was performed for two-hour timeframes on a 2-year dataset, achieving meaningful improvements compared to using six- or less-month datasets. They proved that their method was more accurate than the ones used by the TSOs at that time.

Dumas et al. [15] combine volume imbalance forecasts with reserve costs. They use a two-step approach, starting with the computation of probabilities for system imbalances and then based on that make predictions regarding the imbalance prices.

In their work [16], Bottieau et al. proved that machine learning techniques were better than conventional methods by using a one-step-ahead forecasting model for system imbalance.

Rojas et al. [17] compared three algorithms for two different lead times. The result was that the random forest algorithm was the most accurate and computationally more efficient compared to the linear regression and standard neural network models, especially in short-term forecasts (15 minutes ahead).

Thanks to all those investigations, it can be concluded that the random forest model stands out as one of the most effective methods for forecasting power system imbalances. Furthermore, optimal results can be achieved by using more than one-year of historical data and targeting the prediction of short-term fluctuations.

In this research the LightGBM algorithm will be examined in detail and tested in several conditions, also being compared to other models to prove its superiority.

CHAPTER 4

METHODOLOGY

4.1 Data Analysis

In this section of the document, a detailed analysis will be conducted on the datasets from MAVIR (the Hungarian TSO) [18], intended for forecasting short-term imbalances. Three datasets are available:

"**Imbalance forecasting main**": which includes data about the historical imbalances, solar power prognosis, prevision of electricity demand, the real-time imbalance of IGCC, and day-ahead market prices.

"**Energy mix real-time data**": which includes parameters such as the gross system load, the net actual generation sum, and the power generated by the most important energy sources (Wind Onshore, Solar, Hydro Run-of-river and poundage, Nuclear, Fossil Gas…)

"**Import-export**": includes only two parameters, the actual and the planned import-export balances.

All these datasets contain data from *2022-03-17 17:45:00* to *2024-01-25 05:15:00*, providing approximately **1 year and 10 months of data**.

On the other hand, the parameter that is set as the target of the forecast is the one called "**IBT15-30**", which is the average of three parameters: "Imbalance T20", "ImbalanceT25" and "Imbalance T30".

After an extensive investigation of the data, this study presents the major insights, relations, and patterns that have been uncovered among these datasets. This analysis will start with the parameter Imbalance T0 (power imbalance at the time of doing the prediction). Since it has a strong relation with the target parameter.

Imbalance T0:

Indicates de imbalance between demand and supply in MW. The X indicates the difference between the time of the index and the time in which the value was measured. 26 parameters of this kind are provided, from "Imbalance T0" to "Imbalance T-125" with 5 minutes of step between parameters.

Let’s understand the nature of these parameters through some graphs:

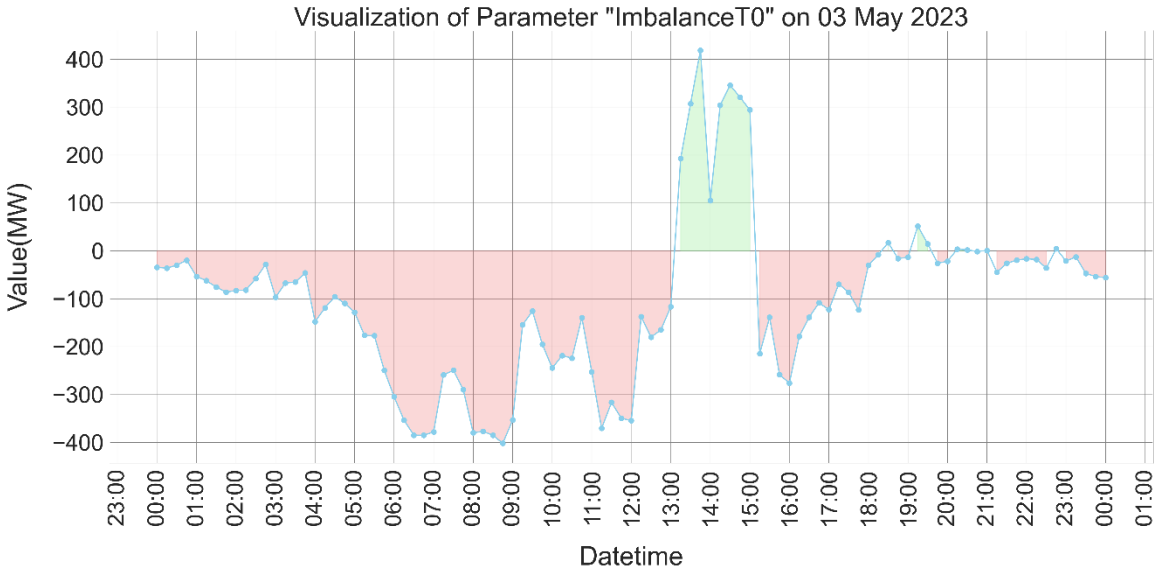

Figure 6: Graph of “Imbalance T0” along a day.

Figure 6 illustrates the evolution of the imbalances during a particular day of the data, in which a dominance of negative values can be clearly appreciated. These fluctuations can be either positive (generation is bigger than consumption) or negative (generation is smaller than consumption) but they are rarely zero (generation is equal to consumption).

The following histograms will reveal if this negative tendency belongs to that day or is a data general trait.

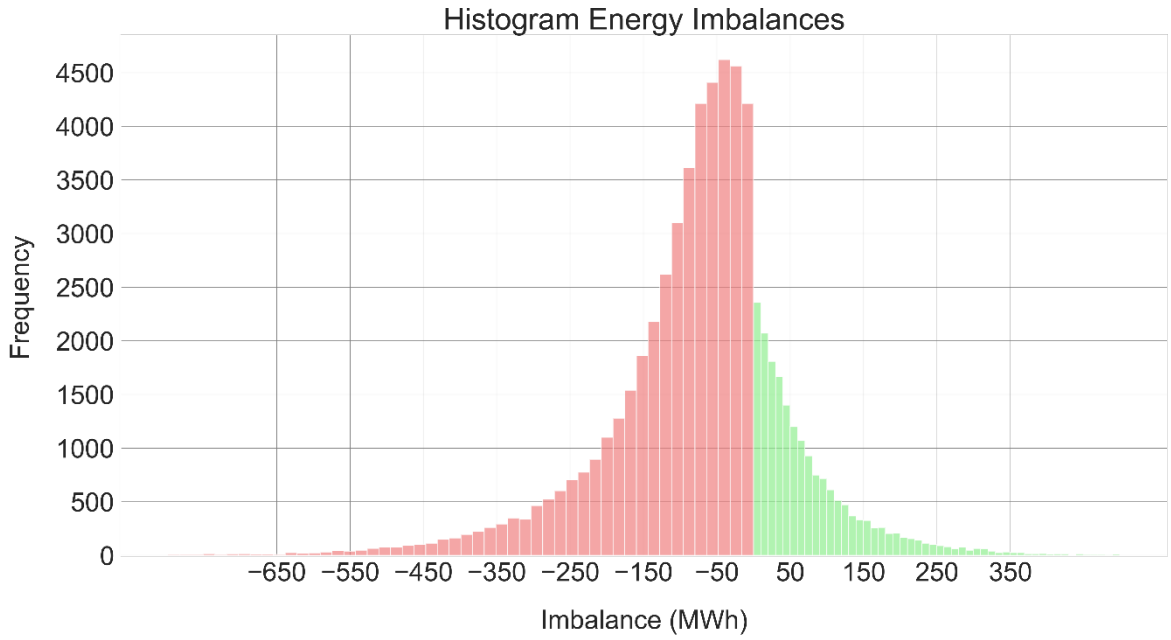

Figure 7: Histogram of "Imbalance T0" (all data)

If the analysis is done for each month, the next histograms are obtained.

