Universiteit Leiden
The Netherlands

Bachelor Thesis

Evaluation methods and model performance
for imbalance price forecasting

Ivar Martens

Supervisors:
Francesco Bariatti & Matthijs van Leeuwen

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

January 15, 2024

Abstract

Presently, the Dutch electrical grid is a topic of significant concern and discussion. Unprecedented energy price and challenges associated with grid congestion have emerged as primary focal points. To help solve these problems on the grid the startup simpl.energy wants to use imbalance price forecasting to optimise grid utilisation. In this research we investigate how we can assess model performance with quantitative measures instead of the expert visually assessing model performance. This adds to the very limited previous research on imbalance price forecasting. Specifically to the assessment of model performance regarding our specific use case of helping companies lower their energy costs and participating in balancing efforts on the Dutch electricity grid.

The first part of this research entails comparing quality measures. For this purpose we compare three more standard error based quality measures and four other quality measures. Two of these other quality measures are developed by us, the first punishing bad predictions called the “Punishment score” and the other calculating the Root Mean Squared Error (RMSE) over slope values instead of the real values called the “slope RMSE””. Together with an expert from simpl.energy we developed fictitious models that reflect expert opinion of bad, medium and good model performance. We consider a quality measure to be good if it is able to separate these three classes. The length of the data sets plays a big role in the performance of the quality measures, which can likely be attributed to averaging effects and its impact on the characteristics of each quality measure. However, the Mean Absolute Error(MAE), relative MAE (rMAE) and Punishment score were able to separate all classes over all lengths of data sets and are thus considered the best.

The second part of this research has two goals. The first goal is to compare different forecasting models using open-source data. This comparison involves a naive model, a statistical regression model, a tree based machine learning model and a neural network model. The expert from simpl.energy ranked neural network algorithm called NHITS as best performing of these 4 models followed by the tree based machine learning model called lightGBM and the statistical regression model called Lasso. The naive model performed the worst. From this experiment we conclude that this open-source data does not have enough correlation to make accurate predictions with the models that are used. The second goal was to test the findings of the first part with these real forecasting algorithms. The RMSE and correlation were the only quality measures to correctly reflected the expert ranking of the different models. Which opposed the findings of the first experiment, where the RMSE and correlation were not able to capture the expert opinion where the MAE, rMAE and punishment score were.

Contents

1 Introduction 1

2 Background Knowledge 2

2.1 Electricity Grid . . . 2

2.1.1 Actors on the electricity grid . . . 2

2.1.2 Merit order . . . 3

2.1.3 Balancing Mechanisms . . . 3

2.1.4 Energy Markets . . . 4

2.2 Use case . . . 6

2.3 Data . . . 6

3 Literature review 9

4 Evaluation measure 11

4.1 Good/Bad predictions . . . 11

4.2 Quality measure . . . 12

4.3 Fictitious model predictions . . . 14

4.4 Determining best evaluation measure . . . 17

5 Forecasting Algorithms 20

5.1 Models . . . 21

5.2 Model performance determined by the expert . . . 22

5.3 Model performance determined by the quality measures . . . 25

6 Conclusion 27

References 29

Appendix A 30

Table 1: Important Terms in the Paper

Term	**Definition**
Net Congestion	Net congestion, also known as congestion, occurs when the electricity grid becomes overloaded due to excessive demand or an excess of electricity supply. With the increasing electrification and growing use of renewable energy sources, net congestion poses a significant challenge to the reliable and efficient delivery of electricity to consumers and businesses. It requires approaches such as upgrading the infrastructure and balancing mechanisms to maintain a stable and efficient electricity system.
Imbalance Prices	The resulting price from the balancing market. Described in more detail in Section 2.1.4.
Grid Frequency	The electricity grid uses alternating current. The frequency for this alternating current on the Dutch grid is 50 Hz. This frequency rises or drops depending on the imbalance of supply and demand on the grid. Big deviations from the 50 Hz can result in power loss or a total blackout might occur.
Ancillary Services	The mechanisms used to restore grid balance. Described in more detail in Section 2.1.4.
Upward (Short)	Supplying energy to the grid. Is reffed to in the context of bids, regulation and price
Downward (Long)	Demanding energy from the grid. Is reffed to in the context of bids, regulation and price
Bid activation	The maker of a bid is asked to supply or consume energy. This can be done in the present or the future depending on the market.

1 Introduction

Presently, the Dutch electrical grid is a topic of significant concern and discussion. Unprecedented energy price and challenges associated with grid congestion have emerged as primary focal points. These issues are very important in the transition towards renewable energy. Companies are grappling with the implications of soaring energy costs. And congestion on the electricity net is hindering their ability to electrify industrial processes. Furthermore, grid congestion acts as a bottleneck, impeding the development of solar-, wind- and other renewable energy projects. While the long-term remedy for these congestion issues involves expanding the electrical grid to accommodate increased supply and demand, such expansion requires substantial investments of both time and financial resources. Consequently, an immediate solution lies in the *optimisation of grid utilisation*, a task in which forecasting plays a pivotal role. Currently, there are no feasible business cases for companies to help circumvent grid congestion and participate in grid balancing efforts. Forecasting facilitates such business cases and thus are part of the solution.

This research project was initiated by an energy management startup called simpl.energy. Their efforts regarding grid usage optimisation are discussed in Section 2.2. This research seeks to contribute to the scientific literature on electricity price forecasting and specifically to the limited research on forecasting of the Dutch imbalance price, a price resulting from balancing efforts detailed in Section 2.1.4. A model with the ability to forecast imbalance would enable simpl.energy in their efforts to aid companies in making informed decisions to support market equilibrium while simultaneously reducing expenditure on energy. Such a model should have a high enough accuracy that clients feel comfortable taking financial risks based on the predictions of the model. This means that simpl.energy should be able to support their claims about model performance to their clients in a confident manner. However, right now the expert from simpl.energy visually judge whether the performance of a model has a high enough accuracy. The reason for this is that standard quality measures such as Root Mean Squared Error (RMSE) do not reflect model performance well enough for the specific use case described in Section 2.2 and there has been no research regarding quality measures in this use case. To tackle the issue of assessing model performance this study aims to find the best quantitative measure for our use case. We do this by addressing the following research questions:

Question 1: *What quantitative measure for model performance, in the realm imbalance price forecasting, best reflects the expert opinion of "sufficient" accuracy?*

Question 2: *Does the proposed quantitative measure correctly reflect model performance when using real forecasting algorithms?*

To address these questions, our research comprises two segments. Firstly, we conduct a comparison of seven quality measures applied to fictitious models across various data set lengths. These fictitious models are developed to represent expert opinion, categorised into the classes bad, medium and good. A good quality measure is able to separate these three classes. The second part involves implementing real forecasting algorithms to test their performance and validate the conclusions drawn from the first part. The expert from simpl.energy ranks the models visualy, and we compare this ranking with the outcomes of all quality measures.

2 Background Knowledge

In this section some important background knowledge will be given in order to understand the terms used in this paper and the mechanisms influencing the imbalance prices. The electricity grid, its actors and markets are discussed. More details about the mechanisms of the balancing market are given. The use case for a energy management company like simpl.energy is discussed. Lastly, the data used in this research is detailed.

2.1 Electricity Grid

This section describes the operation of the North-Western European electricity grid. Details might differ for each country in this region. The details provided, such as company names, are those of the Dutch electricity grid.

2.1.1 Actors on the electricity grid

In the electricity sector, various entities assume specific roles and responsibilities to ensure the efficient operation of the electrical grid. The active parties on the electricity grid are [20]:

- Transmission System Operator (TSO)
- Balance Responsible Parties (BRP)
- Balance Service Providers (BSP)
- Nominated Electricity Market Operators (NEMOs)

TenneT is the Transmission System Operator (TSO) for the Netherlands, which is tasked with securing the grid's power supply. TenneT also invests in critical electrical infrastructure and manages bids for the balancing market, which is detailed below.

Balance Responsible Parties (BRPs), companies like Eneco and Vattenfall, operate on the electrical grid. These companies serve as the link between electricity consumption and generation, taking responsibility for addressing imbalances resulting from deviations between electricity consumption and generation. They provide forecasts for energy consumption and generation a day in advance through a so called E-program. This E-program is divided into 15-minute intervals known as Imbalance Settlement Periods (ISP). This E-program can be changed up until 1 hour before energy delivery.

Balance Service Providers (BSPs) contribute to grid stability by offering ancillary services. These services are offering short-term balancing services, for example a company allowing usage of a battery or a generator to (partly) restore balance, if the final E-programs still contain imbalances. BSPs can submit bids a day in advance for automatic participation in the Frequency Restoration (aFFR) mechanism discussed in Section 2.1.3. These bids remain modifiable until half an hour before the ISP of delivery.

Lastly, the Nominated Electricity Market Operators (NEMOs). They have the responsibility for running the so called spot markets (the day-ahead market and the intraday market). For the Dutch market the two NEMOs are EPEX spot and Nordpool.

2.1.2 Merit order

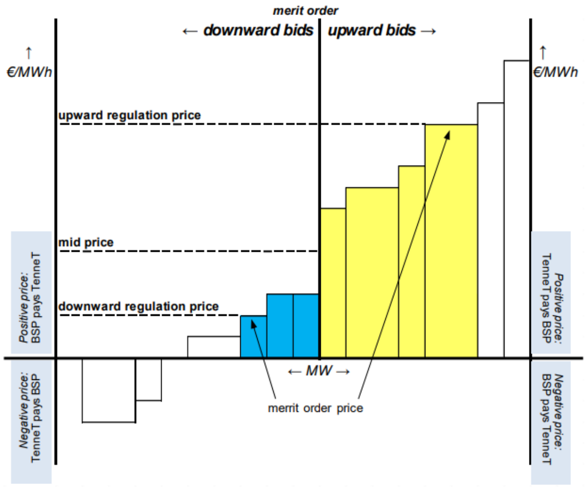

merit order
← downward bids
upward bids →
↑ €/MWh
↑ €/MWh
upward regulation price
mid price
downward regulation price
Positive price: BSP pays TenneT
Negative price: TenneT pays BSP
Positive price: TenneT pays BSP
Negative price: BSP pays TenneT
← MW →
merrit order price

Figure 1: Marginal price setting based on the merit order for upward and downward regulation [19]

The workings of a merit order system is important for understanding the way energy pricing is done for the markets detailed in Section 2.1.4 and is best described by Figure 1[1]. This figure shows the workings of a merit order for upward (supplying energy to the grid) and downward (demanding energy from the grid) regulation. At the right side of the figure are the upward regulating bids. These bids are sorted based on the values of the bids. The first bid to be activated (The bid maker is asked to supply energy to the grid. This can be in the present or future depending on the market.) is the lowest upward regulating bid (the most left yellow bar). When this bid is fully active the next bid is activated and so on until the demand is met. The yellow bids are all the activated bids for this merit order. The upward regulating price is determined by the last activated bid (the most right yellow bar). All the activated bids receive the upward regulating price for their energy. The downward regulated price is determined by the lowest activated downward bid (the most left blue bar), the exact reverse mechanism of the upward regulating price. Lastly, there is the so called mid price, which is the average of the highest activated downward bid (the most right blue bar) and the lowest activated upward bid (the most left yellow bar). This mechanism ensures fair pricing as each bidder bids as low/high as possible to be activated, but high enough to ensure profitability.

2.1.3 Balancing Mechanisms

The supply and demand on the grid needs to be balanced at all times to prevent outages. Due to the increased use of more volatile renewable energy sources this balance becomes harder to maintain. This puts much more importance on the Dutch ancillary services used for restoring balance. Figure 2 gives a comprehensive overview of the of the way the different mechanisms operate together. These ancillary services consist of three mechanisms [19]:

- Frequency containment reserve (FCR)
- Automatic Frequency Restoration Reserve (aFRR)
- Manual Frequency Restoration Reserves (mFRR)

[1] TenneT. Imbalance pricing system: How are the (directions of) payment determined? 2022

Frequency Containment Reserves (FCR) is a first line of defence. In compliance with its international balancing responsibilities, TenneT contracts a prescribed quantity of FCR, as stipulated by EU regulations. The allocation and activation of FCR operate on an automatic basis, contingent upon the prevailing frequency[2], rather than being administered by the TSO. This mechanism ensures that FCR are seamlessly engaged in response to frequency deviations without the need for direct TSO intervention.

