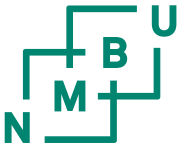


Norwegian
University of
Life Sciences

Master's Thesis 2024 30 ECTS
Faculty of Science and Technology

Analysing Norwegian Electricity Prices During the Energy Crisis: A Data-Driven Forecasting Approach with Explainable AI

Oscar P. Viken
Environmental Physics and Renewable Energy

Abstract

The Russian invasion of Ukraine has significantly impacted energy resources across Europe, leading to a widespread increase in electricity prices. In an effort to mitigate this, European nations have significantly expanded their renewable energy projects, aiming to lessen their reliance on Russian energy and to confront the global climate crisis. As a result of these initiatives, electricity price volatility has dramatically increased across the continent. Due to the interconnected nature of the European electricity market, the surge in prices and increased volatility also affect Norway, which in recent years has become even more coupled with the European market due to new subsea HV-DC connectors.

The new market dynamics and elevated prices have made forecasting electricity prices increasingly challenging. This thesis evaluates the effectiveness of machine learning models, particularly focusing on Long Short-Term Memory (LSTM), for forecasting day-ahead prices in Norway's NO2 bidding zone under these difficult conditions. Alongside the LSTM, a Deep Neural Network (DNN) model is developed, and both are compared with an established benchmark. Following an extensive hyperparameter optimisation process, including cross-validation and multi-objective Sequential Model-Based Optimisation, both models achieved similar performance, with the LSTM model attaining a mean absolute error of 15.48 EUR/MWh and the DNN model reaching 15.62 EUR/MWh. The performance of our models aligns with similar studies, though direct comparison is limited due to our models' strict alignment with the day-ahead market timings, an aspect often overlooked.

Furthermore, the thesis employs Explainable Artificial Intelligence techniques, specifically SHapley Additive exPlanations (SHAP), to analyse electricity price predictions. It examines two separate Extreme Gradient Boosting models, each using data from different periods, one before and one after the onset of the energy crisis. This analysis highlights the influence of key market-specific variables, such as residual loads, hydro reservoir levels, and gas prices, on electricity price forecasts. By integrating SHAP values, this study quantifies the importance of these features, enhancing the transparency and understanding of the driving factors behind electricity prices in Norway. The dual-dataset approach further allows for an examination of market differences pre- and post-energy crisis, providing insights into the changing market dynamics.

Our results indicate oil and gas prices as the primary drivers of electricity prices in Norway, with the influence of gas prices becoming even more pronounced following the onset of the energy crisis. As anticipated, there is also an increased dependency on German residual load after the crisis, which coincides with the introduction of the NordLink subsea HV-DC cable between Norway and Germany. Additionally, before the crisis, the results reveal an odd non-linear dependency between electricity prices and the net position. Prices decrease when exports are below 2000 MW and sharply increase above this threshold. Post-crisis, this behaviour becomes more consistent with a pure economic view where imports lower prices, while exports raise them.

Sammendrag

Den russiske invasjonen av Ukraina har hatt en betydelig innvirkning på energiressursene i Europa, og har medført en markant økning i strømprisene. For å motvirke dette har europeiske nasjoner akselerert utbyggingen av ny fornybar energi, med mål om å redusere avhengigheten av russisk energiforsyning samt å takle den globale klimakrisen. Disse tiltakene har resultert i en dramatisk økning i volatiliteten i strømprisene over hele kontinentet. På grunn av det tett sammenkoblede europeiske elektrisitetsmarkedet, påvirker denne prisøkningen og økte volatiliteten også Norge, som i senere tid har blitt enda tettere integrert med det europeiske markedet gjennom nye undervanns HV-DC-forbindelser.

Som følge av konflikten har endringene i markedet og prisøkningen gjort det stadig vanskeligere å predikere strømpriser. Denne oppgaven evaluerer ytelsen av maskinlæringsmodeller, hovedsakelig Long Short-Term Memory (LSTM), for å predikere day-ahead priser i det norske budområdet NO2 ved disse utfordrende markedsforholdene. I tilegg til LSTM-modellen utvikles et Dypt Nevralt Nettverk (DNN), hvor begge modellene sammenlignes med en etablert referansemodell. Modellene gjennomgår omfattende hyperparameteroptimalisering, som inkluderer kryssvalidering og sekvensiell modellbasert optimalisering. De oppnår lignende ytelse, med et gjennomsnittlig absolutt avvik på 15,48 EUR/MWh for LSTM-modellen og 15,62 EUR/MWh for DNN-modellen. Ytelsen til modellene er tilsvarende lignende studier, selv om direkte sammenligning er utfordrende på grunn av modellenes strenge tilpasning til day-ahead-markedet, et aspekt som ofte overses i andre studier.

Videre brukes forklarbar kunstig intelligens, spesifikt SHapley Additive exPlanations (SHAP), for å analysere prediksjoner laget av to XGBoost-modeller. Disse modellene er utviklet basert på datasett fra periodene før og etter energikrisens start. Analysen fremhever hvordan nøkkelvariabler som blant annet, residuallast, fyllingsgrad og gasspriser påvirker strømprisprognosene. Gjennom bruk av SHAP-verdier kvantifiserer studien betydningen av disse variablene, noe som gir en forståelse av de drivende faktorene bak strømprisene i Norge. Den todelte datasett tilnærmingen gir også muligheten til å utforske forskjeller i markedet før og etter energikrisen, og gir innsikt i hvordan markedet har endret seg.

Resultatene viser at olje- og gasspriser er de primære driverne bak strømprisene i NO2, hvor gassprisens innflytelse blir enda mer markant etter starten på energikrisen. Som forventet øker også avhengigheten av tysk residual last etter krisen, noe som sammenfaller med introduksjonen av undervanns HV-DC-kabelen NordLink mellom Norge og Tyskland. Videre avslører resultatene en ikke-lineær sammenheng mellom strømpriser og nettoposisjon før krisen. Her synker prisene når eksporten er under 2000 MW og øker kraftig når den overstiger dette nivået. Etter krisen blir denne sammenhengen mer konsistent med et rent økonomisk synspunkt hvor import senker prisene, mens eksport øker dem.

Acknowledgements

That my years of studying are now coming to an end is almost incomprehensible. The five years at university have absolutely flown by, and have been some of the most rewarding and challenging years of my life. The final semester of thesis writing definitely adds to that.

Writing this thesis would not have been possible without my supervisor, Leonardo Rydin Gorjão, not just because of his expertise in the subject matter, but also due to him being such a kind and welcoming person. A big thank you for always answering my questions, no matter how big or small, and for calming my nerves during periods of high stress.

Also, I need to acknowledge my family for their immense support through these years. This degree really feels like a team effort and would not have been possible without you.

Finally, a massive thanks to all my friends who have made these years so enjoyable.

Contents

List of Figures

List of Tables

List of Acronyms

EPF Electricity Price Forecasting

TSO Transmission System Operator

DAM Day-Ahead Market

BZN Bidding Zone

SDAC Single Day-Ahead Coupling

ML Machine Learning

FFNN Feed-forward Neural Network

RNN Recurrent Neural Network

LSTM Long Short-Term Memory

GB Great Britain

MAD Mean Absolute Deviation

TPE Tree-structured Parzen Estimator

DNN Deep Neural Network

XAI EXplainable Artificial Intelligence

GBT Gradient Boosted Tree

MAE Mean Absolute Error

RMSE Root Mean Square Error

sMAPE symmetric Mean Absolute Error

rMAE relative Mean Absolute Error

SMBO Sequential Model-Based Optimisation

SHAP SHapley Additive exPlanations

XGBoost eXtreme Gradient Boosting

LIME Local Interpretable Model-agnostic Explanations

BPTT Backpropagation Through Time

LASSO Least Absolute Shrinkage and Selection Operator

LEAR Lasso Estimated Auto-Regressive

1 Introduction

1.1 Overview

During the last couple of years, Norwegian consumers have been exposed to electricity prices significantly higher than the historical norm. Traditionally, Norway has maintained prices below those of our neighbouring countries, largely attributed to a strong energy balance and the abundant production of hydroelectric power [1]. While governmental subsidies have reduced the price effects for domestic consumers, the economic ramifications for power-intensive industries are still substantial [2]. One of the main contributors to the increasing electricity prices is the Russian invasion of Ukraine. The dependency of many European countries on Russian oil and natural gas has led to a scarcity in the European energy supply [3]. In the Norwegian context, the impact is not characterised by a direct shortage, but rather implicit implications, such as price contagion [1].