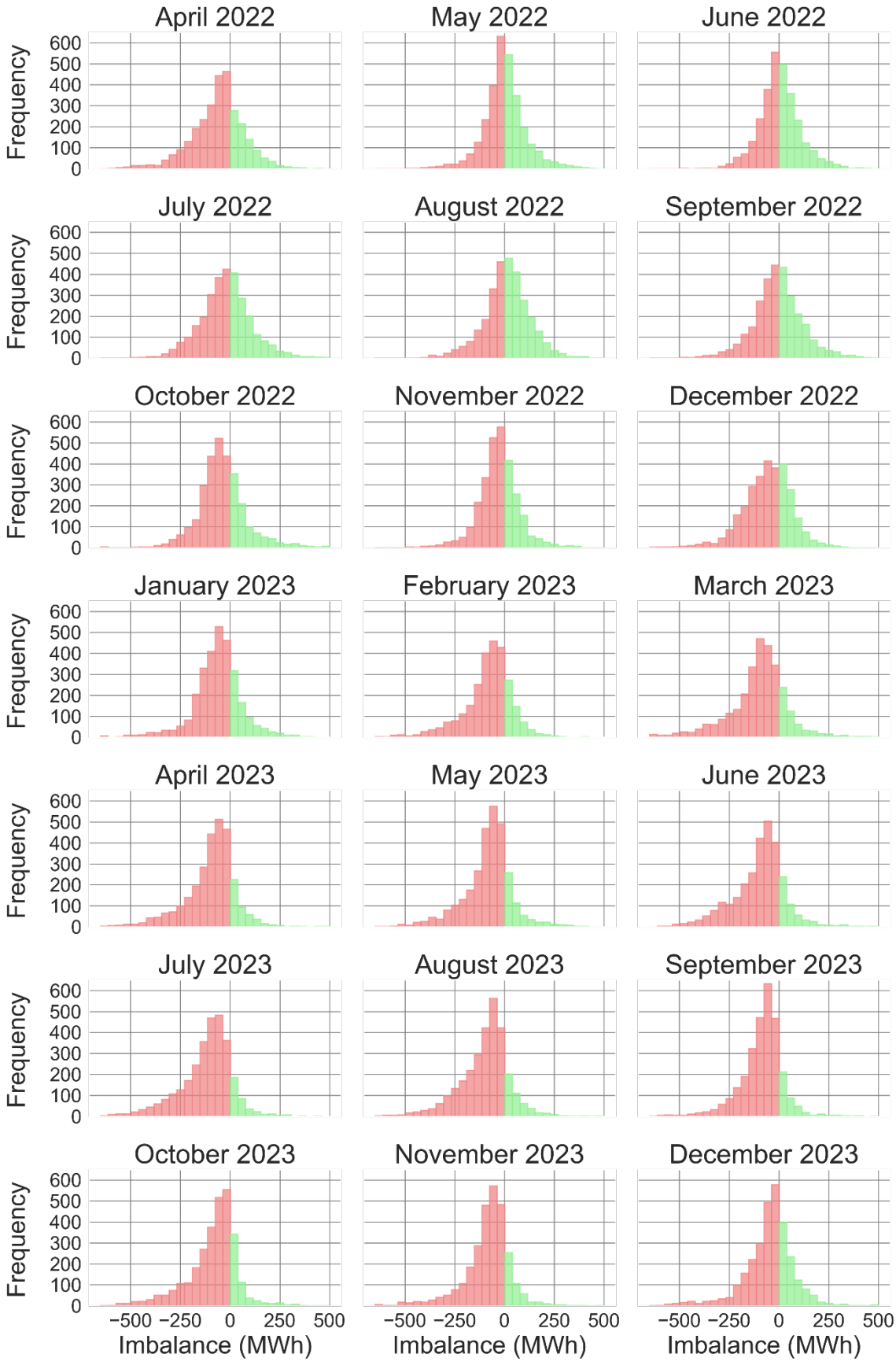

Figure 8: Histogram comparison of "Imbalance T0" for each month.

These histograms show that that negative tendency is not a coincidence, most months present it. Now can be conclusively stated that there exists a tendency toward negative imbalances in the data.

As explained in the theoretical framework section, this phenomenon is due to the charges that impose TSOs on BRPs, which are higher when the control is done upwards. So is completely reasonable to find this tendency in the data.

Once the nature of the imbalance has been uncovered, let's explore if there are parameters that have a strong correlation with the target one (IBT15-30), with a correlation matrix.

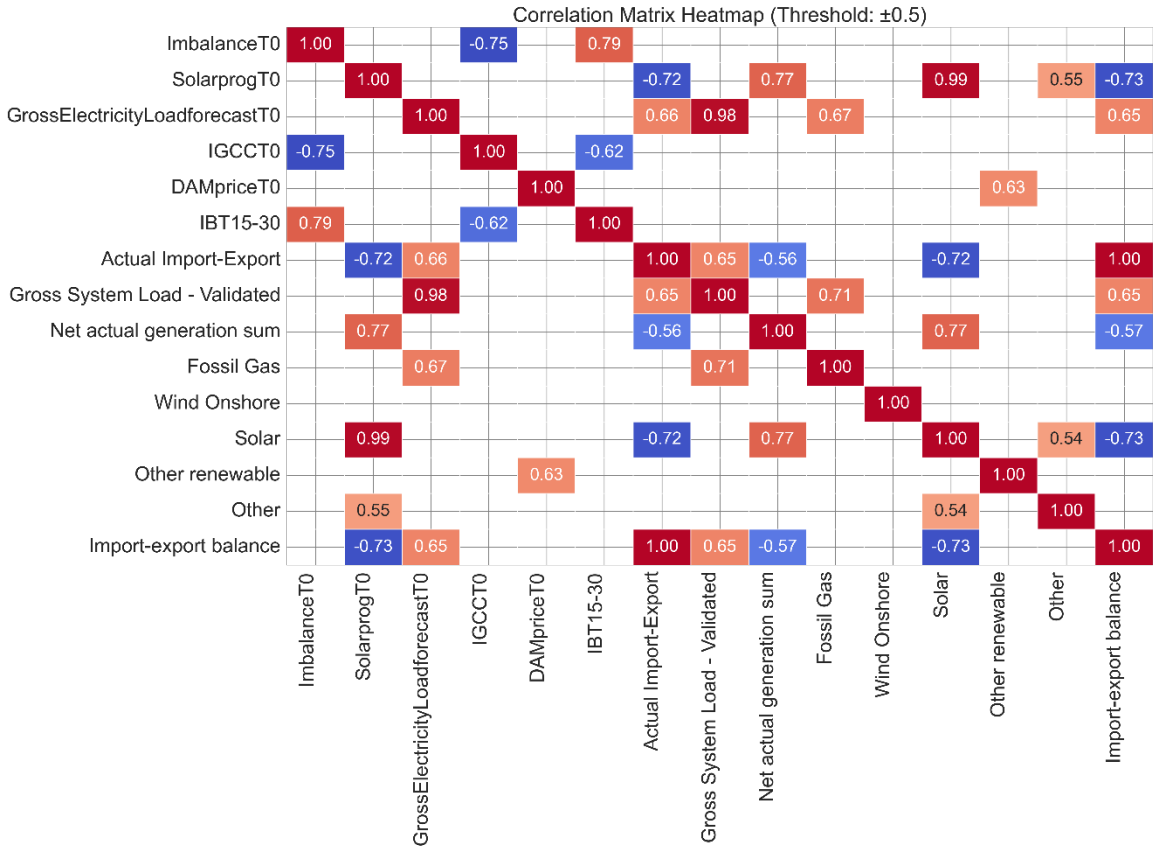

Figure 9: Correlation matrix.