Automatic Frequency Restoration Reserves (aFRR) are acquired through a mechanism known as "bid-obligations." Under this arrangement, Balance Service Providers (BSPs) are contracted to furnish TenneT with bids for balancing energy at certain times. Additionally, BSPs not under contract have the option to submit what are referred to as "free-bids," which are considered alongside contracted bids within a common merit order list. This bid-obligation system ensures the consistent availability of an ample pool of balancing bids, thereby promoting grid stability, while concurrently allowing market forces to govern the pricing dynamics. Consequently, the market-driven price mechanism remains intact, contributing to economic efficiency in the provision of aFRR services. Activation of aFRR can, at times, be circumvented by means of imbalanced energy offsetting facilitated by the International Grid Control Cooperation (IGCC). Through IGCC, TSOs can prevent the simultaneous activation of aFRR in opposing directions within adjacent grid zones. Nevertheless, the practical utility of IGCC is contingent upon the existing cross-border transmission capacity.

In the context of Manual Frequency Restoration Reserves (mFRR) also known as incident reserve, the procurement process differs from aFRR, as it primarily revolves around capacity contracts rather than bid-obligation contracts. The contracted BSPs are required to maintain a certain capacity continuously available, ensuring its readiness for activation by TenneT whenever the need arises. Unlike aFRR, mFRR activations do not rely on a merit order list. Instead, TenneT directly engages the previously contracted capacity to restore frequency balance in response to grid requirements.

2.1.4 Energy Markets

Day-ahead market There are three energy markets that operate sequentially. The first is the day-ahead market. On this market the BRPs trade the energy needed to satisfy their E-program based on the day-ahead forecasts. This markets operates 36 hours to 12 hours in advance of energy delivery, with clearing occurring at noon on the day preceding energy delivery. Bids and offers are harmonised using a merit order system, described in Section 2.1.2, where all market participants receive the same price. All NEMOs maintain price parity, meaning entities such as EPEX and Nordpool share identical day-ahead prices. The trading on the day-ahead market takes place in blocks of 1 hour, resulting in a different price for every hour in the day.

Intraday market The intraday market opens directly after the day-ahead market closes. If features continuous matching of bids and offers up until five minutes before the energy exchange

[2] The electricity grid uses alternating current. The frequency for this alternating current on the Dutch grid is 50 Hz which rises or drops depending on the imbalance of supply and demand on the grid. Big deviations from 50 Hz can result in power loss or a total blackout might occur.

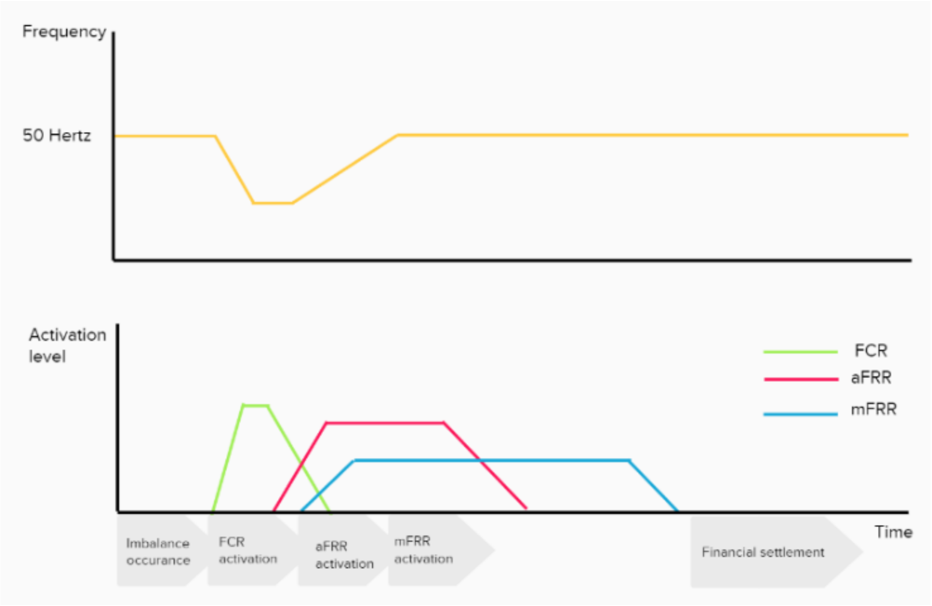


Figure 2: Overview of the balancing process[19]

as opposed to one clearing time like the day-ahead market. This allows BRPs to adjust their order volumes according to up to date forecasts. This market has grown in significance due to the increasing prominence of renewable energy sources, characterised by greater volatility in comparison to conventional energy sources leading to greater imbalances.

Balancing market Lastly, the balancing market. This market works differently to the two spot markets described above. When a change in grid frequency is detected the TSO has to act to prevent outages. The TSO does this by relying on the ancillary services described in Section 2.1.3. For this research we are most interested in the imbalance price as it is the price each BRP has to pay for the difference between there actual energy exchanges and their final E-program. This price is a result of the aFRR and mFRR mechanisms. The paragraph below describes the price mechanism of the imbalance price.

BSPs have the flexibility to make offers up until half an hour before the energy exchange. These offers are combined into a merit order system for each ISP. The results of this merit order system determines the imbalance price for that ISP. This price mechanism is slightly different from the previously described merit order by using 'marginal pricing' in the following way. A merit order is created for the activated aFFR bids of a ISP. The upward regulating imbalance price is the result from the aFRR merit order or, if it is higher, the price for upward incident reserve (mFRR) in that ISP. For the downward regulating imbalance price the mechanism is the exact opposite. When no balancing energy is activated in one of the directions in an ISP, the mid price determines the imbalance price. In cases where a price for either upward or downward regulation is not available, the price is determined based on the price of the preceding ISP for the respective direction. It is important to note that these prices may be a negative value, meaning market participants get paid to consume energy and have to pay to deliver energy back to the grid.

2.2 Use case

Simpl.energy is an energy management company. It specialises in the development of software solutions aimed at assisting their clients in issues related to net congestion and the reduction of energy costs. This software gives insight into the energy flows of their clients and optimise their energy schedules. One potential improvement of their software involves utilising imbalance forecasts to enhance energy consumption and delivery schedules. These schedules govern various actions, such as the charging and discharging of batteries, managing electric vehicle fleets charging, optimising heat pump utilisation, and more. In this context, it is not of utmost importance that the imbalance predictions are absolutely accurate. What truly matters is the ability to predict the occurrences of high and low imbalance prices, enabling the energy schedule to be appropriately adjusted. To illustrate this concept, consider an example: let the typical price level be $y = 100$ at a given moment in time. If at time t=00:15:00, the actual imbalance price was $y_{t=00:15:00} = 2000$, a valid prediction made at t=00:00:00 might be $\hat{y}_{t=00:15:00} = 250$, as this differs enough from the typical price level to trigger an alteration of the energy schedule. However, the ability to also predict the amplitudes correctly does make this use case more profitable.

This use case is contingent on the clients having a specific contract. The most common contract for a company to have is a so called dynamic price contract. Here the company pays the day-ahead prices (every hour a different price) for their energy, meaning the BRP with which the contract is signed is responsible for the imbalance on the market. However, the necessary contract for this use case allows the company themselves to become responsible for their own imbalance, essentially becoming it's own BRP. This means that the company provides day-ahead forecasts to the BRP in quarter hour intervals. The company pays the day-ahead prices for this energy regardless of the actual consumption the next day. If their consumption differs from these forecasts they have to buy or sell their imbalance against the imbalance prices. This mechanism allows the software of simpl.energy to create imbalance when the imbalance price differs from the day-ahead price. If the imbalance price is lower than the day-ahead price it could be profitable to use "too much" energy against the imbalance price by storing it in a battery or shift energy consumption if possible. If the imbalance price is higher than the day-ahead price it could be profitable to use "too little" energy and sell the remaining energy against the imbalance price by discharging a battery or postponing energy consumption if possible.

2.3 Data

The data used in for this research is all gathered from open source platforms[3]. Data is gathered from two sources: The Dutch TSO TenneT [19][4] and European Network of Transmission System Operators for Electricity also known as ENTSOE-E [4]. TenneT publishes data with a one to three minute delay regarding the activation of the ancillary services. This data has 10 features consisting of the date time values, upward and downward IGCC data, activation of upward and downward aFRR, activation of upward and downward mFRR, and the upward-, downward- and mid price.

[3] There is no open source data available for the Dutch intraday market. This is why this data has not been included in this research.

[4] https://www.tennet.org/bedrijfsvoering/Systeemgegevens_uitvoering/Systeembalans_informatie/balansdeltaIGCC.aspx#PanelTabTable

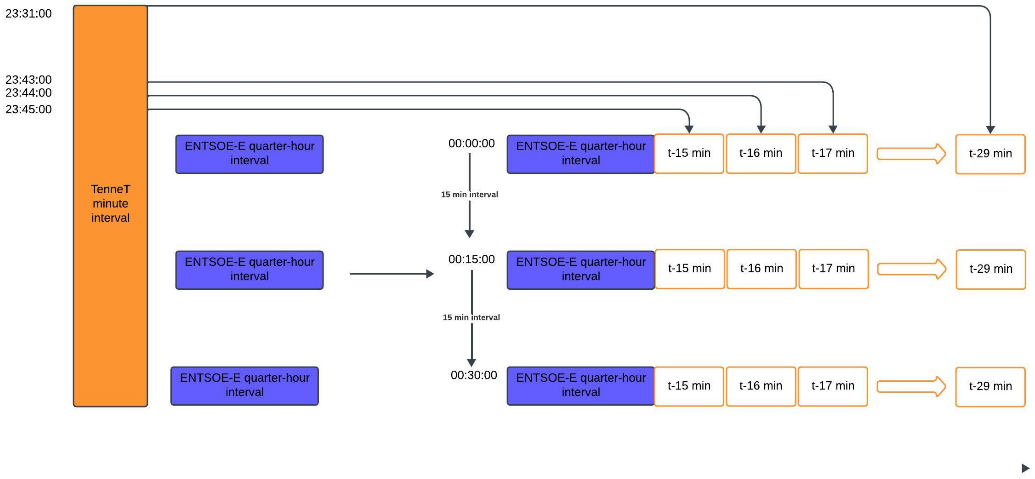


Figure 3: Visual representation of data manipulation method

We gathered the historical data of these variables for the period of 2022-01-01 until 2023-08-31[5]. In addition to this, data from the ENTSOE-E transparency platform is used. From this platform we gather historical data over the same period as the data of TenneT. This data has 7 features and includes the date time values, imbalance prices, day-ahead prices, load data, and generation data, all in quarter hour intervals except the day-ahead prices which are given at an hourly rate. This data is accessible through an API. This study uses the python package `entsoe-py` with an API-key provided by simpl.energy.

Data prepossessing We have developed a way of combining these to sources of data. We use the ENTSOE-E data, with a quarter hour interval, supplemented with TenneT data. This method simulates the model trying to predict a quarter hour in the future. Given that our model would be trying to predict future prices, the time of the target being $t = 0min$, the models can only receive data from $t = -15min$ and earlier. The only exception is the resulting imbalance prices, which serve as the target variables. This time difference is what we simulated with the data structure.