The energy transition, initiated by the European Union to combat the climate crisis, gained additional momentum following the Russian invasion of Ukraine. This shift is crucial in tackling environmental challenges and aims to reduce reliance on oil and gas, thereby decreasing Europe's dependency on Russian energy sources. However, this transition presents several challenges. It is primarily facilitated by increased electrification and the expansion of generation from intermittent renewable sources like solar and wind. The generation from these sources is highly volatile, complicating the balance between generation and demand.

Electricity is a special commodity because it is economically non-storable, and power system stability requires a constant balance between production and consumption [4]. The demand varies seasonally, due to factors such as weather conditions and different consumption patterns during, day and night, weekdays and weekends, etc. Along with being variable, the demand is also strongly inelastic, i.e. it is not strongly influenced by price signals [5]. On the supply side, generation is associated with uncertainty, often deviating from planned production due to factors that can not be accounted for. Weron states that "these specific characteristics lead to price dynamics not observed in other markets, exhibiting seasonality at daily, weekly and annual levels, and unanticipated price spikes" [4].

Electricity markets crucially balance the supply and demand of energy, while minimising the cost to consumers. However, this was not always the case, with the European power sector being predominantly monopolistic and government-controlled until the 1990s. Norway emerged as a pioneer in deregulating its electricity system in 1991, motivated by growing dissatisfaction with performance, particularly in the realm of investments. This decision has resulted in several positive outcomes, including a decrease in prices and more equitable pricing across consumer groups [6].

To ensure balancing, electricity is now traded on markets operating on different time horizons. The electricity market is generally bifurcated into two main cat-

egories: retail and wholesale. In the retail market, electricity is directly sold from producers to end-consumers. This involves bilateral contracts, often characterised by long-term commitments and relatively stable volumes. In Norway, the retail market accounts for approximately one-third of the total consumption, with customers ranging from industry to larger end-consumers such as hotels and retailers [7]. Further, the wholesale market is subdivided into three key segments: day-ahead, intraday, and balancing markets. Notably, the day-ahead and intraday markets transact on power exchanges, with the Day-Ahead Market (DAM) boasting the highest trading volumes. Meanwhile, the balancing market is operated by the Transmission System Operator (TSO), tasked with balancing momentary supply and demand [7].

Predicting electricity prices has become increasingly more difficult in recent years, primarily due to the rise in price levels and volatility. Obtaining accurate electricity price forecasts is crucial for market participants, as it allows them to allocate resources efficiently, improve bidding strategies, and hedge against potential risks. Also, it enables consumers to make a profit by anticipating price movements and optimising smart systems correspondingly [2]. Various models and methodologies have been employed for Electricity Price Forecasting (EPF) over the years. In the earlier part of the 2010s, conventional approaches centred around simple linear regression models and artificial neural networks with a limited set of features [8], but with the massive increase in data and computational power in recent years, models have grown larger and more advanced Machine Learning (ML) approaches have shown promising results [8].

However, ML models are no silver bullet. Although they enable us to capture complex non-linear relations between features, their effectiveness relies on the data provided truly representing the function we are trying to estimate. Issues can arise, such as having excessively short time series or dealing with non-stationary data where the characteristics change over time, thereby making the data insufficient for fitting a complex model for the task [9]. ML models are often also criticised for their black-box nature. This is due to the difficulty in understanding the mechanisms behind their decision-making process. Especially in areas that require a high degree of reliability, it might be difficult to trust and implement such models. One approach to address this issue is through the adoption of EXplainable Artificial Intelligence (XAI), which aims to enhance transparency by revealing how features contribute to predictions [10].

In this thesis, our objective is to forecast electricity prices in the Norwegian DAM during challenging market conditions using ML models. Our primary forecasting model is an Long Short-Term Memory (LSTM). While LSTM models have been applied in the field of EPF previously, there remains a significant research gap concerning the Norwegian market in the aftermath of the energy crisis. Our aim is to fill this gap and demonstrate the effectiveness of employing ML models for EPF in this context.

Additionally, we leverage ML models combined with XAI to analyse the primary

drivers of electricity prices in Norway. Specifically, we utilise SHapley Additive exPlanations (SHAP) to attribute significance to individual features and examine their impact on price predictions. This method allows us to assess the influence of different factors on the market and explore how the Norwegian electricity market has changed in response to the energy crisis.

1.2 Problem Statement

This master's thesis seeks to explore the following two questions:

- Can machine learning models be used to accurately forecast spot prices in Norway during periods with high electricity price volatility?
- What are the main drivers of the electricity prices in the Norwegian electricity market, and how have recent changes in energy markets affected them?

2 The Norwegian Electricity Market

This chapter aims to give a concise introduction to the Norwegian electricity market, central to the focus of this thesis. It is worth noting that markets in other European countries share similar characteristics, making them somewhat analogous. The first section provides a broad overview of the Norwegian market structure. Subsequently, we explore the mechanisms through which day-ahead prices are settled. Finally, we examine the integration of the European market, and how prices are affected through Single Day-Ahead Coupling (SDAC).

2.1 Market Structure

All electricity producers contribute to the same power grid, and once delivered, the supply becomes indistinguishable. The amount of electricity a producer sells at a given time is not necessarily tied to their actual generation. This is because, in order to maximise profits while fulfilling sales commitments, producers have the flexibility to engage in buying and selling across various markets. Only large players such as producers, traders and industry, trade electricity on power exchanges or through bilateral contracts, mainly due to high transaction costs [7].

The Norwegian electricity market is segmented into several Bidding Zones (BZNs) to address congestion issues within major grid bottlenecks. This division allows for the reflection of localised supply and demand dynamics as well as grid limitations in wholesale electricity prices. As a consequence, price variations across these zones generate congestion revenue, which the TSO allocates towards operating costs and grid capacity enhancements. Currently, Norway comprises five distinct BZNs, labelled from NO1 to NO5. This segmentation is shown in Figure 2.1, alongside BZNs of some neighbouring northern European countries.

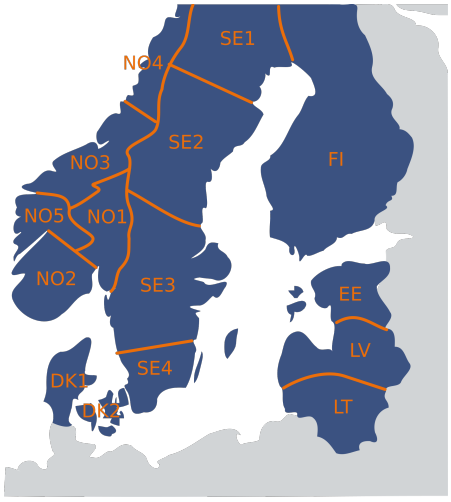


Figure 2.1: Overview of the northern European bidding zones. Image taken from [11] under CC BY-SA 4.0 license.

As briefly mentioned in section 1.1, the Norwegian electricity market is segmented into various markets, these include the spot market, where electricity is traded for immediate physical delivery, and financial markets, where the delivery is at a later date. Figure 2.2 shows the overall structure and the approximate time frame within which these markets operate.