This matrix only shows the parameters that show correlations with other ones, it has been filtered to increase the clarity of the graph. We can observe that the target parameter correlates with two parameters "ImbalanceT0" and "IGCCT0", both measure imbalances in real-time. Let's explore them deeper:

ImbalanceT0: Present a correlation of "0.79". Its strong positive relation was to be expected since the target is the evolution of this parameter over time and, as can be seen in Figure 6, it follows tendencies most of the time. This correlation is clearly represented in the following scatter diagram:

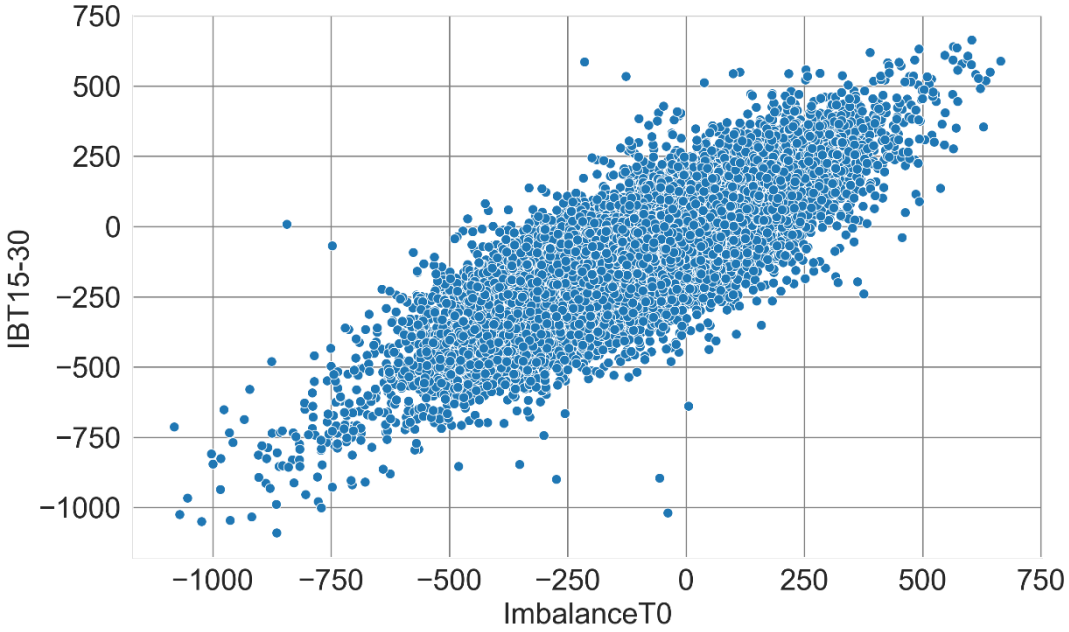

Figure 10: Scatter Diagram: ImbalanceT0 vs IBT15-30.

IGCCT0: As it has been explained in section 2.4, IGCC is a cooperation that aims to coordinate efforts among several TSOs from several European countries, in order to manage the grid imbalances. So, this parameter shows how much aFRR demand has been netted between the cooperating partners.

The value of the matrix that relates IGCCT0 with the target parameter is negative: "-0.62". This is due to the fact that IGCC has a different sign convention for practical reasons.

This is how its scatter diagram looks like:

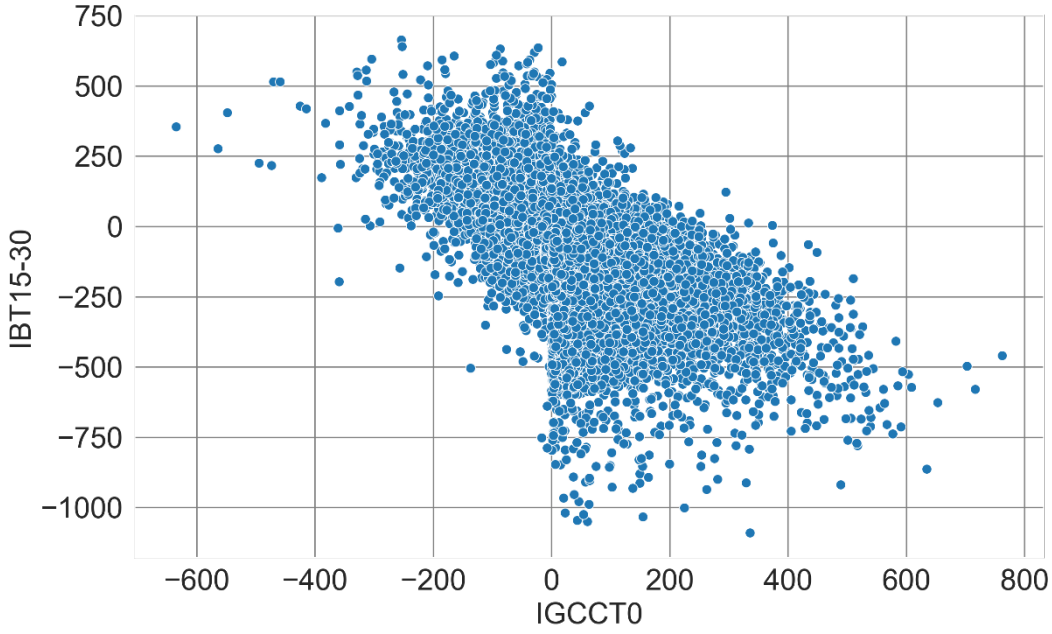

Figure 11: Scatter Plot: IGCCT0 vs IBT15-30.

As was stated in the introduction, renewable energy sources affect the grid stability. In order to see how much relation they have with the target parameter; a correlation matrix will be presented. This matrix contains all the parameters related to renewable energies that are meant to be used as predictors.