The data obtained through the combination method is visually represented in Figure 3. The resulting data set comprises a row for every quarter hour in the ENTSOE-E data. This includes the target variable. The ENTSOE-E data, excluding the actual imbalance prices, is available a day in advance, and thus maintains alignment with the simulated time difference mentioned earlier. Each row also holds the last two known imbalance prices (the imbalance price from $t - 15$min and $t - 30$min). This data is complemented with TenneT data from $t - 29min$ to $t - 15min$. Thus, a row corresponding to 00:30:00 contains ENTSOE-E data from 00:30:00 along with 6 columns of TenneT data from 00:01:00, followed by 6 columns from 00:02:00, and so on, until 00:15:00, each with 6 columns representing data for every minute.

[5] The start of this project was early September 2023. We chose the last day of the preceding month as the last date of our data, being 2023-08-31.

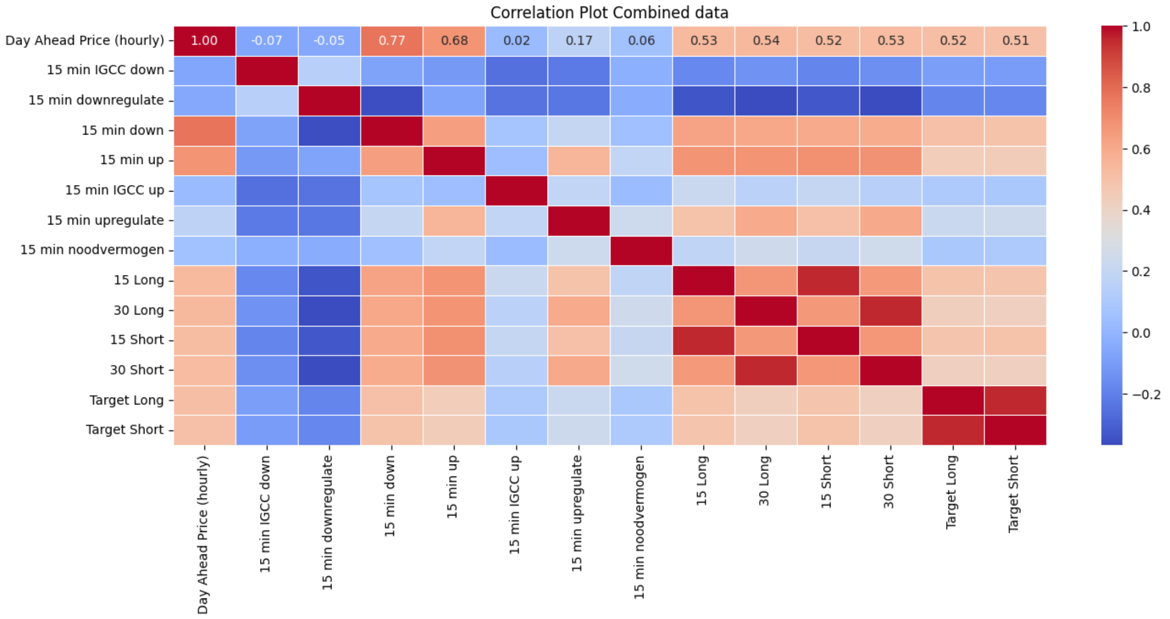
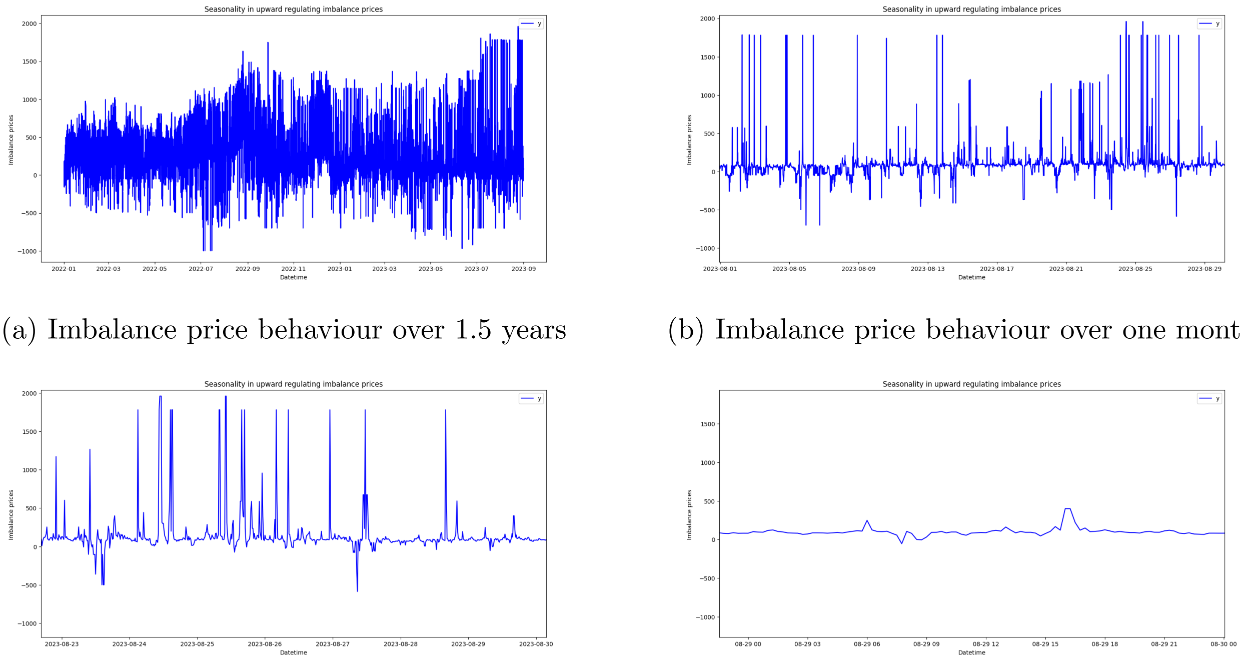


Figure 4: Correlation plot for combined data

Data analysis This data preprosessing results in a data set featuring 115 columns (See Appendix A for more details about each feature.) and 58365 rows all of which have zero missing values. Figure 4 shows the correlation of the most important of these 115 columns. The columns with 15 min before the name refers to the TenneT data from $t - 15$ and the 15 long, 30 long, 15 short and 30 short columns refer to the last know downward and upward regulating imbalance prices. The figure shows that the highest correlation is between the last known imbalance price and the actual imbalance price. This is likely due to the fact that there are mostly gradual changes in imbalance prices. This can be seen in Figure 5 where the price is mostly in the "normal" range. The exceptions to this are the peaks, however this happens much less frequently than an imbalance price in the "normal" range. Figure 4 also shows that the day ahead price has a great influence on the imbalance price. This is in line with expectations as the day-ahead market is the first market to close and thus gives a price indication for the other markets. Lastly, we see that 15 min down and up[6] have a high correlation with the target variables. This is also as expected as these values give an indication as to where the result of the merit order is headed.

The imbalance price displays different behaviour over time, as can be seen in Figure 5. This difference mainly occurs from month to month. As can be seen in Figures 5b, 5c and 5d, the stable price is a relatively straight line. This is not the case from month to month, which is most prevalent in the last 5 months of 2022 as can be seen in the middle of Figure 5a. This shows us that there is no single "normal" imbalance price. We observe another interesting behaviour occurring in this figure. This regards the differences between 2022 and the end of 2023. The peaks in 2023 are greater

[6] This refers to the highest or lowest value of the activated bids up until that point. These values are set equal to the price of the activated bid or, if no bid is activated in this minute, the mid price.



(a) Imbalance price behaviour over 1.5 years

(b) Imbalance price behaviour over one month

(c) Imbalance price behaviour over one week

(d) Imbalance price behaviour over one day

Figure 5: Behaviour of upward regulating imbalance price

than ever before, which highlights the importance of imbalance price forecasting as stated in the introduction. And the peaks are more consistent than ever before, especially in the last two months.

As mentioned above there are two imbalance prices. One for upward regulation and one for downward regulation. We observed that both prices have a similar predictability. In order to keep consistency we use the upward regulating imbalance price throughout this paper.

3 Literature review

In this section previous research is discussed. This review goes over the markets researched, approaches the researches took, models they used and how this research adds to the existing literature.

Traditionally, the predominant focus in electricity price forecasting has been on the day-ahead market [1]. However, there is an increasing body of research in the shorter time frame forecasts of the intraday market and imbalance prices, largely driven by the growing importance of renewable energy sources. It is speculated [20] that this bias towards the day-ahead market may be attributed to the later development of the balancing market and the heightened complexity associated with predicting imbalance prices due to substantial market volatility.

The prevailing trend in electricity price forecasting involves comparing state-of-the-art statistical models with simpler machine learning models and vice versa, resulting in potentially unfair model comparisons as stated in this [7] review. Additionally, it is noteworthy that a substantial portion of the research has relied on relatively small data sets, occasionally limited to just a week's worth of data. Thus, the researches of that review hoped to improve the comparability in field of electricity price forecasting. For this purpose they reviewed the existing literature and proposed benchmark algorithms, data sets to test model performance and propose a set of best practices guidelines.

The bulk of the research on imbalance price forecasting has been conducted within the German and Nordic markets. The German market has gained prominence due to recent expansion and the availability of publicly accessible data. The Nordic market is well-researched, owing to its prevalence of hydro energy, which is a more flexible form of renewable energy, reducing the market's volatility and making it more amenable to forecasting as can be seen in this review [6]. The aim of this review was to benchmark 1 hour ahead and day-ahead imbalance price forecasting models for the Nordic market. They found that day-ahead imbalance price forecasts are not possible as all available information, at that point in time, is reflected in the day-ahead price. There has been some, although very limited, research into the Dutch balancing market. This research [15] modelled the Dutch balancing market aiming to gain insights in the effects of certain features on the imbalance price. This research [20] tried to forecast the Dutch day-ahead price and the imbalance price. It used different AutoRegressive Moving Average with eXogenous inputs (ARMAX) models and Artificial Neural Network (ANN) models for the forecasting. It concluded that ARMAX models performed better than ANN models, both in performance and explainability.

There have been a lot of approaches in the field of electricity price forecasting without consensus on which one is the best. However, most research uses a some linear regression model, an artificial neural network based model or both. This is exemplified by the research into forecasting imbalance prices. This research [10] used the Lasso model to try and predict intraday prices where this research [14] compared the performance of neural based models for predicting intraday prices. The use of a naive model as a benchmark is also common practice in electricity price forecasting. We choose to follow this trend by using a naive model, the Lasso model and the NHITS model, further discussed in Section 5.1. In addition to this, we also decided to use the LightGBM model based on the forecasting experience from simpl.energy.

This research diverges from previous studies in several ways. We exclusively employ state-of-the-art models and utilises a data set spanning over a year. Furthermore, this research is tailored to the practical use case of imbalance price prediction, necessitating a novel perspective on model performance evaluation. This review [2] has shown that Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and RMSE are the most commonly used evaluation metrics in the field of electricity price forecasting. However, their relevance to our specific use case merits further consideration highlighting a gap in the scientific literature. As there has not been any research, to our knowledge, in evaluating forecasts performance for a similar use case. Focusing on the Dutch electricity market, this research aims to contribute to the limited existing research in this context as well.