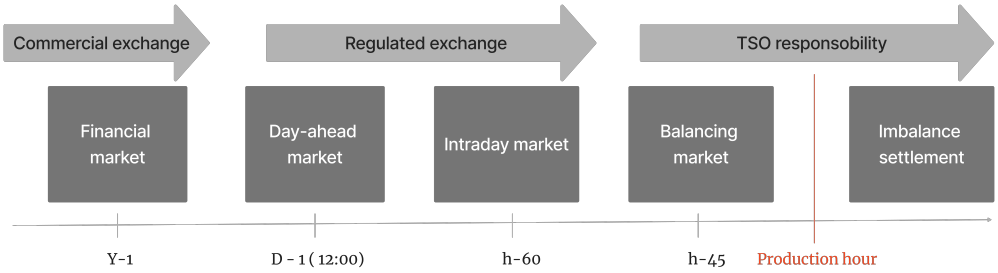


Figure 2.2: Overview of the electricity market, with approximate operating time frames.

The financial electricity market in Norway is currently managed by NASDAQ OMX, which acquired NordPool's market share in 2010 [12]. Financial trading enables participants to adjust their risk profile and hedge against possible price fluctuations by trading financial instruments [6]. Commonly traded financial instruments include future and forward contracts, where the electricity price is predetermined, and the contracts' performance is determined by the difference between the agreed price and the reference spot price (system price). This happens without physical delivery, and also occurs well in advance of spot price settlement, spanning from days to years prior.

The DAM serves as the primary market for trading physical electricity contracts, with NordPool being the main platform for such trading in Norway. Participants engage in the market by submitting bids and offers during the opening hours from 08:00 to 12:00. The TSO publishes available grid capacities at 10:00, and bids and offers are submitted hourly for the entire following day. The auction concludes at 12:00, and prices for each hour of the following day are determined based on the equilibrium achieved by aggregating bids and offers. The settled trades in the DAM constitute the binding generation and consumption plan that is delivered to the TSO [13].

While DAMs typically function as the primary mechanism for balancing supply and demand, unforeseen events can disrupt predetermined production and consumption plans. Factors such as inaccurate weather forecasts and issues with generation plants may arise. To address such changes, market participants have the option to engage in the Elbas market. Elbas, the intraday market in the Nordics (and certain other European countries), allows continuous trading up to one hour before delivery. This provides market actors with the opportunity to adjust their positions to avoid

paying fines in the balancing market. Unlike the DAM where prices are cleared each hour, Elbas operates on a 'pay-as-bid' system, similar to regular stock markets, due to the continuous nature of trading [14].

Disturbances may also occur within the operating hour. In Norway, Statnett is responsible for ensuring the momentary balance of the power system. This equilibrium must be maintained to prevent significant frequency deviations from the nominal value, which is 50 Hz in the Nordic synchronous area. To achieve this, TSOs like Statnett purchase flexibility from producers or large consumers who are willing to up- or down-regulate their generation and/or consumption. Various frequency control measures exist and operate at different time intervals. The specific pricing mechanisms associated with these measures fall outside the scope of this thesis, which focuses solely on the DAM.

2.2 Price Formation

After the closure of the DAM, NordPool calculates a system price for all the Nordic countries [15]. The calculation aggregates all bids and offers for a particular hour, without considering transmission restrictions in the grid. The system price is determined at the intersection of the supply and demand curves. This intersection point reflects the marginal cost of the last accepted production type. In the DAM, production types are selected in order of increasing cost until demand is met. Variable renewable resources are usually offered cheaply into the market, while dispatchable power sources normally have a higher marginal cost. The difference in the marginal cost of different production types creates the merit order, as shown in Figure 2.3. Importantly, the market operates on a 'pay-as-cleared' system, whereby all market participants receive the marginal cost of the last accepted offer within their respective BZN.

In reality, there are limitations to transferring electrical power, and power balance can vary greatly depending on geographical location. To manage major grid bottlenecks while maintaining local system balance, Norway is divided into five . Each BZN is subject to its own price (area price), with price differentials between zones determining the electricity flow. Consider a simplified system with two BZNs connected by a transmission line. If one zone has a surplus of inexpensive electricity, it would naturally have a lower price. To equilibrate prices and direct power where its value is maximised, electricity flows through the transmission line. However, the transmission lines' capacity is not infinite. If a bottleneck occurs, the power transferred is insufficient to equalise prices across zones. Hence, the zone with the surplus of inexpensive electricity would experience lower prices, while the other zone would see higher prices. From a demand-supply perspective, the power transferred from the cheaper zone is perceived as a demand increase in that zone's demand curve. Conversely, the recipient zone experiences an increase in its supply of more affordable electricity. In the ideal case, the area prices would converge to the system price.

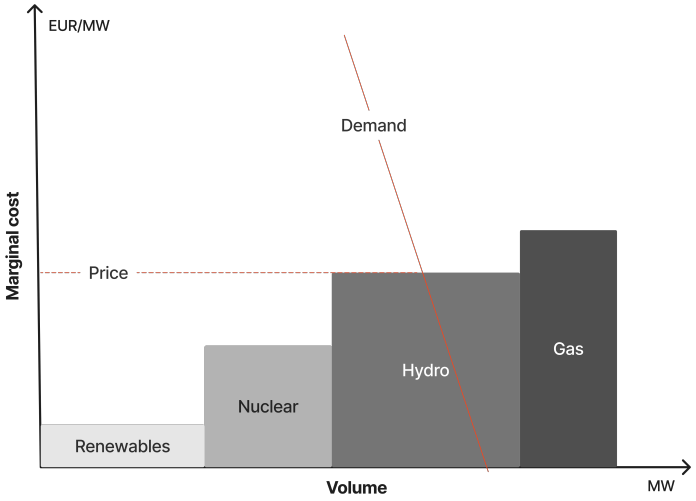


Figure 2.3: Simplified merit order curve showing the market clearing price (MCP) at the intersection of supply and demand, marked by two red lines.

Limited transmission capacity is a key factor contributing to the substantial price difference between Norway's northern and southern regions.

2.3 Single Day-Ahead Coupling and European Integration

The determination of area prices in Norway is not carried out in isolation. Due to the interconnections not only with other Nordic countries but also with high-voltage direct current (HV-DC) cables to continental Europe and Great Britain, Norway is integrated into a larger market. SDAC seeks to create a unified DAM in Europe, aiming to improve trading efficiency and enhance the utilisation of energy resources [16].

The EUPHEMIA algorithm is responsible for the calculation of area prices. Relevant information about transmission capacities, bids and offers is provided by the participating TSOs and power exchanges. Using these data, the EUPHEMIA algorithm calculates the power flows and determines which orders to execute in each BZN in order to maximise social welfare. Social welfare is defined as the sum of consumer surplus, producer surplus, and congestion rent across all regions [17]. This is all done while ensuring transmission constraints are not exceeded. Constraints do not only include transmission line capacities but also ramping restrictions, which refer to the rate at which the direction of power in transmission lines can be changed.

Norwegian electricity prices are predominantly influenced by the supply and demand dynamics within the Nordic power market and the power balance across other central European countries [18]. Key components of the Nordic region's generation capacity are hydropower and nuclear energy, both of which have low production

costs. The volume of hydropower generated, largely dependent on precipitation and reservoir inflow, significantly affects the total output potential and electricity prices. As the Nordic market becomes more integrated with the rest of Europe, European price signals increasingly influence Nordic prices. Historically, European electricity generation has been dominated by thermal power plants like coal and gas-fired stations. However, from 2004 to 2022, there has been a substantial increase in renewable energy sources, with their share in total energy generation doubling [19]. In the future energy market, the penetration of intermittent renewables is expected to increase, leading to more price variations due to limited storage capacity and fluctuating production levels [1]. The impacts of these changes have already started to emerge, with instances of negative prices across Europe and unprecedented price spikes. These changes in the energy market make prices more volatile, and prediction more difficult.