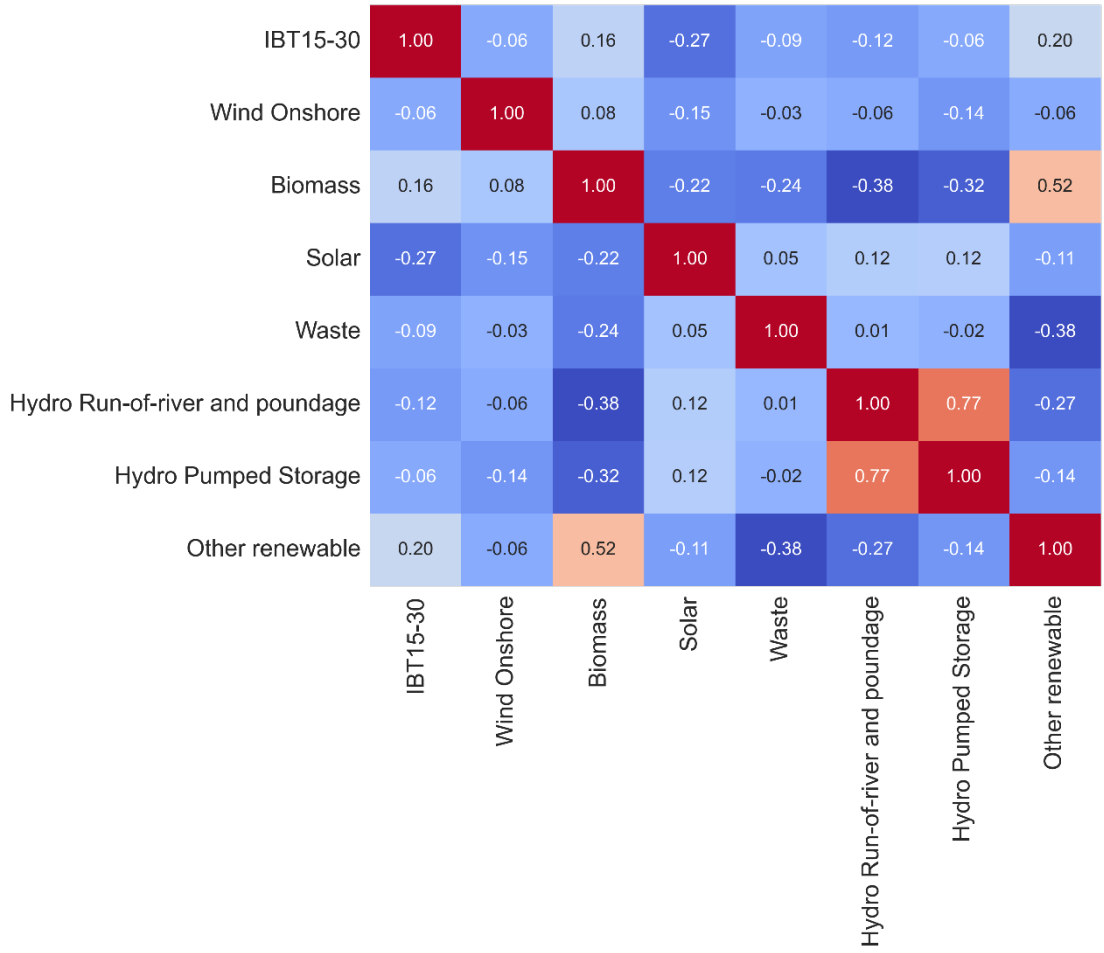

Figure 12: Correlation Heatmap (Renewable Energy Sources).

The correlation matrix shows that renewable energy sources have a relatively low impact on the target parameter. It should be noted that the Solar source is the one that has the highest value of correlation with IBT15-30, that's because it represents a great piece in the distribution of electricity generation in Hungary (as shown in the graph) and because of its high variability.

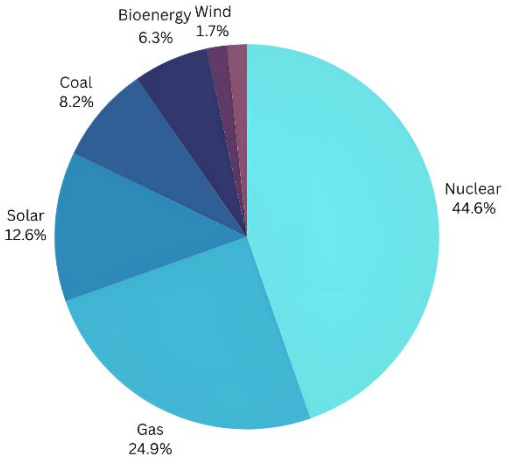

Figure 13: Distribution of electricity generation in Hungary in 2022, by source [19].

4.2 Pre-processing

Once the data that's going to be utilized is correctly analysed, the next step is to prepare it to provide it to the model as input. Several procedures must be done to do so:

Set the same time section for the datasets: Since each of the datasets has data from different time sections, a start and end date must be established in order to have a common period across all datasets. In the study, have been the following date limits have been set:

Start date: '2022-03-17 17:45:00+00:00'

End date: '2024-01-25 05:15:00+00:00'

Dealing with NaN values: The data that is being used in this study present some missing values that are filled with "NaN". To address this situation, these missing values have been replaced by the mean of the two previous and the two next values, obtaining that way a good representation of the parameter tendency at this point.

Scaling the data: This action ensures that the parameters have similar magnitude, avoiding the model's sensitivity to the data scale. In this research, a method from scikit-learn is used: StandardScaler. It transforms the data such that every feature has a mean of 0 and a standard derivation of 1. Moreover, this modification helps the algorithm to learn and process faster.

4.3 Feature Engineering

This section aims to help the algorithm uncover patterns in the data by adding features or transforming them. These adjustments have been applied in this research:

DateTime feature extraction: This technique transforms the index of the rows in information that can be used as inputs in the model. In this case, the index has been transformed into "hour", "day", "day of the week" and "month" parameters. This can provide very useful insights in this study since grid imbalances have a relation with power consumption and generation, which are very sensible to the time of the day, week, or month.

Working day variable: As the power consumption and generation also are very influenced by whether it is a holiday or not, a parameter that considers it has been created. This binary feature takes into account only the Hungary holiday schedule.

4.4 Machine Learning Introduction

As introduced earlier, machine learning will be employed to forecast power system imbalances, improving our ability to anticipate them more accurately.

4.4.1 WHAT IS MACHINE LEARNING?

[20] Machine learning emerged from the necessity to manage huge amounts of data, surpassing human capabilities. Data analysis allows comprehension of phenomena, modelling of behaviours, or making forecasts.

In the past, humans would analyse data, design algorithms, and then apply them through machines to solve problems. Nowadays, humans simply input data, enabling machines to learn from themselves. Obtaining relations and conclusions that many years ago would have required extensive study.

It's also important to stand out that machine learning has strong a relationship with statistics, data science, and artificial intelligence.

The machine learning method can be described in 5 steps:

1- Gather the data: Clean and pre-process the data if it has noise or improve its format. The quantity and quality of data significantly impact model performance.
2- Feature engineering: Select, transform, or create input variables to improve the model's performance, which has to be chosen carefully.
3- Training. Select the section of our data to make the model better and better at predicting.
4- Testing: Select a different section than the previous one to evaluate the model's accuracy.
5- Prediction: Use the trained model to forecast future data.

It goes without saying that this process is an iterative one, requiring many attempts and modifications until a robust model is achieved.

Although several machine learning approaches exist supervised, unsupervised, semi-supervised, and reinforcement learning, only the first one will be explained. This decision is based on the data employed in this study belonging to that category.

Supervised learning involves training the model with labelled data, where the outcomes are known, enabling the machine learning model to uncover patterns within the data. Once trained, the model can forecast results from inputs where the outputs are unknown.

4.4.2 MACHINE LEARNING BASICS

This section will help to understand the most basic concepts of machine learning in order to facilitate the comprehension of the algorithm. These concepts are as follows:

Decision trees: Algorithm that makes predictions by splitting the data into subsets, based on the value of a chosen parameter. Each subset is also split into other subsets until a stopping criterion is reached. Finally, a decision tree is created, and each input can be classified according to the values of its parameters. An example of a simple decision tree is shown in Figure 14.

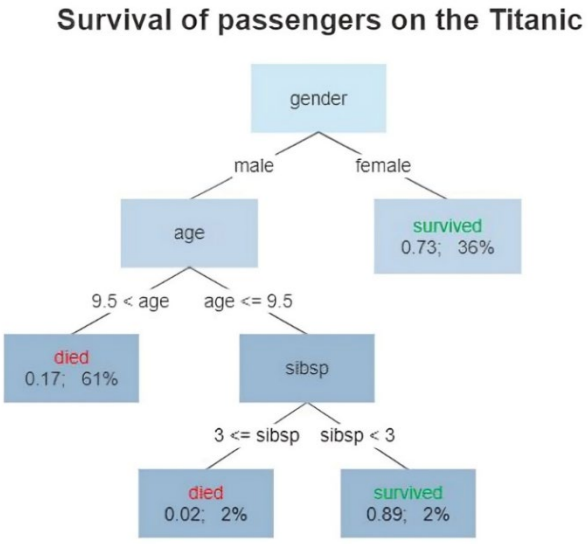

Figure 14: Decision tree example. [21]

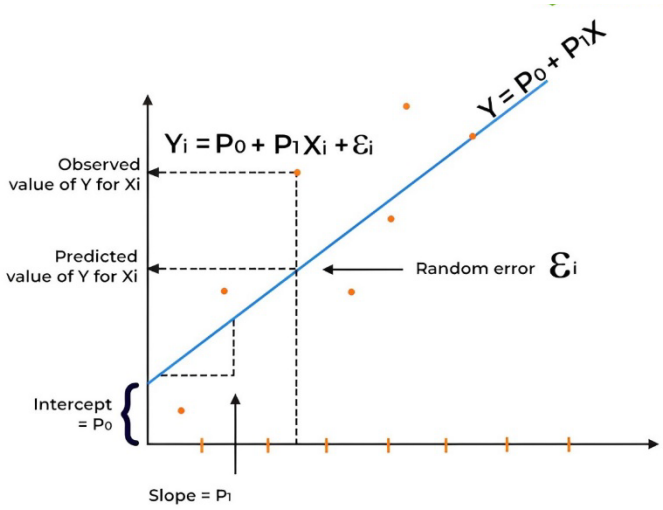

Figure 15: Linear regression graph. [22]

Loss functions: This is how the difference between forecasted and observed values is measured. It is used during training to evaluate the model's performance and penalize its errors. The accuracy of the model will be reflected in how low this function is. Depending on the kind of problem that is being faced, the loss function will be formulated in one way or another. A common one is the Root Mean Squared Error (RMSE), for regression problems.