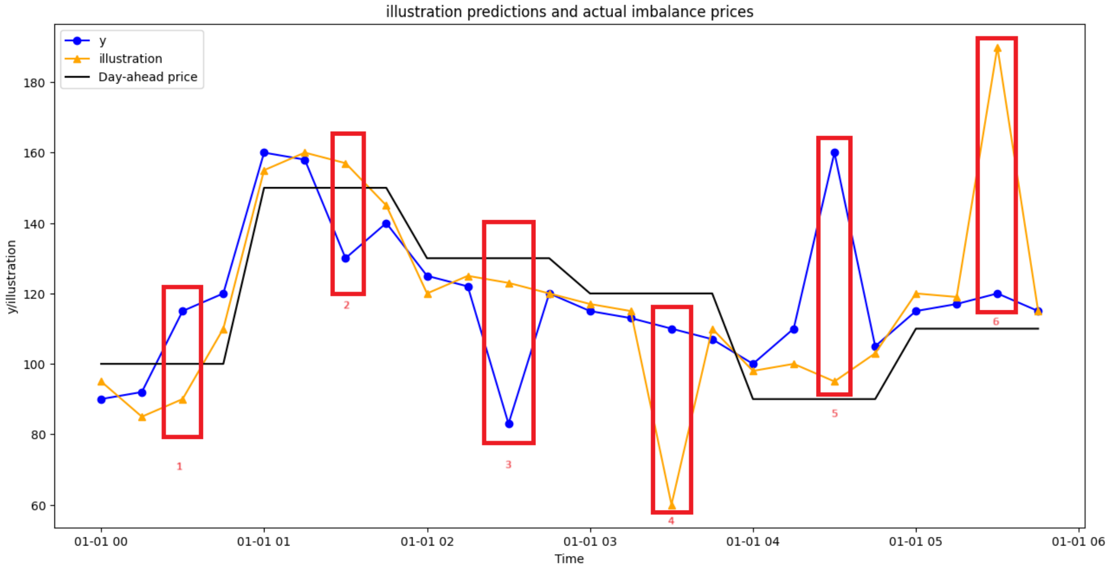


Figure 6: Examples of bad predictions

4 Evaluation measure

In this section we aim to determine which quantitative measure is best suited for our use case. A definition of a good and bad prediction is given, the quality measures which are tested are discussed, the fictitious models used to test the quality measures and how the quality measures are compared to each other is given. Lastly, the results of this experiment are discussed.

In this section we use the following notation: y is the actual imbalance price, $\sum y$ is the sum of all the actual imbalance prices, $\bar{y}$ is the mean imbalance price, $\hat{y}$ is the predicted imbalance price, $\sum \hat{y}$ is the sum of all the predicted imbalance prices, $\sum y \cdot \hat{y}$ is the sum of the product of all the imbalance prices and predicted imbalance prices, $\sum y^2$ is the sum of all the squared values of y, $\sum \hat{y}^2$ is the sum of all the squared values of $\hat{y}$, N is the size of the data set and x is the date time value converted to integers (2022-01-01 00:00:00 is 1, 2022-01-01 00:15:00 is 2 etc.).

4.1 Good/Bad predictions

For this research it is useful to define what constitutes as a good and what constitutes as a bad prediction in the use case described in Section 2.2. We outline three cases that constitutes bad predictions, if none of these cases are true we consider the prediction to be good. These cases are visualised in Figure 6 where every bad example is highlighted by a red box. All of the cases are based on the imbalance price differing from the day-ahead price. This is because the contract outlined in the use case entails that the client has bought a specific amount of energy against the day-ahead price for each quarter hour of the day. Imbalance price predictions would allow to use this energy optimally in such a way that the shortage or surplus created can be traded against the imbalance price.

The first bad case is where a prediction is on the wrong side of the day-ahead price compared to the real imbalance price. This case is exemplified by red boxes 1 and 2 in Figure 6. This costs money because more expensive energy will be used as a result of this wrong prediction. The notation we use to indicate this case is: $sign(\hat{y} - DA) \neq sign(y - DA)$.

The second bad case is where the prediction shows a peak where there is no peak. We define a peak by a value being outside the range of $[DA - 100,\ DA + 100]$ (from now on the term "peak" is used to indicate both a high and a low point). This case is exemplified by the fourth and sixth red box in Figure 6. Such a prediction gives a wrong profit estimation (e.g. charging a battery 2 MWh where 1 MWh would have been optimal) and thus saves less money than would be possible with a correct peak prediction. The notation we use to indicate this case is: $\hat{y} \notin [DA - 100, DA + 100], y \in [DA - 100, DA + 100]$.

The third bad case is where the prediction shows no peak where there is a peak. Again, we define a peak by a value being outside the range of $[DA - 100,\ DA + 100]$. This case is exemplified by the third and fifth red box in Figure 6.This gives a wrong profit estimation (e.g. only charging a battery 1 MWh where 2 MWh would have been optimal) and thus saves less money than would be possible with a correct peak prediction. The notation we use to indicate this case is: $\hat{y} \in [DA - 100, DA + 100], y \notin [DA - 100, DA + 100]$.

4.2 Quality measure

For our specific use case it is not so straight forward to asses model performance with a quantitative measure due to the fact that the main concern is the predictions of peaks and to a lesser degree the actual magnitude of those peaks. This is the reason the more standard error-based quality measures, such as RMSE, R-squared or MAPE, might not give the most accurate representation of the quality of a model. The reason for this is that an error in a bad prediction can be quite small[7], resulting in better scores than the actual model performance. This is why we study 4 quality measures for evaluating model performance alongside RMSE, R-squared and MAPE. The first two are known quality measures not commonly used in the realm of electricity price forecasting. The final two are new quality measures proposed by us.

RMSE (Root Mean Square Error) is the first commonly used error based quality measure. It is a statistical measure that gauges the average magnitude of the errors between predicted and observed values. The formula for RMSE is given in Equation 1. Lower RMSE values signify better predictive accuracy.

$$RMSE = \sqrt{\frac{\sum_{i=0}^{N-1} (y_i - \hat{y}_i)^2}{N}} \tag{1}$$

[7] A prediction on the wrong side of the day-ahead price might only have an error of 50. This is a relatively small error resulting in a decent error-based score, but still costs money as stated in Section 4.1.

R-squared is the second commonly used error based quality measure. It is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model. The formula for R-squared is given in Equation 2. The values range mostly from 0 to 1, where 0 indicates that the model does not explain any variance, and 1 indicates perfect explanation. A negative value indicates that the model performs worse than using the average as a predicted value.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{2}$$

The last commonly used error based quality measure is Mean Absolute Percentage Error (MAPE). It measures the average percentage difference between predicted and observed values. It is calculated by taking the absolute percentage difference for each data point, summing these differences, and then averaging them over the entire data set. The formula for MAPE is given in Equation 3. MAPE is expressed as a percentage and provides a clear indication of the model's performance in terms of the relative magnitude of errors across all predictions.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{3}$$

The first known quality measure that is not often used in electricity price forecasting is Pearson correlation. Correlation is sensitive to outliers when outliers are present in both variables (in our case y and $\hat{y}$) as shown in this [5] paper. So our reasoning for choosing correlation as a quality measure is that this sensitivity should mean that it is good at punishing false and missed peaks, which are two of the three bad cases as discussed in Section 4.1, and thus be a suitable quality measure for our use case.

$$correlation = \frac{N(\sum y \cdot \hat{y}) - (\sum y)(\sum \hat{y})}{\sqrt{(N\sum y^2 - (\sum y)^2)(N\sum \hat{y}^2 - (\sum \hat{y})^2)}} \tag{4}$$

The second quality measure that is not often used in electricity price forecasting is the relative MAE (rMAE). This quality measure was deemed as being the best practice for model evaluation with regards to electricity price forecasting by [7] . This measure devides the MAE of a model by the MAE of a naive model. The naive model we use, which is explained in more detail in Section 5.1, predicts the actual imbalance price of the exact same time the day before. A rMAE < 1 means that the model performs better than the benchmark method, and a rMAE > 1 means that the model does not outperform the benchmark model. The formula for rMAE is given by Equation 5.

$$rMAE = \frac{\frac{1}{N}\sum|y - \hat{y}|}{\frac{1}{N}\sum|y - \hat{y}_{naive}|} \tag{5}$$

The first proposed quality measure punishes the bad predictions a model can make. Predicting on the wrong side of the day-ahead price is the worst case as described in Section 4.1 as it actually costs money. The other two cases are less bad as they do not cost money and are thus punished less harshly. Lastly, to take magnitude of the error into account, we add a scaled absolute error. Scaling by a factor of 1000 ensures that most error terms will be between 0 and 1, making sure the magnitude stays a secondary concern. The formula for the punishment is given in Equation 6. These punishments are added up to total the row punishment. The final score is the average of all these row punishment and is given by Equation 7.

$$\begin{aligned} punishment_i = & 2 \cdot \mathbb{1}_{\text{sgn}(\hat{y}_i - DA) \neq \text{sgn}(y_i - DA)} + \mathbb{1}_{\hat{y}_i \notin [DA-100, DA+100], y_i \in [DA-100, DA+100]} \\ & + \mathbb{1}_{\hat{y}_i \in [DA-100, DA+100], y_i \notin [DA-100, DA+100]} + \frac{|y - \hat{y}|}{1000} \end{aligned} \tag{6}$$

$$punishment\_score = \frac{\sum_{i=1}^{N} punishment_i}{N} \tag{7}$$

The last proposed quality measure uses slope values instead of the actual values. We calculate the RMSE with the y and $\hat{y}$ values replaced by the slope between each y and $\hat{y}$ data point as can be seen in Equation 8.The idea is that the error in the slope should give a better indication if the models predicted the peaks correctly than the actual error of the prediction.

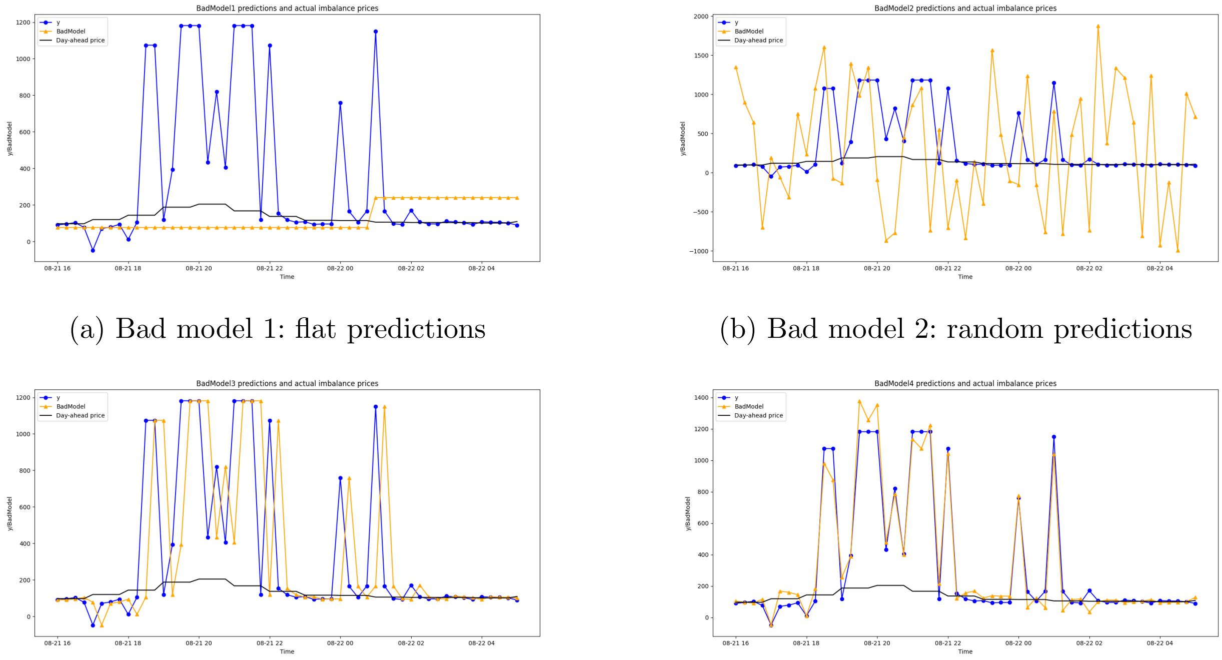
$$slopeRMSE = \sqrt{\frac{\sum_{i=0}^{N-1} (\frac{y_i - y_{i-1}}{x_i - x_{i-1}} - \frac{\hat{y}_i - \hat{y}_{i-1}}{x_2 - x_1})^2}{N}} \tag{8}$$

4.3 Fictitious model predictions

In order to evaluate the effectiveness of the previously mentioned quality measures we create eight distinct fictitious models to generate our data sets, which can be categorised into three classes: “Bad”, “Medium” and “Good”. These fictitious models are developed together with the expert from simpl.energy to capture the expert feeling of a bad, medium and a good model. Thus, the better a quality measure is a separating these classes, the more suitable the quality measure is for our use case. For a better understanding of the fictitious models we provide visualisations in figures 7, 8 and 9, where the black line shows the day-ahead price, the orange line the predictions, and the blue line the real imbalance price.