Norway shares interconnectors with several European countries, including Denmark, Finland, Russia, The Netherlands, Germany, and Great Britain. Recent substantial additions to these interconnections are the HV-DC cables NordLink and North Sea Link, both operational since 2021, connecting Norway to Germany and Great Britain, respectively. The total capacity for power exchange across all these interconnections is approximately 9000 MW, distributed as follows: 4000 MW to Sweden, 1400 MW each to Germany and Great Britain, 1600 MW to Denmark, and 700 MW to The Netherlands [20]. In theory, these interconnections allow for a power transfer of 80 TWh annually. However, the actual power exchange is considerably smaller in practice. Despite this, it is important to note that the power exchange with other countries has seen an increase with the introduction of the latest interconnectors, NordLink and North Sea Link [20].

While closer integration with the European market tends to level out price disparities, thus mainly leading to higher overall prices in Norway, it also introduces certain advantages. As stated by [7], "power sharing allows countries to derive mutual benefits from their distinct natural resources, diverse methods of electricity production, and varying consumption patterns." This is particularly relevant for Norway, whose energy system heavily relies on hydropower and consequently is subject to weather conditions. The integrated market enables Norway to import power when its capacity is limited and export when there is a surplus. This is reflected in real trading activities, with the highest import levels typically occurring in the winter when reservoir inflow is low, and the highest export levels during the summer when inflow is high [7].

3 Main Concepts of Machine Learning

Numerous ML models exist for forecasting electricity prices. This chapter aims to develop a clear understanding of these models, particularly those employed within this thesis. We begin with the fundamental concepts of ML, progressing to an examination of Feed-forward Neural Networks (FFNNs), and subsequently delving into more sophisticated architectures, such as the LSTM and the eXtreme Gradient Boosting (XGBoost) architectures. The concluding section gives a mathematical description of Shapley values, and how they can be used to explain ML models, both locally and globally.

3.1 Fundamentals of Machine Learning

The area of ML focuses on the development of algorithms that enable systems to learn from data without explicit instructions. By using historical data as input, ML models can perform a wide variety of tasks, including predictions, classification, dimensionality reduction, and even generation of new content [21]. The field is split into three main categories: supervised, unsupervised, and reinforcement learning. In this thesis, the main focus is on supervised learning.

The aim of supervised learning is to learn relations between input data and some desirable output. Formulated mathematically, we want to find a function $f : X \to Y$, that maps observations from the feature space (X) to an output space (Y). Finding the function $\hat{f}$ that represents the true mapping is often challenging. However, supervised learning provides methods to approximate $\hat{f}$, by utilising labelled samples $(x, \hat{f}(x))$.

In supervised learning, we try to find an approximate function f by searching through a space of candidate functions F. We aim to find the function f in F that yields the most accurate predictions for the samples given in our dataset. To evaluate the performance of different candidate functions, we use a loss function $\mathcal{L}$. For a given labelled sample this function measures the difference between the predicted output $y = f(x)$, and the ground truth $\hat{y} = \hat{f}(x)$. The nature of the labels varies depending on the specific problem at hand. There are two primary forms of supervised learning: regression, where the objective is to predict continuous output values, and classification, where the goal is to predict discrete values. In classification, for instance, one might predict whether the sentiment of a movie review is positive or negative. Formally, the task of finding an optimal approximation, given a set of n samples, can be expressed as

$$f = \underset{f \in F}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f(x_i), \hat{y}_i). \tag{3.1}$$

In most cases the set of available labelled data is finite, and given that our search space of functions F is sufficient, it will always be possible to find an optimal solution to equation (3.1). However, the primary objective is not merely to find the mapping

between already known data points. The main concern is to find a solution that also provides good predictions on unseen data, or put differently, to find a model that generalises well. To achieve this, we commonly partition the dataset into two parts: the training data, which is used to navigate through the search space, and the validation data, which provides an estimate of the model's generalisation ability. The validation data is typically used during the training to guide the search process, providing insights into the model's ability to generalise. However, to prevent the model from becoming overly adjusted to the validation data, and to ensure true generalisation, a separate test set is usually included. This provides an unbiased evaluation of the model, as the test data is completely isolated from the training process, thereby offering a realistic measure of the model's predictive performance on unseen data.

When a model performs significantly better on training data than on the test data, it is an indication of overfitting. Overfitting happens when the model becomes too complex and gives accurate predictions for the training data but not for new unseen data. This occurs because the model has high variance, meaning its predictions would drastically change if it were trained on a different subset of the same overall distribution. On the other hand, if a model performs poorly on both training and test data, it is underfitting. Underfitting occurs when the model fails to capture the relationships between input and output variables, leading to a high bias. Striking a balance between bias and variance is crucial in finding a model that generalises well. In practice, finding this balance is difficult given the inverse relationship between the two quantities.

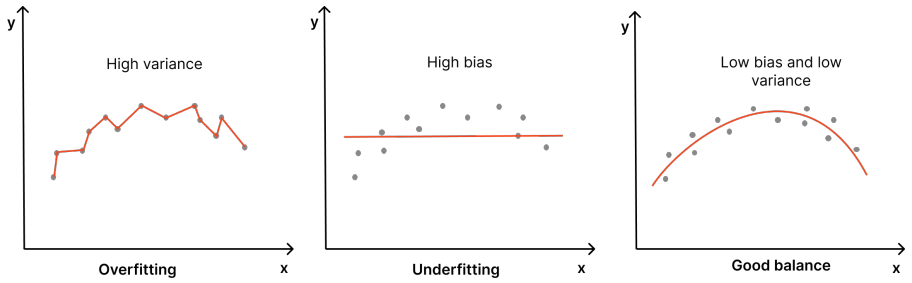


Figure 3.1: Visual representation of model overfitting (left), underfitting (middle), and an appropriately fitted model that captures the general trends of the data (right).

Techniques like regularisation and adjusting model complexity are essential for managing this trade-off, and enhancing the model performance [22]. The effectiveness of these techniques depends on hyperparameters, which are set before the training process begins and remain static throughout. Hyperparameters include the learning rate, the number of layers in a neural network, and the number of nodes in each layer, among others. These parameters dictate the model's architecture, i.e., the overall framework that defines how the model operates. For instance, more layers can

increase a model's capacity to learn complex patterns, reducing bias but potentially increasing variance if the model becomes too complex. Regularisation techniques, such as adding a penalty for large coefficients, can help reduce variance by promoting simpler models that are less likely to overfit.

Proper tuning of hyperparameters is crucial for achieving an optimal balance between bias and variance, ensuring that the model performs well on both seen and unseen data. Figure 3.1 provides a graphical representation of overfitting, underfitting, and a well-balanced model, illustrating the effects of these adjustments.

3.2 Feed-forward Neural Networks

FFNNs are function approximators that in theory can perform any arbitrary mapping from one vector space to another [23]. Their structure closely resembles that of a human brain, consisting of interconnected nodes, also known as neurons. Each neuron, representing a vector-to-scalar function, processes input vectors of arbitrary size [24]. These nodes are organised into layers, where the output of one neuron is transmitted to all neurons in the subsequent layer. This layered arrangement is illustrated in Figure 3.2. The computations of neurons in a single layer are often combined using matrices, allowing the output $\mathbf{a}^{(l)}$ of layer l to be expressed by equation (3.2), where $\mathbf{a}^{(0)}$ corresponds to the input vector $\mathbf{x}$.

$$\mathbf{a}^{(l)} = g^{(l)}(\mathbf{z}^{(l)}) = g^{(l)}(\mathbf{W}^{(l)\top}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)}). \tag{3.2}$$

In this equation, $\mathbf{W}^{(l)}$ denotes the weights connecting layers l and $l-1$, and $\mathbf{b}^{(l)}$ represents the bias term. The function $g^{(l)}$ serves as the activation function, introducing non-linearity to the model. While the input to the function $\mathbf{z}^{(l)}$ is linear, the application of a non-linear activation function transforms this linear input into a non-linear output. This property enables the model to capture more complex non-linear relationships. Commonly used activation functions include the rectified linear unit (ReLU), the hyperbolic tangent (tanh), and the sigmoid function.