Overfitting: It's a phenomenon that happens when a model performs well in the training data but fails its predictions with new data. Its explanation lies in the fact that the model has captured noise or random variations instead of real underlying patterns.

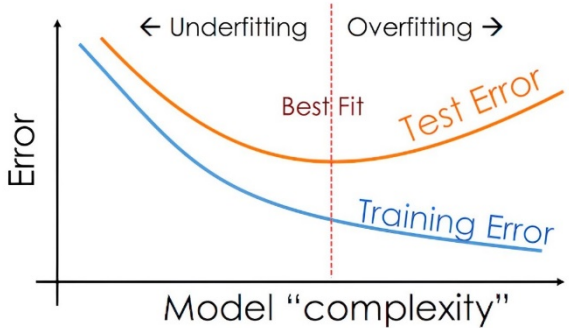

Figure 16: Overfitting graph. [23]

4.5 Light Gradient Boost Machine (LightGBM)

To understand LightGBM, it is essential to first study which are its foundations: **Gradient Boosting Machines (GBM)**.

4.5.1 GRADIENT BOOSTING MACHINES (GBM)

[15] Gradient Boosting Machines are machine learning algorithms that are used to solve both regression (our study case) and classification problems, through a gradual, additive, and sequential manner. Its methodology consists of combining sequentially weak learners (typically decision trees) to create a strong one and minimize a defined loss function.

GBM is considered an algorithm that provides accurate results with a high prediction speed, especially with large and complex datasets, like the one that has been analysed in this study.

To get a clearer vision of this kind of method, a step-by-step explanation is going to be done:

1- To do a first prediction based on the observations in the training dataset, the GBM will build a base model, which is usually an average of the target parameter on the regression tasks. This first approach has a very low prediction value. However, it will be useful to consider it as the baseline from which the error could be reduced gradually.
2- The algorithm will calculate the residuals (errors) which are: (observed value - predicted value). In our case, this is the difference between the observed value and the average value that has been calculated in the previous step.
3- The algorithm will build a model to forecast these residuals. So, it will use a weak learner (such as a decision tree) to obtain the first prediction.
4- Predictions from the decision tree are scaled using a parameter called the learning rate (this prevents overfitting). Then these scaled predictions are added to the average value and the next step is calculating the new and improved residuals. The evaluation of each forecast will determine the direction and magnitude of the needed adjustments to minimize the loss.
5- The model will do iterations by adding new decision trees, each of which is trained to predict just the residuals from the previous prediction by rectifying the errors.

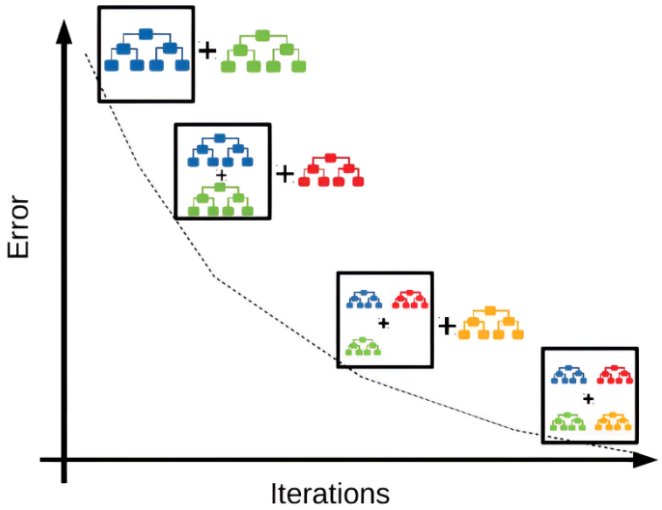

Figure 17: Evolution of a GBM algorithm. [24]

Once the foundations have been understood, the LightGBM can be explained in detail.

4.5.2 LIGHTGBM FEATURES

[25] Light Gradient Boosting is an open-source distributed; gradient boosting framework developed by Microsoft [26]. Designed for efficient, scalable, and high-performance machine learning tasks, particularly in the realm of decision tree-based algorithms.

The following are the key improvements over other GBM models:

Leaf-wise Growth: LightGBM adopts a leaf-wise splitting strategy, in contrast with other boosting algorithms that use a level-wise approach. It selects the leaf to split that it believes will provide the most significant reduction in the loss function. This leads to deeper trees with fewer nodes, resulting in higher accuracy and faster training.

By prioritizing splits based on their impact on the global loss rather than just the loss on a specific branch, it often will learn lower-error trees "faster" than level-wise (used by XGBoost). This speed is particularly beneficial for large-scale datasets and real-time applications.

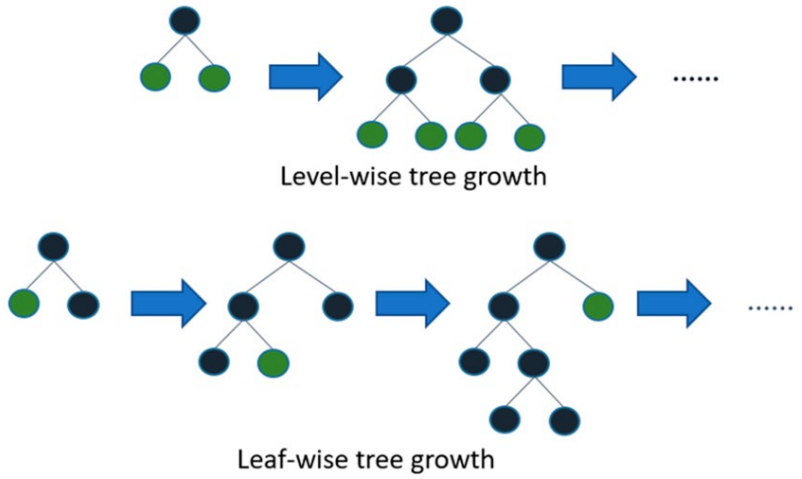

Figure 18: Level-wise vs leaf-wise tree growth strategy. [27]

Histogram-based Splitting: It's a technique used for efficiently finding optimal split points in decision trees during the training process. Instead of considering all possible split points for each feature individually, LightGBM constructs histograms of feature values and then selects the best split points from these histograms. This approach reduces computational complexity and memory usage, resulting in faster training times and improved scalability, especially for large datasets.

Gradient-based One-Side Sampling (GOSS): This method uses a different sampling method that can achieve a good balance between reducing the number of data instances and keeping the accuracy for learned decision trees. LightGBM will put more focus on undertrained data points during tree construction, leading to more efficient and accurate model training.

In a traditional gradient boosting algorithm, the gradient (residual error) for each observation provides useful information. For instance, if an observation has a small gradient, it indicates

a small training error, suggesting that it's already well-trained. Therefore, a simple method to decrease the number of instances is to remove observations with small gradients and concentrate on those with larger gradients, which means they have larger errors, so the data points are not learned well yet.

However, this approach can alter the data distribution and negatively impact the model's accuracy. To address this, GOSS employs a new sampling technique that retains all observations with large gradients while down samples those with small gradients. To reduce the effect on data distribution, GOSS introduces a constant multiplier for the observations with small gradients when calculating the information gain.