The fictitious “Bad” models are created to resemble model outputs we consider bad. The first “Bad” model yields flat predictions, as illustrated in Figure 7a, aligning with the mean value of the actual prices. The mean is updated every days worth of data with the mean of the day before, to account for changes happening over time as mentioned in Section 2.3. This model is considered bad because this model is not able to capture any changes in the market. The second “Bad” model, depicted in Figure 7b, generates random predictions spanning the range between the highest and lowest values of the actual imbalance prices. This model is bad because it does not reflect the market at all. The third “Bad” model illustrated in Figure 7c offers delayed predictions by forecasting real values from a previous time step with a shift of X steps. This results in a model that is not bad, but it does not allow the company to get ahead of the market. This research uses a shift of 1, meaning the model always has a correct prediction, but one time point too late. Lastly, the fourth “Bad” model is illustrated in Figure 7d and follows Equation 9. One of the worst cases, as stated previously in Section 2.2, is when a model would predict on the wrong side of the day-ahead price. The typical quality measures do not pick up on these worst predictions. This is why "Bad" model 4 is important to add. This model is a good model, predicting real imbalance price with Gaussian noise when the imbalance price is far away from the day-ahead price, but if the real value is around the day-ahead price it predicts on the other side of the day-ahead price with the exact same distance to the day-ahead price. We observed that most real algorithms underpredict peaks and not overpredict. We use Gaussian noise with a mean of 0.8 and a standard deviation of 0.1 to capture this behaviour.

$$\hat{y}_{\text{bad4}} = \begin{cases} y + 2 \cdot (DA - y), & \text{if } y \in [DA - 100, DA + 100] \\ y \cdot \mathcal{N}(0.8, 0.1), & \text{if } y \notin [DA - 100, DA + 100] \end{cases} \tag{9}$$



(a) Bad model 1: flat predictions

(b) Bad model 2: random predictions

(c) Bad model 3: delayed predictions

(d) Bad model 4: wrong side of zero

Figure 7: The fictitious Bad models

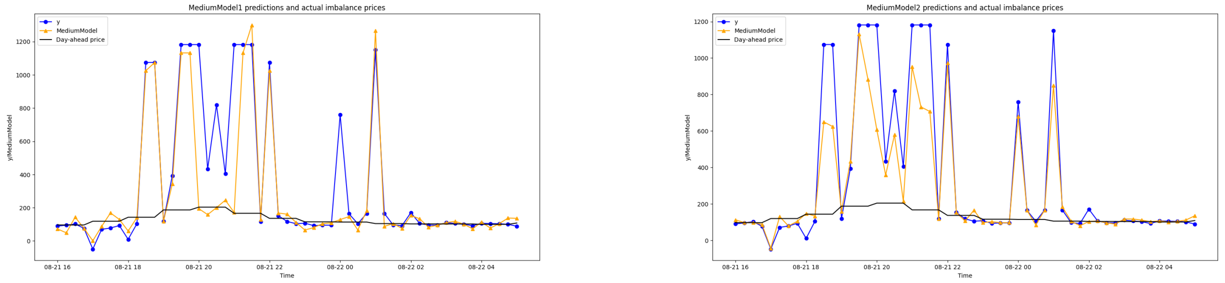
The two "Medium" models predict some of the peaks correctly while sometimes missing them as can be seen in Figure 8. Predictions are generated based on the actual imbalance price. The actual values, for peaks, addition/subtraction and probability, in these two fictitious models are logically chosen and further fine-tuned in collaboration with the expert until the right feeling was captured.

The first "Medium" model follows Equation 10 and is visualised in Figure 8a. This model uses the following notation: P(X) meaning a probability of X and $x \sim U[m, n]$ meaning a random integer between m and n with a uniform distribution. This model works as follows: when there is no peak (Note that the definition of a peak, which was $y \notin [DA - 100, DA + 100]$, has changed for just this model to be a bit more sensitive to $y \notin [DA - 75, DA + 75]$.) we add a small error to the actual imbalance price. When there is a peak there is a 50% chance of adding a small deviation from the day-ahead price and a 50% chance of adding a deviation from the actual imbalance price. This simulates missing half of the peaks, capturing half and having random behaviour around the day-ahead price.

$$\hat{y}_{\text{medium1}} = \begin{cases} y + x \sim U[-50, 50], & \text{if } DA - 75 \leq y \leq DA + 75 \\ \begin{cases} DA + x \sim U[-50, 50], & \text{with } P(0.5) \\ y - 50, & \text{with } P(0.5) \end{cases}, & \text{if } y > DA + 75 \\ \begin{cases} DA + x \sim U[-50, 50], & \text{with } P(0.5) \\ y + 50, & \text{with } P(0.5) \end{cases}, & \text{if } y < DA + 75 \end{cases} \tag{10}$$

The second "Medium" model uses Equation 11 and is visualised in Figure 8b. The second model uses the same notation as the first "Medium" model: P(X) meaning a probability of X and $x \sim U[m, n]$ meaning a random integer between m and n with a uniform distribution. This model works as follows: where there is no peak (Note that the definition of a peak, which was $y \notin [DA - 100, DA + 100]$, has changed for just this model to be a bit less sensitive to $y \notin [DA - 200, DA + 200]$.) we add a small error to the actual imbalance price. When there is a peak there is a 30% chance of adding a small random deviation to the actual imbalance price and a 70% chance of adding a big random deviation. This model simulates a different randomness to the behaviour of "Medium" model 1.

$$\hat{y}_{\text{medium2}} = \begin{cases} y + x \sim U[-30, 30], & \text{if } DA - 200 \leq y \leq DA + 200 \\ \begin{cases} y - x \sim U[0, 30] \cdot 10, & \text{with } P = (0.3) \\ y - x \sim U[0, 30] \cdot 25, & \text{with } P = (0.7) \end{cases}, & \text{if } y > DA + 200 \\ \begin{cases} y + x \sim U[0, 30] \cdot 10, & \text{with } P = (0.3) \\ y + x \sim U[0, 30] \cdot 25, & \text{with } P = (0.7) \end{cases}, & \text{if } y < DA - 200 \end{cases} \tag{11}$$



(a) Medium model 1

(b) Medium model 2

Figure 8: The fictitious Medium Models

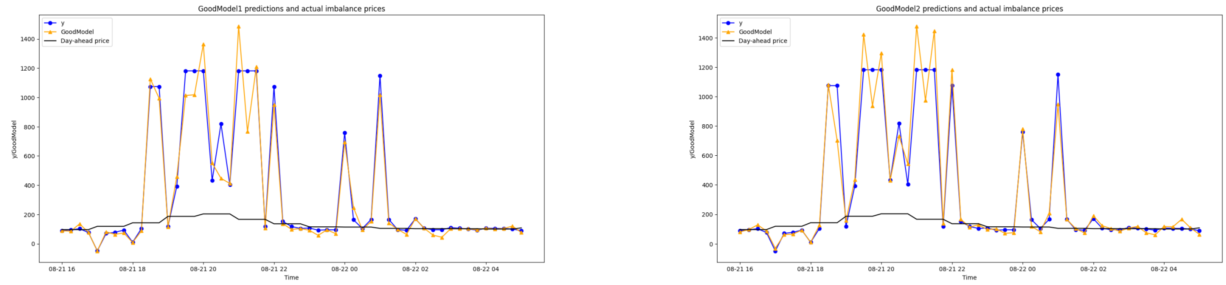
To maintain satisfactory performance for medium models, a correction is introduced, as defined by Equation 12. This correction is applied to all the cases of both "Medium" models where a resulting prediction is on the wrong side of the day-ahead price. It corrects 70 % of these cases by multiplying the actual imbalance price by a random factor between 0.9 and 1.1 or setting it equal to the day-ahead price if the multiplication is not successful at correcting the mistake. The other 30 % is left as a prediction on the wrong side of the day-ahead price.

$$correction(\hat{y}) = \begin{cases} \begin{cases} max(y \cdot x \sim U[0.9, 1.1], DA), & \text{if } y > DA \\ min(y \cdot x \sim U[0.9, 1.1], DA), & \text{if } y < DA \end{cases}, & \text{with } P(0,7) \\ \hat{y}, & \text{with } P(0.3) \end{cases} \tag{12}$$

The two good models follow the imbalance prices almost perfectly with some minor errors as can be seen in Figure 9. Each prediction is based on the actual imbalance price with some added Gaussian noise. These models also ensure that every prediction is on the correct side of the day-ahead price, which is imperative of a good model. The first "Good" model uses Equation 13 and is visualised in Figure 9a. "Good" model 1 multiplies with a factor that is normally distributed with $\mu = 0.9$ and $\sigma = 0.2$. The second "good" model uses Equation 14 and is visualised in Figure 9b."Good" model 2 instead multiplies with a factor that is normally distributed with $\mu = 1.0$ and $\sigma = 0.3$. This second model results in bigger deviations from the actual imbalance price. This makes the comparison with "Good" model 1 and the other classes more interesting.

$$\hat{y}_{\text{good1}} = \begin{cases} max(y \cdot \mathcal{N}(0.9, 0.2), DA), & \text{if } y \geq DA \\ min(y \cdot \mathcal{N}(0.9, 0.2), DA), & \text{if } y < DA \end{cases} \tag{13}$$

$$\hat{y}_{\text{good2}} = \begin{cases} max(y \cdot \mathcal{N}(1.0, 0.3), DA), & \text{if } y \geq DA \\ min(y \cdot \mathcal{N}(1.0, 0.3), DA), & \text{if } y < DA \end{cases} \tag{14}$$



(a) Good model 1

(b) Good model 2

Figure 9: The fictitious Good Models

4.4 Determining best evaluation measure

In order to determine which quantitative measure is the most effective for the context of our study, an analysis is conducted. This analysis involves the implementation of all models outlined in Section 4.3. The performance of each model is evaluated using the metrics outlined in Section 4.2. The measures are compared on their ability to differentiate all classes of models. Additionally,

Table 2: Quality measures for fictitious models on 1.5 years worth of data

Model	RMSE	R-squared	Correlation	MAPE	MAE	rMAE	Punishment score	Slope RMSE
Bad Model 1	205.802039	0.190312	0.448527	6.666556×10^{13}	123.073654	0.786897	1.353473	186.978010
Bad Model 2	934.439363	-15.692461	-0.010846	3.073442×10^{14}	790.444273	5.051443	2.505885	1222.741364
Bad Model 3	186.855971	0.332529	0.666265	6.576996×10^{13}	90.664528	0.579602	0.762757	301.702899
Bad Model 4	80.005798	0.877634	0.941835	8.392357×10^{12}	59.491390	0.380194	1.476444	87.460310
Medium Model 1	109.781118	0.769605	0.877927	1.095729×10^{13}	51.404758	0.328555	0.445216	136.300083
Medium Model 2	113.533848	0.753584	0.869569	1.929100×10^{12}	58.248866	0.372224	0.474893	135.112938
Good Model 1	59.523363	0.932268	0.967597	1.510098×10^{-1}	32.339916	0.206855	0.131236	78.869959
Good Model 2	79.713765	0.878526	0.945851	1.912440×10^{-1}	41.537253	0.265693	0.163530	112.043240

we will investigate whether the length of the data sets impacts the quality measures by using the full 1.5 years worth of data, only the last month worth of data and the last 400 records[8]. This investigation is crucial because some quality measures may exhibit averaging effects over larger data sets.