During the calculation of the model's output, equation (3.2) is sequentially applied L times, where L represents the number of layers, or the depth of the network. This process, known as forward propagation, involves the flow of information from layer to layer in a forward direction. The output of the network can be expressed as the composite function,

$$\mathbf{y} = f(\mathbf{x}) = f^{(L)}(f^{(L-1)}(\cdots f^{(1)}(\mathbf{x}))). \tag{3.3}$$

In order to optimise the network's prediction, we need a way to find the optimal weights and biases for the network. This is normally done using a gradient-based optimisation approach, where the gradient of the cost function C is used to update each individual weight. The cost function is just an average of the loss functions over a collection of samples. In order to calculate this gradient a backpropagation algorithm is applied. This enables efficient calculation of derivatives of the cost

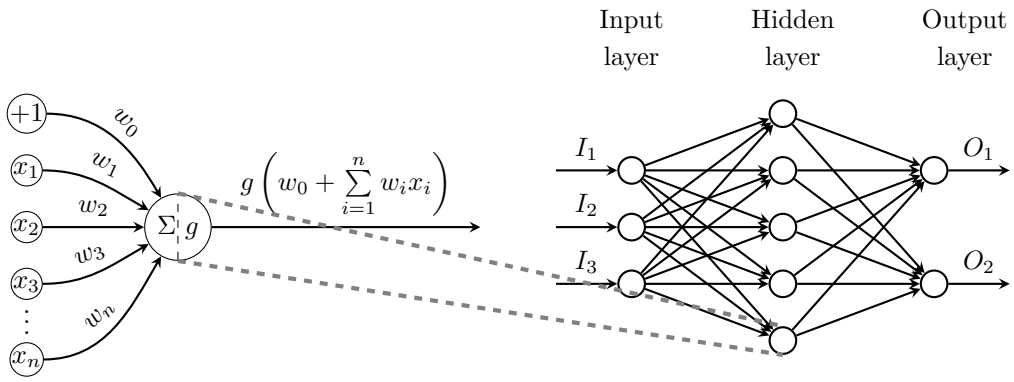


Figure 3.2: Illustration of a three-layered neural network (right). Each connecting line corresponds to a weight in the network, with circles representing neurons in each layer. The neuron computations are broken down and visualised in the left plot, with g acting as the activation function. Adapted from [25].

function with respect to each weight of the network. This is achieved by propagating the error of the output layer,

$$\boldsymbol{\delta}^{(L)} = \nabla_a C \odot g'(\mathbf{z}^L) \tag{3.4}$$

backwards through the network. Here, $\nabla_a C$ is a vector representing the rate of change of C with respect to the output activations. The error of a previous layer can be written as

$$\boldsymbol{\delta}^{(l)} = ((\mathbf{W}^{(l+1)})^T \boldsymbol{\delta}^{(l+1)}) \odot g'(\mathbf{z}^{(l)}). \tag{3.5}$$

The rationale behind this equation lies in the backward propagation of the error through the network. By multiplying the error at the $l+1^{\text{th}}$ layer with the transpose of the weight matrix, we effectively trace the impact of each neuron's output on the error, providing a measure of the error at the l^{th} layer's output. Subsequently, taking the Hadamard product ($\odot$) with the gradient of the activation function moves the error backwards through the activation function in layer l, giving us the error in the weighted input to layer l [26]. The partial derivatives of the weights and biases can then be calculated as a function of the error terms,

$$\begin{aligned} \frac{\partial C}{\partial b_j^l} &= \delta_j^l, \\ \frac{\partial C}{\partial w_{jk}^l} &= a_k^{(l-1)} \delta_j^l. \end{aligned} \tag{3.6}$$

Here, j denotes a neuron in layer l and k a neuron in layer $l-1$. The fact that the gradient only depends on the error terms and that the error terms themselves can be calculated recursively dramatically increases computational efficiency. To update the weights, an optimisation technique is used, commonly gradient descent. The idea

of gradient descent is that we move in the opposite direction of the cost function's gradient. The gradient gets multiplied by the learning rate η, which determines the size of each update step. A smaller learning rate results in smaller steps, potentially preventing overshooting, but it may lead to slower convergence, while a larger learning rate can speed up convergence but risks overshooting the optimal values. The rule for updating the parameters of the network according to gradient descent is:

$$\begin{aligned} \boldsymbol{W}^{(l)} &\leftarrow \boldsymbol{W}^{(l)} - \eta \nabla_{\boldsymbol{W}^{(l)}} C. \\ \boldsymbol{b}^{(l)} &\leftarrow \boldsymbol{b}^{(l)} - \eta \nabla_{\boldsymbol{b}^{(l)}} C. \end{aligned} \tag{3.7}$$

3.3 Recurrent Neural Networks

One limitation of FFNNs is their inability to capture temporal information. Each input is processed independently, and there is no internal state to remember information from previous inputs. In the context of processing sequences, such as sentences, the order of elements holds meaningful information that FFNNs are unable to capture. For example, when processing a time series, the same features at different time steps are separated into different features, and the temporal information is lost. In other words, the time-dependent relationship between features, which could be crucial for accurate predictions or understanding the data, is overlooked.

Recurrent Neural Networks (RNNs) aim to fix the loss of temporal information by processing data sequentially. The architecture of the network includes an internal state that retains information that the network has processed thus far. This is facilitated by a looped connection, which allows the internal state to be passed on to the next layer. The recurrent nature of this network is characterised by the ability to feed the outputs from preceding layers as inputs, in conjunction with the actual input at each time step. The internal state of the RNN is reset between two independent sequences so that they can be treated as separate data points [27]. In practice, this is not always the case as data is usually processed in batches.

An important attribute of RNNs is parameter sharing, as it allows for the network to process sequences of different lengths and generalise across them [24]. Simply put, parameter sharing involves applying the same weight matrices across all different time steps. This allows the model to capture relevant information independently of where in the sequence it appears [24]. In a single-layered RNN, there are three shared weight matrices [28]:

- $\boldsymbol{W}_{xh}$: The weight matrix between the input, $\boldsymbol{x}^{(t)}$ and hidden layer $\boldsymbol{h}$.
- $\boldsymbol{W}_{hh}$: The weight matrix associated with the recurrent connection.
- $\boldsymbol{W}_{ho}$: The weight matrix between hidden layer $\boldsymbol{h}$ and output layer $\boldsymbol{o}$.

The hidden state $\boldsymbol{h}^{(t)}$ functions as the RNN's memory. During training, it creates a compressed summarised representation of the information it has seen up until

the current time step. This summary is 'lossy' due to it mapping the information to a lower dimension [24]. The hidden state can be represented in two different ways: either recurrently, with the function f mapping a previous hidden state to the current one, or by a single function k mapping the entire input sequence $(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t-1)}, \boldsymbol{x}^{(t-2)}, \ldots, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)})$ to the current time step. Mathematically, this can be expressed as

$$\boldsymbol{h}^{(t)} = f(\boldsymbol{h}^{(t-1)}, \boldsymbol{x}^{(t)}; \boldsymbol{\theta}) = k^{(t)}(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t-1)}, \boldsymbol{x}^{(t-2)}, \ldots, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)}). \tag{3.8}$$

The advantage of the recurrent representation lies in the ability to learn a single model f without requiring a separate model k for each time step, enabling the parameter sharing as discussed above [24]. Graphically, we can visualise the RNN architecture in two ways, either as a compact circuit or as an unfolded graph, both of which are depicted in Figure 3.3.