Exclusive Feature Bundling (EFB): It's a feature engineering technique aimed at improving the efficiency and effectiveness of decision tree construction, especially for datasets with many categorical features. EFB identifies groups of categorical features that never occur together in the same data instances (mutually exclusive) and bundles them into a single feature during tree construction. This reduces the number of decision rules needed to split the data, leading to faster training times and more interpretable models.

Categorical Feature Support: This characteristic allows the algorithm to directly handle categorical variables without requiring one-hot encoding. Resulting in faster training times and improved accuracy.

Customizable Parameters: LightGBM offers a wide range of customizable parameters to fine-tune model performance and adapt to different machine learning tasks.

4.5.3 WHY LIGHTGBM WAS CHOSEN OVER OTHER ALGORITHMS?

In the context of short-term forecasting of system power imbalances, several machine learning algorithms were considered, including traditional linear models, decision trees, random forests, gradient boosting machines, and neural networks. After a comprehensive evaluation, LightGBM was chosen for several reasons:

- High computational efficiency, which is critical for real-time forecasting applications.
- Ability to handle large datasets, like the one used in this study.
- Incorporation of mechanisms to reduce overfitting.

Let's look at its competitors:

- Linear Regression has a limited ability to capture non-linear patterns, which are so present in the imbalance data.
- Other tree-based methods such as Decision Trees, Random Forest, or other GBM are outperformed in terms of speed and efficiency while maintaining or exceeding accuracy levels. This will be proven in the results rection section 5.5.
- Tree-based models still outperform deep learning in dealing with tabular data. [28]

CHAPTER 5

EXPERIMENTAL SETUP

5.1 Model Specifications

Machine learning and Light GBM algorithm have been explained earlier. Now it's time to particularise it to the study case. These are the adopted specifications:

Deterministic forecasting: The model provides one single output for each input, however, unlike probabilistic forecasting, it does not offer information about the uncertainty of the prediction. Despite that, this method is used because its processing time is shorter since fewer parameters are computed.

Although probabilistic forecasting is also an option to be considered, this procedure is not used in this research and will be proposed in the conclusion section in order to continue with the study.

Timeseries split cross-validation (TSCV): As has been explained in the introduction to Machine Learning, the data should be split into two sections: training and evaluating. TSCV is a technique specifically designed for time series data samples, like the ones used in this study. Its virtue is that, unlike other cross-validation methods, TSCV respects the temporal order of the data.

This method divides the dataset into multiple folds, ensuring that the model is trained on past data and evaluated on future data, simulating a real-word situation. So, using this methodology is necessary to accurately evaluate the performance of models based on time series data.

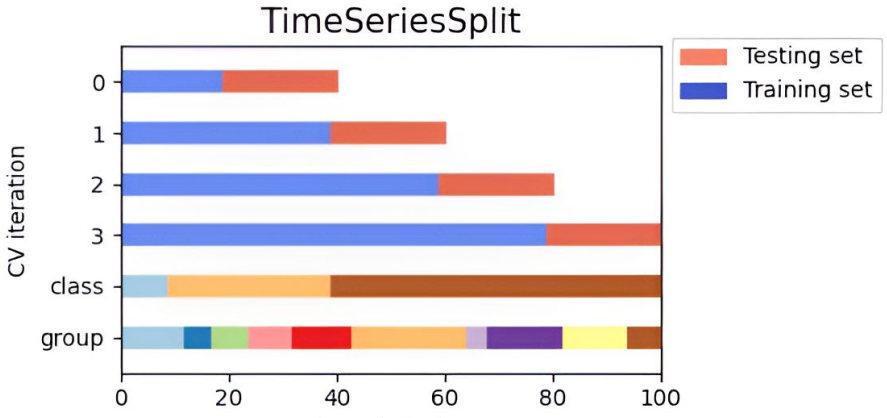

Figure 19: Time series split example. [29]

Hyperparameter tuning: Optuna will be used for optimization, with a range of 100 to 500 trials. Optuna is an efficient hyperparameter optimization framework that uses a combination of algorithms, such as Tree-structured Parzen Estimator (TPE) and Bayesian optimization, to intelligently sample the hyperparameter space. It selects sets of parameters from a

predefined search space and evaluates their performance using TSCV. Optuna will repeat this process as many times as specified to find the configuration that minimizes the loss function.

Unlike RandomizedSearch, which tests combinations of hyperparameters randomly, Optuna does this process using the history data of trials completed thus far. That way this last method can more effectively navigate the hyperparameter space, potentially leading to better model performance in fewer iterations.

5.2 Evaluation Metrics

Machine learning models need some evaluation metrics for several reasons: performance measurement, comparison of models, hyperparameter tuning, detecting overfitting, evaluation of business impact…

In the case of this study, five different metrics have been employed:

Root Mean Squared Error (RMSE): Commonly used in regression tasks, RMSE provides a measure of the average errors between forecasted and actual values. It's characterized by its sensitivity to large errors, which are more penalized than the small ones, due to the squaring of differences. On the other hand, even though this measurement is expressed in the same units as the target value, it is not easy to interpret its value in isolation. So, it must be compared with the RMSE of other models. This is its mathematical formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

As in the following formulas, $\hat{y}$ represents the forecasted values and y represents the actual values.

Mean Absolute Error (MAE): As RMSE, MAE is also widely used in regression problems. However, unlike RMSE, MAE penalizes all errors proportionally; measuring the average absolute difference between predicted and observed values. Moreover, MAE it's easier to interpret than RMSE because it has a clear meaning: the average error of the model. Its mathematical expression is the following.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Normalized Mean Absolute Error (nMAE): It's a modification of MAE that provides its normalized value. The difference between the previous one is that nMAE is calculated by dividing MAE by the mean of the actual values and multiplying the result by 100, to give it a percentage format. It provides that way, a percentage that allows the comparison across models, datasets, and scales, giving an easily interpretable value. Mathematically, it's expressed like this:

$$nMAE = \frac{MAE}{Mean\ of\ Actual\ Values} \times 100$$

Variance Ratio: It's used in regression models to measure the ratio of the variance of the forecasted values to the variance of the actual values in the dataset. It should be as near to 1 as possible, meaning that the predictions vary similarly to the actual values. Its formula is the following:

$$VR = \frac{var(\hat{y})}{var(y)}$$

Correlation: This parameter quantifies the degree to which changes in one variable are associated with changes in another variable, in this case between forecasted values and the actual ones. It's easy to interpret values near to 1 mean positive correlation, values near to -1 indicate negative correlation, and values close to 0 mean weak or no correlation.

Direction accuracy: Determines de percentage of predicted values that have the same sign as the actual values. If the actual value is positive and the forecasted value is also positive, the direction is considered to be correct. Similarly, if the actual value is negative and the forecasted value is also negative, the direction is considered to be correct. The more it approximates to 100% the better (although this may be a sign of overfitting), values far away from 1 mean a very poor accuracy.

CHAPTER 6

RESULTS

This chapter will show all the results achieved under different conditions, varying the selected input data, the number of iterations, the hyperparameters search space, and other parameters. Moreover, other models will be tested to demonstrate improvements over the baseline and the high efficiency of LightGBM. In order to have a clear idea of the parameters that are going to be analised, let's take a look at their definitions:

max_train_size: This parameter specifies the maximum number of training samples used in each split of time series cross-validation. It limits the size of the training set to prevent it from growing too large as more data becomes available over time. Depending on the amount of disponible data its value will have to be bigger or smaller.

n_splits: This parameter defines the number of folds or splits the data is divided into during time series cross-validation. Its value depends on the dimension of the max_train_size.

n_trials or **n_iterations**: These terms refer to the number of times the hyperparameter search algorithm runs to find the best set of hyperparameters for a model. Each trial or iteration involves selecting a different set of hyperparameters, training the model, and evaluating its performance, aiming to optimize the model's performance by finding the best hyperparameters. More iterations mean more probability of finding the best possible hyperparameters, but it will require more computational cost as a price.