The results obtained from these evaluations are organised into Tables 2, 3 and 4. This systematic approach aims to yield insights into the strengths and weaknesses of the various quantitative measures, ultimately guiding the identification of the most suitable metric for gauging prediction accuracy in the specific context of imbalance price prediction.

Behaviour on full data set Most of the quality measures are quite good at differentiating the Bad, Medium and Good models (i.e. all the scores of the bad models are worse than those of the medium models, which are worse than the good models), but Bad model 4 is the exception. For example, the RSME score in Table 2 of Bad model 4 is around 80, where the good models have a RSME score of 60 and 80. Based on these values Bad model 4 resembles a good model. This is a big problem, as Bad model 4 has a lot of occurrences where the prediction is on the wrong side of the day-ahead price and thus will cost a lot of money. However, this is in line with our expectations about error based quality measures, further proven by the same tendencies for R-squared and MAPE. Correlation and Slope RMSE exhibit the same inability to separate Bad model 4 from the Medium and Good models. This opposes our reasoning for choosing as them as a quantitative measure, as we expected them to better capture the prevalence of false and missing peaks. But there are 3 quality measures that are able differentiate all Bad models from the Medium models and the Good models. Those being the MAE, rMAE and the Punishment score. The Punishment score has the biggest relative difference between the bad and the medium models. The highest punishment score for a medium model is 0.47 and the lowest bad model score 0.76. For the MAE and rMAE respectively those scores are 58.25 compared to 59.49[9] and 0.37 compared to 0.38. Thus, our proposed punishment score is considered the best when using the entire data set.

Impact of length of data set It is also important to discuss whether the length of the data sets have an impact on the effectiveness of the quality measures. From this experiment we can conclude that the length of the data set does have an impact. In Table 3 can be seen that there

[8] The real models used in the second experiment described in Section 5 yield 400 predictions.

[9] The magnitude of the MAE scores is roughly 50 times higher. If we scale the difference down to the same magnitude as punishment score and rMAE we get a difference in the order of magnitude 10^{-2}.

Table 3: Quality measures for fictitious models on 1 month worth of data

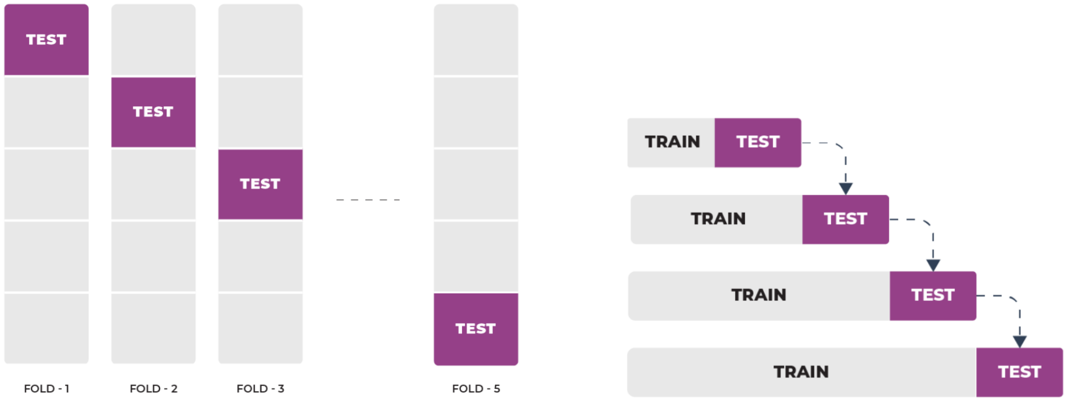
Model	**RMSE**	**R-squared**	**Correlation**	**MAPE**	**MAE**	**rMAE**	**Punishment score**	**Slope RMSE**
Bad Model 1	251.584632	-0.017040	0.097908	1.449168	110.382845	0.742888	1.313437	254.916480
Bad Model 2	968.042299	-14.057700	-0.019948	14.827653	810.171750	5.449403	2.711092	1234.720816
Bad Model 3	255.069412	-0.045410	0.477296	0.898100	84.561983	0.554267	0.635451	412.038597
Bad Model 4	76.668442	0.905550	0.958196	1.283568	57.505675	0.390408	1.714523	78.906979
Medium Model 1	138.908628	0.689952	0.831020	0.607799	43.548633	0.296103	0.414722	182.532221
Medium Model 2	91.987547	0.864035	0.938097	0.680119	40.732175	0.274522	0.388895	108.892509
Good Model 1	54.066271	0.953030	0.978838	0.155853	21.318261	0.142606	0.044243	71.139125
Good Model 2	83.959015	0.886733	0.946053	0.195036	29.029519	0.195795	0.066239	120.966858

Table 4: Quality measures for fictitious models on 400 records

Model	**RMSE**	**R-squared**	**Correlation**	**MAPE**	**MAE**	**rMAE**	**Punishment score**	**Slope RMSE**
Bad Model 1	141.890667	-0.018113	-0.012685	1.165134	54.181621	0.723561	1.330863	163.770859
Bad Model 2	938.644686	-43.554402	0.011588	16.718399	796.431225	11.705281	3.647862	1212.826594
Bad Model 3	163.790373	-0.356641	0.321722	0.626960	44.982200	0.625093	0.459964	266.668242
Bad Model 4	69.351631	0.756779	0.918750	0.867418	53.548349	0.751267	1.914597	56.297334
Medium Model 1	102.931376	0.464223	0.681485	0.603339	32.790627	0.420833	0.398167	125.527230
Medium Model 2	62.727746	0.801021	0.900797	0.912996	29.277018	0.434466	0.443554	78.841077
Good Model 1	47.110095	0.887768	0.951206	0.156637	17.503144	0.247272	0.057506	65.439956
Good Model 2	34.815763	0.938703	0.978872	0.200587	19.213874	0.281019	0.068428	51.274695

are 4 quality measures with the ability to separate all classes when using only a months worth of data: MAPE, MAE, rMAE, and Punishment score. With an even smaller data set, given by Table 4, we again observe that the Punishment score, MAE and rMAE have the ability to separate all classes. On only 400 records we see that the Punishment score has the smallest relative difference, when compared to the MAE and rMAE, instead of the biggest on the entire data set. The lowest Punishment score for a bad model is 0.46 for Bad model 3, where the highest medium model scores 0.44. The MAE and rMAE have a bigger difference of 45.0 versus 32.8 and 0.63 versus 0.43.

Best evaluation method The only quality measures consistently able to separate the classes over all lengths of data sets are the Punishment score, MAE and the rMAE. We consider our proposed quantitative measure, the Punishment score, the best performing despite it having the smallest relative difference on the small data set. The reasoning behind this is that the Punishment score has trouble defining Bad model 3 as a bad model (it is able to do so, only with very small margins). However, it is very good at labelling Bad model 4 as being bad. This model represented the core problem that lead to this research. The inability of most quantitative measures to represent model performance when the magnitude of the errors are not the most important thing. For instance, Table 4 shows that the RMSE has a score of 69.35 for Bad model 4 versus the highest medium score of 102.93. Bad model 4 has a Punishment score of 1.9 versus the highest medium score being 0.44. For the rMAE, which is also able to separate all the classes all of the times, this difference is only 0.63 versus 0.43. Even though we do not consider the rMAE the best, we do believe it should be used alongside the Punishment score as it also is able to separate all the classes all of the time and it is proposed as the new electricity price forecasting standard by [7].



(a) k-fold cross-validation

(b) Rolling cross-validation

Figure 10: Comparison of cross-validation methods[17]

Discussion It is important to note that these scores do not serve as a baseline. Currently there is no indication of what performance is feasible with applications of real forecasting models. This means that we cannot conclusively say that a model is only good if it has a Punishment score lower than 0.1 and that we should optimise until this threshold is met. These scores are only used for comparison in the quantitative measure experiment. Further research, the experiment described in Section 5 and other future research, should give more insight into the actual feasibility of model performance.

The MAPE is not a very suitable quality measure for this use case as the imbalance prices sometimes approach zero. This can result in huge MAPE scores as we divide the error by the imbalance price as shown in Equation 3. A further look at the data shows that there are no occurrences of the imbalance price approaching zero in the last month worth of data. This is the reason only Table 2 shows extremely high values for the MAPE and Tables 3 and 4 do not.

5 Forecasting Algorithms

In this section we aim to asses model performance of real forecasting algorithms and test our findings of Section 4 on these algorithms. First the used models used are given. We discuss the performance of the models determined by the expert and lastly our findings on the quantitative measures.

Please note that the main focus lies on the evaluation of model performance, not to find the most optimal model for imbalance price prediction. As the forecasting of imbalance prices merits it's own research which would take a lot of time and effort, we chose not to optimise for model performance. This means that we did not perform hyper-parameter tuning, we focused on well implemented models during model selection and we did not enhance our data with closed-source data.

5.1 Models

cross-validation In this research, we use the `cross_validation` function from the mlforecast[12] and neuralforecast[13] packages. This function regards time series cross-validation, not to be confused with “normal” cross-validation. All “normal” cross-validation versions follow the workings of the so-called k-fold cross-validation to some degree. Thus we use the k-fold cross-validation to explain the workings of a “normal” cross-validation. The k-fold cross-validation[9], visualised in Figure 10a, works by dividing the data set into k random, equal parts called folds. One of these folds is chosen as the test set, the remaining folds serve as the training set. This is repeated k times until every fold has served as a test set, each itteration saving the performance estimates. The final performance evaluation using k-fold cross-validation is the mean of the performance estimates of each iteration.

However, with time series forecasting, the order of the data set is of importance. Thus, we need to use time series cross-validation, also called rolling cross-validation. The workings of the rolling cross-validation is best described using Figure 10b. Instead of randomly shuffling data, it sequentially divides the time series into training and testing sets. This division is called a window. The size of the test set is determined by the so called forecast horizon. The model is training set is earlier data and test set is later data. After each iteration, the training set is expanded to include the test set of the previous window, the test set is moved forward in time, and the performance estimate is save. This process repeats for multiple rounds until the predetermined number of windows is met. Similar to the k-fold cross-validation, the final performance evaluation is the mean of the performance estimates of each iteration. This rolling cross-validation allows us to simulate an in production algorithm that gets updated with every prediction or every few predictions[10]. In this research we use a forecast horizon of 1. The reason for this is that [1] showed that performance significantly drops the further ahead you try to predict imbalance prices.

models The naive model serves as a baseline to compare other models to as well as showing the complexity of the problem we are trying to solve. The naive model works as follows: A prediction for $[day = 2, t = 00:00:00]$ will be equal to the known imbalance price of the day before, being the imbalance price of $[day = 1, t = 00:00:00]$. This differs from the approach form [11]. This research used the intraday prices as a naive predictor of imbalance prices. This naive model performed well. However, the intraday prices are not publicly available for the Dutch market, so this is not an option for us. The predictions of the naive model are cut to the same size as the other models to ensure comparability.

The second model is a statistical regression model. Least Absolute Shrinkage and Selection Operator (Lasso) is first introduced in [18]. This model performs both variable selection and regularisation which improves the accuracy over other statistical regression models. This variable selection, where the Lasso model set coefficients it does not find interesting to zero, is especially interesting for our research as we use a data set with a large amount of features. Our research uses the implementation of this model from the scikitlearn package [16]. The predictions are generated via the `cross_validation` function from the mlforecast[12] package using a forecast horizon of

[10] It is currently not known what updating frequency is feasible with a model in production. It greatly depends on the training time of the implemented model and the computing power available

1 and 400 windows, thus resulting in 400 predictions. All other hyper-parameters are set to the default.

The third model is a tree based machine learning model. LightGBM is a model originally developed by Microsoft. LightGBM is short for light gradient-boosting machine and is a tree based learnign algorithm like the better known XGBoost. Advantages of this model over other tree based learning algorithms are that it has a faster training speed, lower memory usage, better accuracy, supports parallel, distributed and GPU learning and it is capable of handling large-scale data. This speed and low memory usage makes it an interesting machine learning model as an algorithm in this use case is run at least every quarter hour. This research uses the regressor of the lightGBM package[8] together with the mlforecast[12] package from Nixtla. The predictions are generated via the `cross_validation` function from the mlforecast package using a forecast horizon of 1 and 400 windows, thus resulting in 400 predictions. All other hyper-parameters are set to the default.