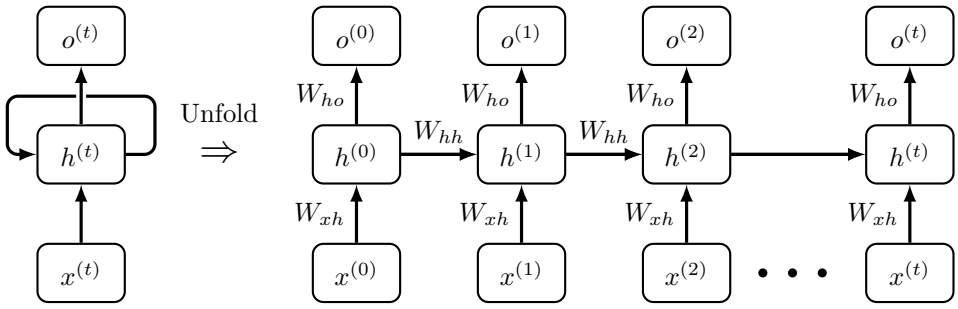


Figure 3.3: Graphical representation of a recurrent neural network, both as a circuit diagram (the left side), and as an unfolded graph (the right side) showing the application of shared weight matrices across time steps.

The computation of activations in RNNs is very similar to that of FFNNs. Initially, the net input $(\boldsymbol{z}_h^{(t)})$ is calculated as a linear combination, "that is, we compute the sum of the multiplications of the weight matrices with the corresponding vectors and add the bias term" [28]. After, the value is passed through an activation function,

$$\boldsymbol{h}^{(t)} = \phi_h\left(\boldsymbol{z}_h^{(t)}\right) = \phi_h\left(\boldsymbol{W}_{xh}\boldsymbol{x}^{(t)} + \boldsymbol{W}_{hh}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_h\right). \tag{3.9}$$

Similarly, when the hidden states are computed, the output activations can be calculated as

$$\boldsymbol{o}^{(t)} = \phi_o(\boldsymbol{W}_{ho}\boldsymbol{h}^{(t)} + \boldsymbol{b}_o). \tag{3.10}$$

Calculating the gradients of the network involves applying the standard backpropagation algorithm on the unfolded graph. While this may seem straightforward, it becomes challenging due to the output's dependency on inputs from previous time steps. The basic idea, however, is to minimise the overall loss, which can be written as

$$\mathcal{L} = \sum_{t=1} \mathcal{L}(t). \tag{3.11}$$

Given that the loss at time t depends on the hidden states of all previous time steps, the gradients of the recurrent weights can be computed as [28]

$$\frac{\partial \mathcal{L}^{(t)}}{\partial \boldsymbol{W}_{hh}} = \frac{\partial \mathcal{L}^{(t)}}{\partial \boldsymbol{o}^{(t)}} \times \frac{\partial \boldsymbol{o}^{(t)}}{\partial \boldsymbol{h}^{(t)}} \times \left(\sum_{k=1}^{t} \frac{\partial \boldsymbol{h}^{(t)}}{\partial \boldsymbol{h}^{(k)}} \times \frac{\partial \boldsymbol{h}^{(k)}}{\partial \boldsymbol{W}_{hh}} \right), \tag{3.12}$$

where

$$\frac{\partial \boldsymbol{h}^{(t)}}{\partial \boldsymbol{h}^{(k)}} = \prod_{i=k+1}^{t} \frac{\partial \boldsymbol{h}^{(i)}}{\partial \boldsymbol{h}^{(i-1)}}. \tag{3.13}$$

A common challenge with neural networks is the issue of gradients either vanishing or exploding, a problem that arises from the numerous multiplications performed during the backpropagation process. This problem can be especially pronounced in RNNs, where backpropagation is extended across time steps. When sequences are long, the gradients can become incredibly small or extremely large. The primary repercussion is that the models' weights and biases are not optimally updated, leading to neglect of the initial parts of the sequence. This issue is exemplified in equation (3.13), where there are $t - k$ multiplications, essentially multiplying the recurrent weight matrix by itself $t - k$ times. If the largest eigenvalue of this matrix is less or larger than one, the gradients will either vanish or explode, respectively. A detailed explanation of the Backpropagation Through Time (BPTT) algorithm, which is utilised for RNNs, is provided in [24].

There are several methods to alleviate the issues of vanishing and exploding gradients in neural networks. One such method is gradient clipping, which sets a limit on the gradients by establishing a specific threshold. If this threshold is surpassed, the gradients are assigned this threshold value, effectively being clipped. Another strategy is truncated backpropagation, which limits the number of steps that gradients are allowed to backpropagate through [28].

3.4 Long Short-Term Memory

A different approach aimed at addressing the gradient problems in traditional RNNs is the LSTM model. Unlike conventional RNNs, an LSTM alters the cell architecture by introducing a cell state, which enables better gradient propagation throughout the network. The LSTM architecture was initially proposed by Hochreiter & Schmidhuber in 1997 [29], and has demonstrated good performance across a diverse range of sequential modelling tasks [24].

The LSTM operates through gated mechanisms, that themselves are FFNNs. These gates play a crucial role in determining information flow within the network, deciding what information to retain, discard, add, and output. Sequences are processed sequentially, just like traditional RNNs. However, the internal operations at

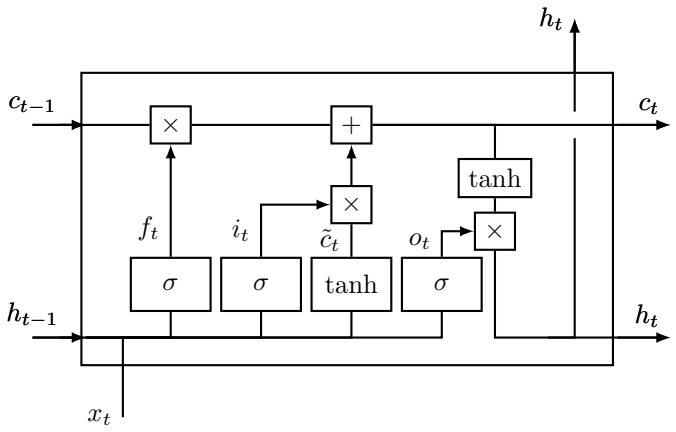


Figure 3.4: Figure illustrating the internal structure of a Long Short-Term Memory (LSTM) cell during a single time step. The hidden state from the previous time step h_{t-1} get concatenated with current input x_t, passed through the different gating mechanism, and updates the cell state c_{t-1} that carries the long-term information through the network.

each time step (in each cell) differ. Each LSTM cell is comprised of three different gates, the input gate, the forget gate, and the output gate. These gates interact with the cell state, $\boldsymbol{C}^{(t)}$, which functions like a conveyor belt, transporting relevant long-term information throughout the network. This mechanism allows information from the sequence to be added to the cell state at any point, carried forward to later time steps, and removed unchanged when needed [27]. A graphical depiction of the LSTM cell is illustrated in Figure 3.4

The role of the forget gate $\boldsymbol{f}_t$ is to determine which information should be discarded from the cell state. Information from both the previous hidden state $\boldsymbol{h}^{(t-1)}$ and the current input $\boldsymbol{x}^{(t)}$ gets passed through a sigmoid activation function σ. This activation function compresses values to a range between 0 and 1. Subsequently, when element-wise multiplied with the cell state, these values represent a percentage of information to retain. This mechanism prevents the cell state from growing indefinitely. The forget gate is mathematically expressed as

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{xi}\boldsymbol{x}^{(t)} + \boldsymbol{W}_{hf}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_f). \tag{3.14}$$

The input gate is responsible for determining what new information to add to the cell state. It comprises two components: a filter i_t, similar to the forget gate, which determines the relevant information to be input, and the computation of a candidate state $\widetilde{\boldsymbol{c}}_t$. This candidate state captures information from both the input and the previous hidden state, and a tanh activation function squeezes its values between -1 and 1 to regulate the network. When multiplied together these two components represent the pertinent information to be added. The equations for the input gate are defined as

$$\begin{aligned} \boldsymbol{i}_t &= \sigma(\boldsymbol{W}_{xi}\boldsymbol{x}^{(t)} + \boldsymbol{W}_{hc}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_i), \\ \widetilde{\boldsymbol{c}}_t &= \tanh(\boldsymbol{W}_{xc}\boldsymbol{x}^{(t)} + \boldsymbol{W}_{hi}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_c). \end{aligned} \tag{3.15}$$