6.1 Baseline Model

As a baseline, it's going to be used a Linear Regression model, which will have as inputs only the imbalances parameters. This model provides the next benchmarks:

RMSE	MAE	nMAE(%)	Variance Ratio (%)	Correlation	Direction Accuracy (%)
64.315	45.770	46.527	80.183	0.864	84.954

Table 1: Baseline model results.

Specifications: As max_train_size=35040 (one year of data) and n_splits=12.

6.2 Different Datasets

To reach the best results the entire dataset has been divided into smaller groups:

Subset Name	Content
IMBALANCES	All parameters of type 'Imbalance TX'
IGCCT0	'IGCCT0'
REST MAIN	'SolarprogT0', 'SolarprogT-50', 'SolarprogT50', 'SolarprogT-20', 'SolarprogT20', 'SolarprogT-35', 'SolarprogT35', 'GrossElectricityLoadforecastT0', 'GrossElectricityLoadforecastgradT30', 'GrossElectricityLoadforecastgradT15', 'GrossElectricityLoadforecastgradT5', 'GrossElectricityLoadforecastgradAVG', 'DAMpriceT-30', 'DAMpriceT0', 'DAMpriceT30', 'DAMpriceT60'
IMP-EXP	'Actual Import-Export', 'Planned Import-Export'
ENERGY MIX	'Gross System Load - Validated', 'Gross System Load - gross.op.meas', 'Net actual generation sum', 'Nuclear', 'Fossil Brown coal/Lignite', 'Fossil Gas', 'Fossil Hard coal', 'Fossil Oil', 'Wind Onshore', 'Biomass', 'Solar', 'Waste', 'Hydro Run-of-river and poundage', 'Hydro Pumped Storage', 'Other renewable', 'Other', 'Import-export balance', 'other correction of DSOs'
DATES	'day', 'month', 'day_of_week', 'hour'
IS WORKING DAY	'IsWorkingDay'

Table 2: Data subsets.

Firstly, in order to see which combination of subsets gives the best forecast the model will be tested with these common parameters: max_train_size = 35040 (a year of data), n_splits = 12, n_trials = 100. These are the results for the different subset combinations:

IMB.	IGCCT0	REST MAIN.	IMP-EXP	EN. MIX	DATES	IS W. DAY	nMAE (%)	VR (%)	Correlation
✓							49.644	74.341	0.837
✓	✓						49.544	74.744	0.838
✓	✓	✓					44.155	78.441	0.874
✓	✓	✓	✓				44.161	78.773	0.873
✓	✓	✓	✓	✓			44.667	74.040	0.874
✓	✓	✓	✓		✓		42.554	77.032	0.878
✓	✓	✓	✓		✓	✓	42.815	77.448	0.878
✓	✓	✓			✓	✓	42.801	77.336	0.878

Table 3: Comparison of the results of different subset combinations.

As it is clearly appreciated, the best nMAE score is achieved with the sixth combination (the red one). These results show that the IMP-EXP, the ENERGY MIX, and the IGCCT0 subset do not provide too much useful information to the model. On the other hand, IMBALANCES, REST MAIN and DATES are the subsets that have more explanation power.

6.3 Max_train_size and n_splits.

Secondly, this research will compare the results of varying the max_train_size and the n_splits. The n_trials will be set to 100 and IMBALANCES is going to be the single subset used in the following experiments.

MAX_TRAIN_SIZE	DAYS	N_SPLITS	nMAE (%)	VR (%)	Correlation
2976	31	12	50.165	69.258	0.833
5760	60	12	49.981	70.131	0.834
5760	60	16	49.902	70.978	0.835
35040	365	5	49.576	74.008	0.837
35040	365	8	49.615	73.988	0.838
35040	365	12	49.644	74.341	0.837
35040	365	16	49.625	74.296	0.837
48000	500	8	49.493	73.995	0.838

Table 4: Comparison of results when varying max_train_size and n_splits.

These results insight that the variance ratio and correlation are influenced by the max_train_size and the change of n_iter does not alter the results significantly. So, for the next experiments, these parameters will be 35040 (a year of data) and 12, respectively. That way, the research considers the variability of the imbalance through an entire year and can collect the patterns of every season.

6.4 Hyperparameter Selection Method

The next step is making a comparison between Randomized Search and Optuna, considering the accuracy and the execution time. For this experiment, the employed datasets are "IMBALANCES", "IGCCT0", "REST MAIN", "IMP-EXP" and "DATES". Which are the ones that have achieved the best score in section 6.2.

The common parameters will be the max_train_size (35040), n_splits (12), and the hyperparameter search space. On the other hand, the variation of the hyperparameter tunning method and the n_iterations will be the parameters that are going to be analysed.

HYPERPARAMETER TUNNING	N_ITER.	nMAE (%)	VR (%)	Correlation	DIR. ACC. (%)	EX. TIME (mins)
Optuna	100	42.554	77.032	0.878	86.075	41
Randomized	100	41.433	78.176	0.879	85.799	52
Optuna	500	42.093	77.090	0.881	86.336	288
Randomized	500	41.848	77.826	0.877	85.814	319

Table 5: Comparison when varying hyperparameter selection method and n_iterations.

Two insights can be drawn from these experiments. On one hand, the results of the Randomized Search method are slightly better than its competitor in terms of nMAE and variance ratio, while they are slightly worse in correlation, direction accuracy, and execution time. On the other hand, considering the variation in the number of iterations, the Optuna method has subtly improved its results with an increase in iterations, whereas the Randomized Search method has shown the opposite trend (that is by chance, typically the more n_iterations the better the prediction, as it explores a larger number of hyperparameter combinations).

So, in the case of this research seems that Randomized Search performs modestly better than Optuna, in general terms. Moreover, the models achieve good performance with only 100 iterations, not improving the results too much with 500 iterations while investing too much more time.

6.5 Comparison with other ML models

Finally, to get to know the performance of the LightGBM model competitors. A common set of parameters have been set to do a comparison:

max_train_size=35040, n_splits=12, hyperparameter tunning method: Optuna, n_iterations=100.

ALGORITHM	nMAE (%)	VR (%)	Correlation	DIR. ACC. (%)	EX. TIME (mins)
Linear Regression	46.204	80.349	0.865	84.885	0.0617
Random Forest	45.939	68.556	0.859	85.085	16
XGBoost	42.749	77.900	0.879	85.952	89
CatBoost	42.985	76.733	0.879	85.845	215
LightGBM	42.554	77.032	0.878	86.075	41
LightGBM + Linear Regression	43.288	80.397	0.883	85.799	46

Table 6: Comparison of LightGBM model competitors.

As was anticipated while explaining LightGBM, this model outperforms both Linear Regression and Random Forest. Nevertheless, when it comes to comparing it to other GBM like XGBoost and CatBoost, Light GBM stands out not so much for its accuracy but for its efficiency and processing speed, which are clearly superior.

Furthermore, an ensemble model combining LightGBM and Linear Regression has been tested. Achieving an improvement in the variance ratio and correlation but scoring a worse nMAE. It's a good model but can't be considered better that the single LightGBM model.

Once the results have been presented and analysed, a final study conclusion could be drawn.

CHAPTER 7

CONCLUSION AND FURTHER STEPS

The results of this investigation demonstrate that LightGBM significantly improves the accuracy of short-term imbalance forecasts due to its advanced techniques. This model outperformed traditional algorithms such as Linear Regression and Random Forest, and it also proved its high efficiency compared to other gradient boosting methods like XGBoost and CatBoost. The results highlight the potential of LightGBM to provide valuable information about the factors that influence energy imbalances, contributing to maintaining grid stability and reliability. Achieving the next results:

MODEL	HYPERPARAMETER TUNNING	nMAE (%)	VR (%)	Correlation	DIR. ACC. (%)	EX. TIME (mins)
LightGBM	Randomized	41.433	78.176	0.879	85.799	52

Table 7: LightGBM best results.