Lastly, we use a neural network model. Neural Hierarchical interpolation for Time Series also known as N-HiTS was first proposed in [3]. NHITS is an improvement on the NBEATS algorithm with a improved accuracy while being times faster and using less memory. The speed and memory makes NHITS an interesting neural network based model. This algorithm has later been implemented into a python package by Nixtla called neuralforecast [13]. The predictions are generated via the `cross_validation` function from the neuralforecast package using a forecast horizon of 1 and 400 windows, thus resulting in 400 predictions. All other hyper-parameters are set to the default.

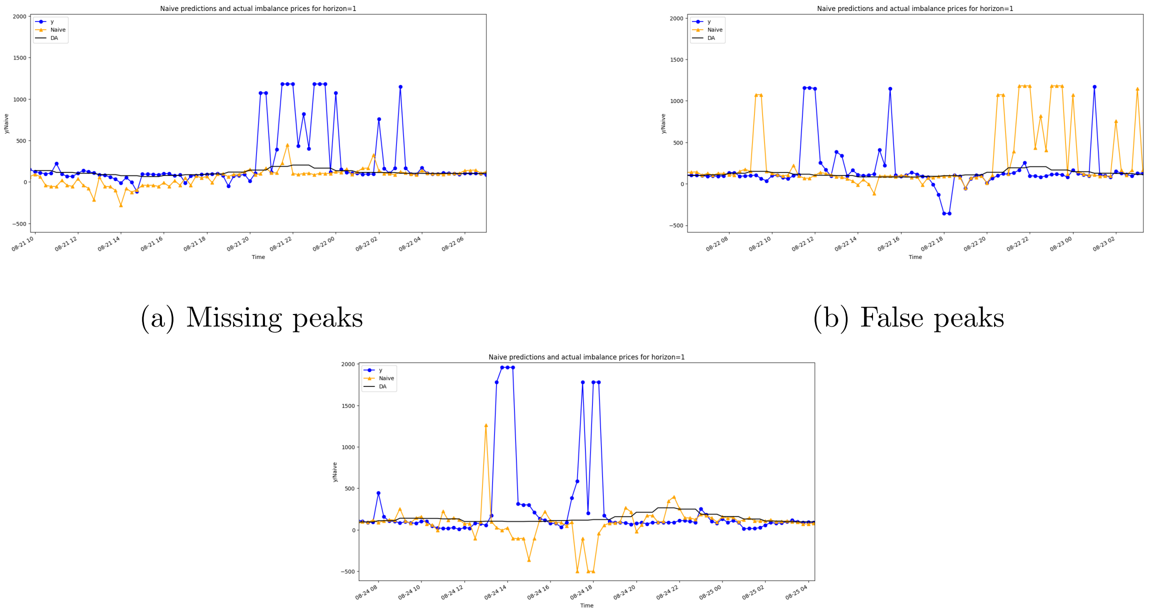
5.2 Model performance determined by the expert

This experiment has two goals[11]. The first goal is to compare model performance in the realm of real imbalance price forecasting. This is be done by visually comparing the models together with the expert from simpl.energy. This comparison is conducted in an in-person meeting with the expert from simpl.energy. We cover all 400 predictions of all the four models, looking at the errors made, the errors in relation to the day-ahead price and the good predictions. From this meeting We gather the feedback regarding model performance and the ranking given by the expert. This feedback and ranking is discussed below. Each figure used in this section is meant to exemplify the remarks made by the expert.

Naive model The performance of the Naive model is deemed very bad by the expert from simpl.energy. The reason for this is that it misses a lot of peaks, as can be seen in Figure 11a. It predicts peaks where there are none, as can be seen in Figure 11b, and it has some big errors on the wrong side of the day-ahead price, as can be seen in Figure 11c. This lead us to conclude that there are no real day to day similarities, as the peaks from one day do not line up with the peaks from the day before. This is in line with our expectations as the balancing market is very volatile.

Lasso The Lasso model seems to have relatively random predictions. We state this because the predicted peaks seem to have no correlation with the peaks of the real imbalance price. This behaviour is best captured in Figure 12a. We also see that when prices are relatively steady the

[11] The second goal is discussed in Section 5.3.



(a) Missing peaks

(b) False peaks

(c) Big errors on wrong side of the DA

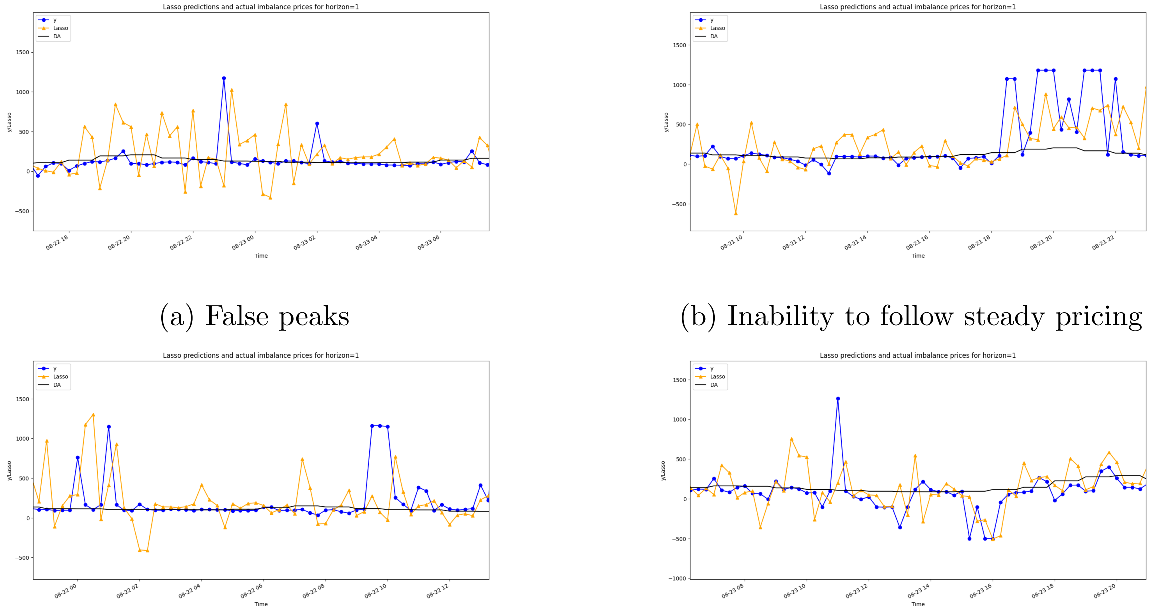
Figure 11: Problems of naive model

Lasso model, more than all other models, still predicts large deviations as can be seen in Figure 12b and 12c. Lasso also predicts on the wrong side of the day-ahead price as can be seen in Figure 12d.

LightGBM The lightGBM model has similar behaviour to some of the bad models from Section 4.3. First of all, it displays delayed predictions as can be seen in Figure 13a, which resembles Bad model 3. Next to this it also displays a lot of random peaks, as shown in Figure 13b, which sometimes are on the wrong side of the day-ahead price, as can be seen in Figure 13c, which is in line with Bad model 2. However, this behaviour is less erratic sometimes as can be seen in Figure 13d.

NHITS The output of the NHITS model is very similar to that of Bad model 3 from Section 4.3, meaning that the NHITS model follows the real imbalance with a delay which is exemplified by Figure 14a and 14b. A possible explanation could be that there is not enough correlation in the open source data used, making the last known imbalance price a disproportional influence on the prediction made. The NHITS model also portrays characteristics of bad model 1 from Section 4.3. This can be seen in Figure 14c, where the predictions are mostly flat, and thus misses a lot of peaks in the imbalance price. It also has some minor errors on the wrong side of the day-ahead price which can be seen in Figure 14d.

The first goal is to compare the performance of all the models. All the models are deemed to be not fit for use. This is because all the models display a big amount of predictions on the wrong side of the day-ahead price. The Naive model is the worst, as it has some really big errors on the wrong side of the day-ahead price and thus will cost a lot of money. The lightGBM and Lasso model