The interaction between these two gates and the previous cell state determines the new cell state, which is then passed to the succeeding cell. The calculation of the new cell state is expressed as

$$\boldsymbol{c}^{(t)} = (\boldsymbol{f}_t \odot \boldsymbol{c}^{(t-1)}) \oplus (\boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t). \tag{3.16}$$

The final gate is the output gate. This gate regulates the amount of information that is transferred from the cell state to the hidden state. The hidden state is used for outputting predictions and also passed to the next cell. The gating mechanism $\boldsymbol{o}_t$ works in the same way as for the two others, but here it regulates the flow of information from the cell state $\boldsymbol{c}^{(t)}$. The hidden state is then calculated as

$$\boldsymbol{h}^{(t)} = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}^t), \tag{3.17}$$

where

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{xo}\boldsymbol{x}^{(t)} + \boldsymbol{W}_{ho}\boldsymbol{h}^{(t-1)} + \boldsymbol{b}_o). \tag{3.18}$$

LSTMs represent one approach for modelling long sequences, and a plethora of variations on this architecture exist. In their 2015 study, Józefowicz et al. [30] attempted to investigate potential modifications to the LSTM equations with the intent of boosting performance. However, their empirical results did not show significant enhancements over the 'conventional' LSTM model. In the same study, the researchers also tried to optimise an alternative model, termed the Gated Recurrent Unit (GRU). The GRU is a simplified version of the LSTM, which consolidates the two separate states of hidden and cell into a single state. Although their inner structures differ, both GRUs and LSTMs operate on similar principles, relying on gating mechanisms.

3.5 Extreme Gradient Boosting – XGBoost

XGBoost leverage a collection of weak learners, typically regression trees, to form a robust ensemble model, ϕ. Unlike bagging, where learners are built independently and their predictions are averaged, boosting is inherently sequential: each new learner is constructed to address residuals from previous ones, thereby progressively reducing the prediction error.

Given a dataset $D = \{(x_i, y_i) \mid i = 1, \ldots, n; x_i \in \mathbb{R}^m, y_i \in \mathbb{R}\}$, the prediction for each instance i using an ensemble of K trees is,

$$\hat{y} = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \tag{3.19}$$

where f_k denotes the k^{th} tree's contribution. The objective of this ensemble model is to minimise the following objective function:

$$\text{obj}(\phi) = \sum_i \mathcal{L}(y_i, \hat{y}_i) + \sum_k \Omega(f_k), \tag{3.20}$$

in which $\mathcal{L}$ represents the loss function evaluating prediction accuracy and Ω is the regularisation term, designed to punish model complexity and mitigate overfitting. Specifically, the regularisation term Ω for a tree f_k is defined as

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda ||w||^2, \tag{3.21}$$

where γ and λ are regularisation parameters that govern the penalty on the number of leaves T in the tree and the magnitude of leaf weights w, respectively. During optimisation, for the j^{th} iteration, the aim is to find a regression tree that minimises,

$$\text{obj}^j = \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{y}_i^{(j-1)} + f_j(x_i)) + \Omega(f_j). \tag{3.22}$$

The above equation can be approximated using a second-order Taylor expansion, from which a relation can be derived for evaluating the gain in the objective function due to a node split [31]. Let I_L and I_R denote the set of instances assigned to the left and right child nodes after a split, and g_i and h_i represent the first and second derivatives of the loss function with respect to the prediction $\hat{y}_i$, the improvement from a split can be expressed as

$$\text{gain} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} + \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma. \tag{3.23}$$

This equation is used to find the best split at a given node. The algorithm can be used to optimise any loss function, given that it is twice differentiable.

In addition to incorporating a regularisation term into its loss function, XGBoost implements additional measures to combat overfitting. It moderates the effect of each individual tree by applying a constant factor, η, to scale down their weights. Additionally, it employs column sampling, selectively utilising only a subset of features for constructing new trees, thereby diminishing the potential of overfitting by reducing the model's reliance on any single feature.

3.6 Explainable Artificial Intelligence – XAI

Despite widespread adoption and strong performance in many predictive tasks ML models often operate as black boxes. These models are akin to extremely complicated functions with thousands or millions of parameters, transforming inputs into outputs in ways that are not transparent. Unlike mathematical functions, where relationships between inputs and outputs are clear and predictable, the internal

workings of complex models like Deep Neural Networks (DNNs) are opaque. XAI tries to solve this problem by making the ML models more transparent, such that the 'reasoning' behind their outputs can be better understood.

One prominent XAI technique that addresses the challenge of understanding individual predictions of ML models is the SHapley Additive exPlanation (SHAP) framework, developed by Lundberg et al. [32]. This model-agnostic approach can explain the output of any black-box model, with certain implementations optimised for specific architectures to enhance computational efficiency. At the core of SHAP lies the concept of Shapley values, a concept from cooperative game theory, which we will delve into in the following section. Furthermore, SHAP extends its utility to providing global explanations by aggregating Shapley values across multiple instances, thus offering a broader perspective on model decisions.

3.6.1 Shapley Values

Shapley values is a concept from cooperative game theory where players work together to generate a common payout [33]. Shapley values give us a way to fairly distribute 'payoffs' between players, by assigning an attribution to each player in the game. The attribution among players follows some properties that ensures fair payout, that is: efficiency, symmetry, dummyness, and additivity [34].

The concept of Shapley values can also be adapted to ML, where a player represents an individual feature, and the payout is the model's prediction. The Shapley value of the j^{th} feature of an instance $\boldsymbol{x}$, when considering a model f, is the weighted average of all possible contributions this feature has to the model prediction. Essentially, this involves averaging the marginal contribution, the difference in the model's prediction with and without that specific feature, across every possible coalition of features. The Shapley value can be calculated as

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S_j}(\boldsymbol{x}_{S_j}) - f_S(\boldsymbol{x}_S) \right], \tag{3.24}$$

where S is the subset of the total feature space F excluding j and $S_j = S \cup \{j\}$ denotes the subset with feature j included. The function f_{S_j} indicates that the model is formed including the feature j, and conversely f_S without it.

Computing exact Shapley values is computationally demanding because it involves assessing an exponentially growing number of coalitions, $2^{|F|}$. For each of these coalitions, the model needs to be fitted anew, making the process impractical, even for scenarios with a not-too-large number of features.

3.6.2 Shapley Additive Explanations

SHapley Additive exPlanations (SHAP) is a method to generate post-hoc explanations of ML models. The method is based on SHAP values, which are Shapley values of a conditional expectation function of the original model. Instead of retraining the

model for each subset of features, SHAP values approximate the model's output by the expectation value of the original model conditioned on that specific subset. This implies that we can calculate the model's output ($f_S(\boldsymbol{x}_S)$) by

$$f_{\boldsymbol{x}}(S) = f(h_{\boldsymbol{x}}(\boldsymbol{z}')) = \mathbf{E}[f(\boldsymbol{x})|\boldsymbol{x}_S]. \tag{3.25}$$

Here $z' \in 0, 1^n$ is a binary coalition vector representing whether a feature is present or absent $(0, 1)$, and h_x is a mapping function from a coalition vector to actual feature values. S is the set of features that are present in z', such that $\boldsymbol{x}_S$ has absent features for those not included in S.

Calculating the conditional expectation function is a complex task. In the original SHAP paper by Lundberg et al. [32], several approaches are proposed for approximating SHAP values. These include both model-agnostic techniques and methods tailored to specific model architectures. The most popular model-agnostic method for approximating SHAP values is KernelSHAP. In KernelSHAP, handling the absence of a feature involves sampling random values from the marginal distribution and then averaging the model's predictions over these values. To put it simply, when explaining a particular instance and a chosen subset of features (coalition), the method retains the values of the features in the coalition from the target instance, while substituting values for the remaining features with random samples drawn from the background dataset.