Specifications: max_train_size=35040, n_splits=12, and n_iterations=100.

Although the model has achieved fine predictions, there is still much room for improvement. Through the development of Machine Learning models or the incorporation of additional input data, such as meteorological factors (e.g., temperature history and temperature prediction), as well as data on reserve activation.

The study faced certain limitations, such as the deterministic forecasting methodology. Instead of this, it could be done with a probabilistic forecasting approach. Which consists of offering a range of possible outcomes with associated probabilities, giving a more comprehensive view of potential future states. This is particularly useful in managing uncertainties. In our research case, this methodology helps TSOs and BRPs to make more informed decisions, improving risk management.

This probabilistic approach has not been pursued due to the limited computational resources available for the study and the real-time application requirements. Deterministic forecasting, on the other hand, requires less computational power and time, making it more suitable for this short-term forecast.

Summarizing, this research highlights the potential of advanced Machine Learning techniques, particularly LightGBM, in managing power systems and dealing with the incorporation of renewable energy into the grid. While the results are promising, future work should be done to explore additional data sources, probabilistic methods, and new improved algorithms. This study marks a step toward achieving a more stable and efficient electrical grid, walking toward a sustainable energy future.

REFERENCES

[1] UN Climate Change Conference (COP21), "Paris Agreement". December 2015. Available: https://unfccc.int/sites/default/files/english_paris_agreement.pdf

[2] Renewable Energy Directive (Key facts). Available: https://energy.ec.europa.eu/topics/renewable-energy/renewable-energy-directive-targets-and-rules/renewable-energy-directive_en

[3] Nano Energies, "Manual Frequency Restoration Reserve (mFRR)" Available: https://nanoenergies.eu/knowledge-base/manual-frequency-restoration-reserve-mfrr

[4] Nano Energies, "Frequency Control Reserve (FCR)" Available: https://nanoenergies.eu/knowledge-base/frequency-control-reserve-fcr

[5] ENTSO-E, "Manually Activated Reserves Initiative". Available: https://www.entsoe.eu/network_codes/eb/mari/

[6] ENTSO-E, "PICASSO". Available: https://www.entsoe.eu/network_codes/eb/picasso/

[7] ENTSO-E, "Imbalance Netting". Available: https://www.entsoe.eu/network_codes/eb/imbalance-netting/

[8] Tomás Oliveira, "Understanding Day-ahead & Intraday Markets". May 2023. Available: https://www.synertics.io/blog/39/understanding-day-ahead-intraday-markets

[9] Henning Thiesen and Clemens Jauch. "Application of a New Dispatch Methodology to Identify the Influence of Inertia Supplying Wind Turbines on Day-Ahead Market Sales Volumes" February 2021. Wind Energy Technology Institute (WETI), Flensburg University of Applied Sciences, 24943 Flensburg, Germany. DOI: 10.3390/en14051255

[10] Nano Energies, "Balance Responsible Party (BRP)" Available: https://nanoenergies.eu/knowledge-base/balance-responsible-party-brp

[11] Maria P. Garcia and D.s. Kirschen, "Forecasting System Imbalance Volumes in Competitive Electricity Markets". 30 January 2006 Power Systems, IEEE Transactions on 21(1):240 - 248. DOI: 10.1109/TPWRS.2005.860924

[12] Š. Kratochvíl, "System Imbalance Forecast". November 2016. Available: https://dspace.cvut.cz/bitstream/handle/10467/67662/Disertace_%c5%a0t%c4%9bp%c3%a1n_Kratochv%c3%adl_2016.pdf?sequence=1&isAllowed=y

[13] Carolina Contreras, "System imbalance forecasting and short-term bidding strategy to minimize imbalance costs of transacting in the Spanish electricity market". July 2016. Available: https://repositorio.comillas.edu/xmlui/bitstream/handle/11531/16621/TFM000596.pdf?sequence=1&isAllowed=y

[14] Tárik S. Salem, Karan Kathuria, Heri Ramampiaro and Helge Langseth. "Forecasting Intra-Hour Imbalances in Electric Power Systems". July 2019. Available: https://ojs.aaai.org/index.php/AAAI/article/view/5021

[15] Jonathan Dumas, Ioannis Boukas, Miguel Manuel de Villena, Sebastien Mathieu, Bertrand Cornélusse. “Probabilistic forecasting of imbalance prices in the Belgian context”. June 2021. Available: https://arxiv.org/pdf/2106.07361 DOI: 10. 1109/ EEM. 2019. 89163 75

[16] Jeremie Bottieau, Louis Hubert, Zacharie De Greve and François Vallee. “Very-Short-Term Probabilistic Forecasting for a Risk-Aware Participation in the Single Price Imbalance Settlement”. March 2020. Power Systems, IEEE Transactions on 35(2):1218 – 1230 DOI: 10.1109/TPWRS.2019.2940756

[17] Ignacio Rojas, Héctor Pomares, Luis Javier Herrera, Fernando Rojas and Olga Valenzuela. “The 9th International Conference on Time Series and Forecasting” *Engineering Proceedings*. P191-200. July 2023. Available: https://mdpi-res.com/bookfiles/book/9193/The_9th_International_Conference_on_Time_Series_and_Forecasting.pdf?v=1715011201

[18] MAVIR, Available: https://www.mavir.hu/web/mavir-en/data-publication

[19] Distribution of electricity generation in Hungary in 2022, by source. Available: https://www.statista.com/statistics/1235432/hungary-distribution-of-electricity-production-by-source/#:~:text=Hungary%20sources%20most%20of%20its,energy%20source%20in%20the%20country.

[20] Ramzi Farhat, Yosra Mourali, Mohamed Jemni and Houcine Ezzedine. “An overview of Machine Learning Technologies and their use in E-learning”. IEE. February 2020. DOI: 10.1109/OCTA49274.2020.9151758

[21] Wikipedia, “Decision tree learning”. Available: https://en.wikipedia.org/wiki/Decision_tree_learning

[22] Preeti Aggarwal. “Understanding Loss Functions”. November 2022. Available: https://heartbeat.comet.ml/understanding-loss-functions-6ad2c0a5bc23

[23] Aakash N S, “Machine Learning with Python and Scikit-Learn – Full Course”. Lesson 5 - Gradient Boosting Machines with XGBoost. November 2023. Available: https://youtu.be/hDKCxebp88A?si=gJ7uQdX0Y7jiP7RV

[24] Aratrika Pal. “Gradient Boosting Trees for Classification: A Beginner’s Guide”. October 2020. Available: https://medium.com/swlh/gradient-boosting-trees-for-classification-a-beginners-guide-596b594a14ea

[25] Masoud Seyyedattar, Sohrab Zendehboudi, Ali Ghamartale and Majid Afshar “Advancing hydrogen storage predictions in metal-organic frameworks: A comparative study of LightGBM and random forest models with data enhancement”. 6 May 2024. DOI: 10.1016/j.ijhydene.2024.04.230

[26] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye and Tie-Yan Liu. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”. December 2017. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[27] Lanfa Liu, Min Ji and Manfred Ferdinand Buchroithner. “Combining Partial Least Squares and the Gradient-Boosting Method for Soil Property Retrieval Using Visible Near-Infrared Shortwave Infrared Spectra”. December 2017. Available: https://www.mdpi.com/2072-4292/9/12/1299 DOI: 10.3390/rs9121299

[28] Léo Grinsztaj, Edouard Oyallon and Gaël Varoquaux. “Why do tree-based models still outperform deep learning on typical tabular data?” Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/0378c7692da36807bdec87ab043cdadc-Paper-Datasets_and_Benchmarks.pdf

[29] Scikit-Learn.org, “Cross-validation: evaluating estimator performance”. Available: https://scikit-learn.org/stable/modules/cross_validation.html