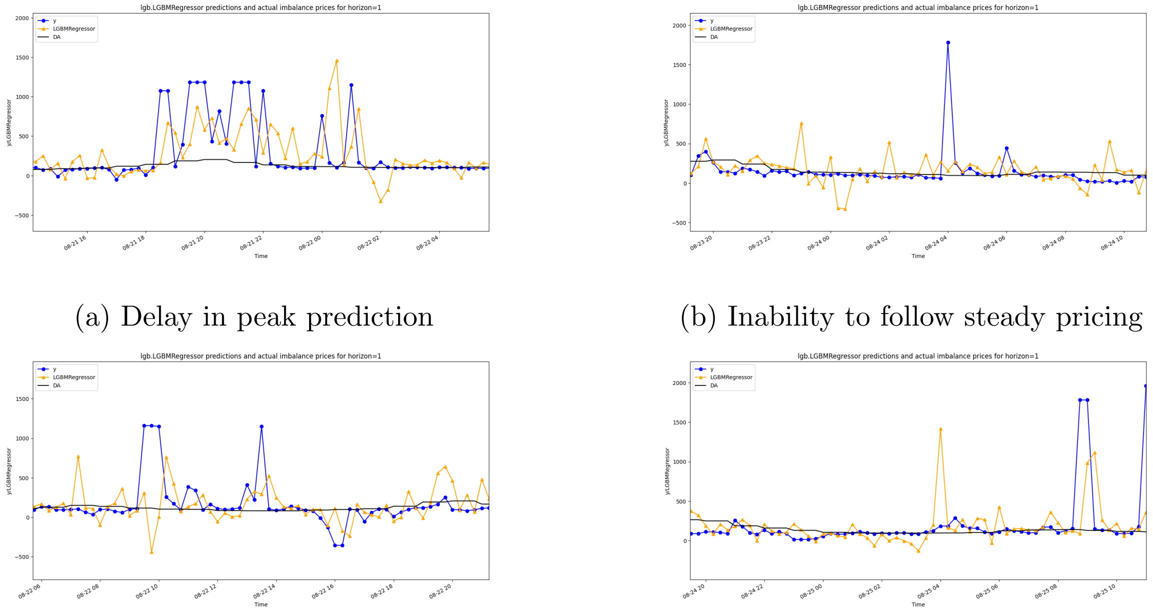
(a) False peaks

(b) Inability to follow steady pricing

(c) False peaks

(d) Wrong side of DA

Figure 12: Problems of Lasso model



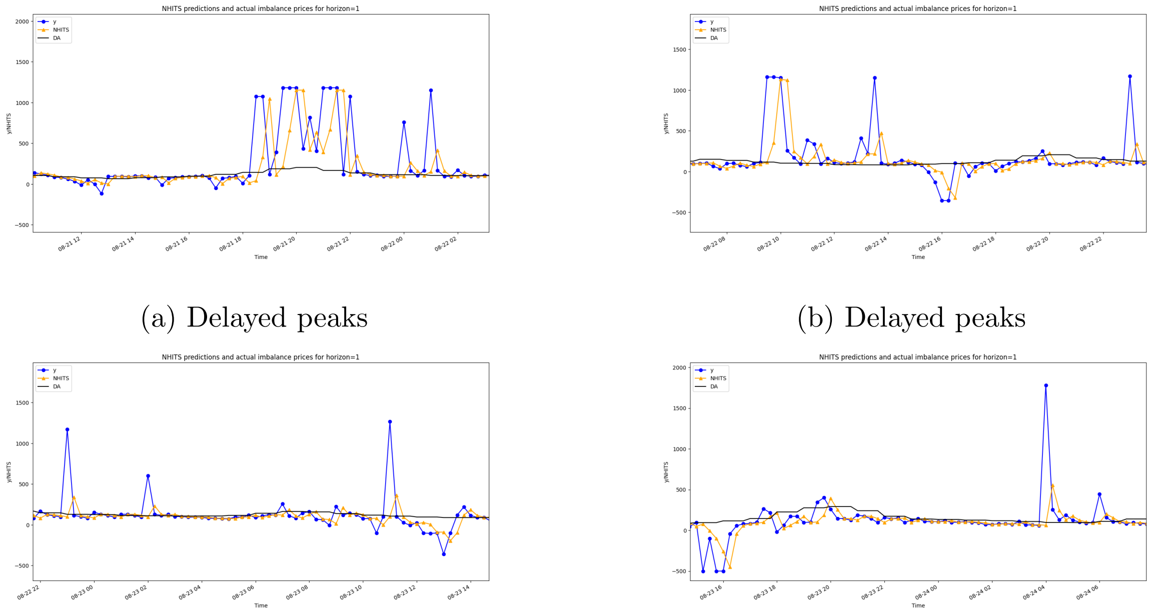
(a) Delay in peak prediction

(b) Inability to follow steady pricing

(c) Predictions on the wrong side of DA

(d) Random and delayed peak

Figure 13: Problems of lightGBM model



(a) Delayed peaks

(b) Delayed peaks

(c) Flat predictions

(d) Wrong side of DA

Figure 14: Problems of NHITS model

have a similar performance to each other. We observe a similar amount of random peaks which are sometimes on the wrong side of the day-ahead price. LightGBM is considered to be slightly better due to the slightly lower amplitude and frequency of false peaks. The best of these four models is the NHITS model. It does not predict on the wrong side of the day-ahead price as often as the other models. Due to the delayed predictions it also gets some peaks correct. However, these delayed predictions are not useful as it does not allow simpl.energy to beat the market and are thus considered bad.

5.3 Model performance determined by the quality measures

The second goal of this experiment is to test our findings of the first experiment with real imbalance price forecasting. This is done by comparing the quality measures scores and see how well they align with the previously stated expert opinion. If the MAE, rMAE and Punishment score capture the expert opinion the best, the findings of Experiment 1 will be more conclusive. Otherwise the results of Experiment 1 needs to be reconsidered in future research. For this purpose Table 5 contains all the quality measure scores for each of the four models. As previously stated, the Naive model is considered the worst, Lasso and lightGBM have similar performance with lightGBM being slightly better and the NHITS model is considered the best. A good quality measure should be able to reflect this ranking.

All the quality measures, with the exception of the Slope RMSE, are correctly able to identify the NHITS model as the best performing. The slope RMSE has a score of 438 for the Naive model where the NHITS model, the second best according to slope RMSE, has a score of 485. Most of the quality measures also score Lasso and lightGBM similarly. Only the slope RMSE has a significantly

Table 5: Model Evaluation Metrics

Model	RMSE	R-squared	Correlation	MAPE	MAE	rMAE	Punishment Score	Slope RMSE
Naive	473.526	-0.639315	-0.077533	1.639312	231.209200	1.000000	1.406209	437.912194
Lasso	382.694902	-0.070731	0.334072	2.236264	238.315113	1.030734	1.450815	501.647281
lgb.LGBMRegressor	381.036871	-0.061473	0.3029	1.830554	215.372478	0.931505	1.397872	485.362587
NHITS	351.137733	0.098574	0.432575	0.800142	143.268840	0.569208	0.788269	484.885798

worse score for Lasso with a score of 502 and 485 for lightGBM. Three quality measures rank the Naive model as the worst, RMSE, R-squared and Correlation. This makes these three the only quality measures that got the correct ranking. This directly contradicts the findings of the first experiment where RMSE, R-squared and Correlation are deemed to not accurately evaluate model performance. All the other quality measures, again with the exception of the Slope RMSE, rank the Naive model in between LightGBM and Lasso.

Discussion The performance of the models, especially the NHITS model as stated previously, indicates that the open source data might not have enough correlation with the actual imbalance price to make accurate predictions. If this is true this could have caused the fact that there are no models that are considered good (enough), which might also be the reason for the contradiction with the first experiment. The error based quality measures become worse when the errors are smaller around the day-ahead price, as showcased by Bad model 4 in the first experiment. We consider it likely that with better performing models the performance of the RMSE, R-squared and correlation will drop and the MAE, rMAE and punishment score will be better able to represent the performance of the different models. This is a limitation of this research, thus it would be wise to reconsider the conclusions from the first experiment in future research.

6 Conclusion

In conclusion, our research focused on finding a robust quantitative measure for assessing the performance of forecasting models, particularly in the context of predicting imbalance prices. Two experiments were undertaken to achieve this objective.

The first experiment involved the evaluation of fictitious models labelled as bad, medium, and good. This evaluation was done with a selection of commonly used, more uncommon and two proposed quality measures. While most quality measures demonstrated efficacy in distinguishing between the labelled classes, a notable exception was observed with Bad model 4. Despite its numerical scores resembling those of good models, its frequent bad predictions underscored the limitations of error-based measures in capturing model performance nuances. However, the Punishment score, MAE, and rMAE emerged as reliable metrics, effectively differentiating all bad models from the medium and good models across various data set lengths.

The second experiment extended the analysis to real forecasting algorithms, including a Naive model, Lasso, LightGBM, and NHITS. The Naive model exhibited poor performance, particularly in generating predictions on the wrong side of the day-ahead price. LightGBM and Lasso demonstrated comparable performance. The NHITS model outperformed the others, although the delayed predictions lack practical utility. The evaluation of these models using multiple quality measures yielded insights, with RMSE and Correlation standing out as the only measures reflecting the correct ranking despite their earlier dismissal.

The comparison between the two experiments highlighted the complexities in evaluating forecasting models. While the first experiment emphasised the limitations of certain measures, the second experiment introduced real-world complexities and showcased the challenges in accurately ranking models.

In summary, the research underscores the importance of selecting appropriate quality measures that align with the specific characteristics of the forecasting task. The proposed Punishment score, along with MAE and rMAE, emerges as a promising set of measures for comprehensive model evaluation, offering valuable insights for practitioners and researchers in the field of electricity price forecasting. Future studies could further refine and validate these measures, taking into account the nuances inherent in forecasting tasks and their real-world applications.

References

[1] J. Browell and C. Gilbert. Predicting electricity imbalance prices and volumes: Capabilities and opportunities. *Energies*, 15:3645, 2022.

[2] S. Chai, Q. Li, M.Z. Abedin, and B.M. Lucey. Forecasting electricity prices from the state-of-the-art modeling technology and the price determinant perspectives. *Research in International Business and Finance*, 67:102132, 2024.

[3] C. Challu, K.G. Olivares, B.N. Oreshkin, F. Garza Ramirez, M. Mergenthaler Canseco, and A. Dubrawski. Nhits: Neural hierarchical interpolation for time series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6989–6997, Jun 2023.

[4] ENTSO. Central collection and publication of electricity generation, transportation and consumption data and information for the pan-european market. https://transparency.entsoe.eu, 2023.

[5] Yunmi Kim, Tae-Hwan Kim, and Tolga Ergün. The instability of the pearson correlation coefficient in the presence of coincidental outliers. *Finance Research Letters*, 13:243–257, 2015.

[6] G. Klæboe, A. L. Eriksrud, and S.E. Fleten. Benchmarking time series based forecasting models for electricity balancing market prices. *Energy Systems*, 6(1):43–61, March 1 2015.

[7] J. Lago, G. Marcjasz, B. De Schutter, and R. Weron. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983, 2021.

[8] lightgbm. Lightgbm.lgbmregressor — lightgbm 4.1.0.99 documentation. https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMRegressor.html.

[9] Rukshan Manorathna. k-fold cross-validation explained in plain english (for evaluating a model's performance and hyperparameter tuning). 12 2020.

[10] G. Marcjasz, B. Uniejewski, and R. Weron. Beating the naïve—combining lasso with naïve intraday electricity price forecasts. *Energies*, 13(7):1667, 2020.

[11] M. Narajewski. Probabilistic forecasting of german electricity imbalance prices. *Energies*, 15(14):4976, 2022.

[12] Nixtla. Mlforecast. https://nixtla.github.io/mlforecast/forecast.html.

[13] Nixtla. Neuralforecast. https://nixtlaverse.nixtla.io/neuralforecast/index.html.

[14] I. Oksuz and U. Ugurlu. Neural network based model comparison for intraday electricity price forecasting. *Energies*, 12(23):4557, 2019.

[15] J. Peters. Modeling the dutch frequency restoration reserve market. Master's thesis, Delft University of Technology, November 2016.

[16] scikit learn. sklearn.linear_model.lasso. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html.

[17] Natasha Sharma. Cross validation: What you need to know, from the basics to llms. https://arize.com/blog/cross-validation-machine-learning/, May 2023.

[18] Robert T. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[19] TenneT TSO B.V. About tennet. https://www.tennet.eu/nl/over-tennet.

[20] N. R. Terpstra. Day-ahead and imbalance price forecasting on the dutch electricity market: a comparison between time series and artificial neural networks models. Master's thesis, TU Eindhoven, 2020.

Appendix A

Feature	Description
Date time	Datetime values with time zone information
Forecasted Generation	The day-ahead forecasts for energy generation
Day Ahead Price (hourly)	Day-ahead price
Forecasted Load	The day-ahead forecasts for energy demand
Delta (Fc generation - Fc load)	Difference in forecasted generation and forecasted consumption
x min IGCC down	IGCC activity for downward regulation x minutes earlier
x min downregulate	aFFR and mFFR activity for downward regulation x minutes earlier
x min down	Lowest activated bid for downward regulation x minutes earlier
x min up	Highest activated bid for upward regulation x minutes earlier
x min IGCC up	IGCC activity for upward regulation x minutes earlier
x min upregulate	aFFR and mFFR activity for upward regulation x minutes earlier
x min noodvermogen	FCR activation
15 Long	Downward regulating imbalance price 15 minutes earlier
30 Long	Downward regulating imbalance price 30 minutes earlier
15 Short	Upward regulating imbalance price 15 minutes earlier
30 Short	Upward regulating imbalance price 30 minutes earlier
Target Long	Downward regulating imbalance price
Target Short	Upward regulating imbalance price

Feature	Range	Accuracy	Standard Deviation	Mean
Date time	[2022-01-01, 2023-08-31]	15 minutes	-	-
Forecasted Generation	[2548, 22442]	1 MW	2177.330231	8275.484399
Day Ahead Price (hourly)	[-500.00, 871.00]	€ 0.01	127.576328	184.901793
Forecasted Load	[439, 28899]	1 MW	2329.614989	10750.798712
Delta (Fc generation - Fc load)	[-12809, 11378]	1 MW	1792.480506	-2475.278148
x min IGCC down	[0, 1037]	1 MW	90.038429	42.065411
x min downregulate	[0, 1017]	1 MW	81.806523	37.879683
x min down	[-999.00, 573.64]	€ 0.01	113.465554	150.247173
x min up	[-11.95, -1781.23]	€ 0.01	140.097208	192.998893
x min IGCC up	[0, 936]	1 MW	92.655289	50.132519
x min upregulate	[0, 655]	1 MW	77.578940	38.719814
x min noodvermogen	[0, 1]	binary	-	-
15 Long	[-999.00, 1959.35]	€ 0.01	229.162313	179.633256
30 Long	[-999.00, 1959.35]	€ 0.01	229.162313	179.633256
15 Short	[-999.00, 1959.35]	€ 0.01	228.715005	192.341367
30 Short	[-999.00, 1959.35]	€ 0.01	228.715005	192.341367
Target Long	[-999.00, 1959.35]	€ 0.01	229.162313	179.633256
Target Short	[-999.00, 1959.35]	€ 0.00	228.715005	192.341367

Note: Each feature with “x min” in the name represents 15 columns with $x = \{15, 16, ..., 29\}$.