TreeShap, which is implemented in this thesis, solves some of the main problems with SHAP values, that is, estimating the conditional expectation effectively, and the exponential nature of equation (3.24). It achieves this by employing a clever algorithm utilising the structure of the trees [35]. Instead of sampling from the marginal distribution, TreeShap approximates the conditional expectation directly. Importantly, there is no need to handle the absence of features, instead, features can be passed directly into the trees and average predictions from unreachable nodes. The algorithm reduces the computational time from exponential to polynomial.

The aim of employing SHAP is to create a simplified explanation model. The explanation model is merely an approximation of the original model, where the local accuracy property of SHAP ensures that the predictions of the simplified model align with those of the original model [32]. The simplified model is expressed as

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{3.26}$$

where M is the number of simplified input features. When this model is used to explain a specific instance, the coalition vector z'_i takes a value of 1 for all features. This allows us to interpret the prediction of the original model as a sum of real values attributed to each input feature. The term ϕ_0 refers to the model's baseline prediction, which is the prediction when no features are present. As a result, our model's predictions can be interpreted as the individual features contributions to the difference from the baseline.

SHAP Interaction Values

The TreeShap algorithm also comes with some valuable extensions, one of which is the SHAP interaction values. SHAP interaction values extend the idea of SHAP values by not only measuring the contribution of individual features to a prediction but also how pairs of features work together to influence the outcome. Essentially, they break down the prediction into parts attributed to each feature alone (main effect) and parts due to interactions between features (interaction effects). This helps to understand not just the importance of each feature but also how the pairs of features affect the predictions. The SHAP interaction values are given by

$$\Phi_{i,j} = \sum_{S \subseteq F \setminus \{i,j\}} \frac{|S|!(F - |S| - 2)!}{2(F-1)!} \Delta_{i,j}(S), \tag{3.27}$$

where

$$\Delta_{ij}(S) = f_x(S \cup \{i,j\}) - f_x(S \cup \{i\}) - f_x(S \cup \{j\}) + f_x(S). \tag{3.28}$$

The main effect is calculated by subtracting the SHAP interaction values, from the SHAP value of that specific feature,

$$\Phi_{i,i} = \phi_i - \sum_{j \neq i} \Phi_{i,j}. \tag{3.29}$$

When calculating SHAP interaction values, $F \times (F-1)$ interactions need to be evaluated, resulting in a computational complexity that scales approximately as F^2. This makes the computation slow, however, the efficiency of TreeSHAP helps alleviate this issue.

Global Explanations

SHAP provides local explanations by assigning an attribution to each feature for every individual instance. The greater this attribution is the larger impact that specific feature value has on the prediction. To derive global insights from these local explanations, one can average the absolute SHAP values across all instances in the dataset for each feature. This process quantifies the average impact of each feature across all the model's predictions, providing a measure of overall feature importance. The formula for computing this global importance score I_j for the j^{th} feature is given by

$$I_j = \frac{1}{n} \sum_{i=1}^{n} |\phi_j^{(i)}|. \tag{3.30}$$

These importance scores can be further normalised, yielding feature importance values that range from 0 to 1, thus making it easier to compare across different models. This metric serves as one of the key measures in the thesis for assessing the relative importance of each feature within our models.

4 Data

In this chapter, we set out to provide an explanation of the dataset used for modelling purposes in this thesis. The first section details each of the features and after we explain the preprocessing steps taken to make our dataset uniform. The techniques and assumptions applied are justified.

4.1 Data Description

Throughout our thesis, we have chosen to focus on the NO2 bidding area, located in southwestern Norway, as our primary area of prediction (see Fig 2.1). This selection is driven by the intricate market dynamics and pronounced price volatility within NO2, a result of its extensive interconnections with other European countries. These connections render NO2's energy market particularly sensitive to the European energy crisis, which has been a catalyst for notable increases in electricity prices since 2021. The surge in prices during this period is a primary focus of this thesis. We argue that analysing electricity markets in a connected BZN like NO2 offers more insights than examining more isolated northern zones, such as NO4, where the price dynamics differ substantially. The difference in prices between NO2 and NO4 is shown in Figure 4.1.

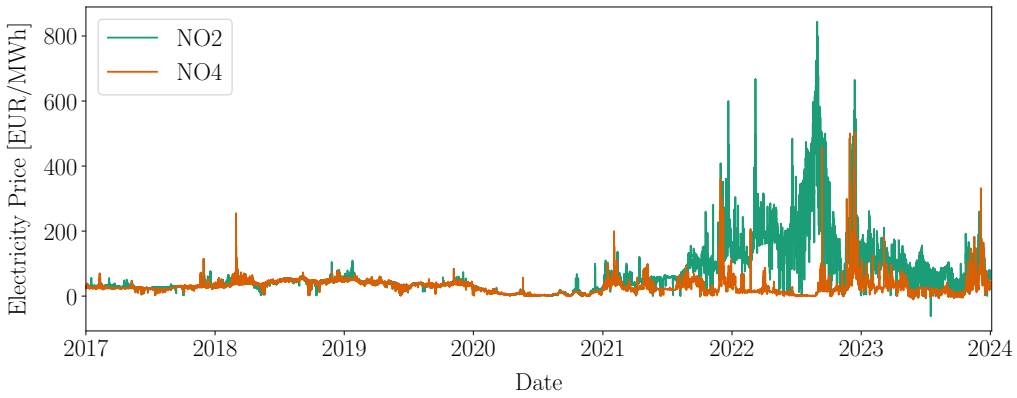


Figure 4.1: Day-ahead electricity prices for bidding zones NO2 and NO4 from 2017 to 2023. NO2 shows a pronounced increase in prices towards the end of 2021, marking the beginning of the energy crisis.

Our dataset includes electricity prices from all Norwegian BZNs (NO1 to NO5), as well as prices from adjacent zones that are electrically linked to the Norwegian grid. We have intentionally excluded Great Britain (GB) due to the prices being determined separately from the EUPHEMIA algorithm. We hypothesised that including it would disrupt our models, but this has not been empirically verified. Also, the data available for GB is of varying quality, containing a considerable amount of missing values. All price data was collected from the ENTSO-E transparency

platform [36].

In addition to price data, we also include a number of power system features in our dataset. For the Norwegian BZNs, this includes hydro filling levels, load forecasts, and generation forecasts. For neighbouring countries, we include the residual load. All these data are publicly available and collected from the ENTSO-E platform. The full list of features is presented in a correlation plot in Appendix A.2, which displays the Pearson correlation coefficients between variables across the entire data range.

The residual load is calculated by aggregating forecasts for wind and solar power generation and then subtracting the forecasted load, for the respective bidding area. This approach allows us to simplify the analysis by integrating what would otherwise be separate time series for wind, solar, and load forecasts into a single, composite residual load feature. For the German and Dutch BZNs, both load and renewable energy forecasts are provided at a quarterly resolution. This is addressed by resampling these forecasts to an hourly resolution using their mean value.

Generation data for renewable energy sources, excluding nuclear power, used to calculate the residual load in neighbouring countries, and the generation forecast in Norway, is not available until 18:00, after market closure. Despite this timing, we opt to include this data in our analysis. This approach is justified by the consideration that including such post-market closure data does not enhance the predictive accuracy of our model. This is because market participants make their decisions based on information available at the time of market closure. Consequently, utilising forecasts that become available later could, paradoxically, lead to less accurate predictions than those based on the forecasts market participants have access to, even though these are not publicly available

We also supply our dataset with fuel prices, including Dutch TTF Natural Gas Futures, sourced from Investing.com [37], and Brent crude oil prices, obtained from Federal Reserve Economic Data (FRED) [38]. Important to notice is that these fuel markets operate within specific trading hours on weekdays and remain closed on weekends and holidays. To avoid data leakage while making the data suitable for our models, we take the closing oil and gas prices and use that as a constant price for the subsequent day(s).

The last feature included in our dataset is the net position of NO2, labelled as 'Exports/imports'. The net-position is the netted sum of exports and imports i.e. the difference between load and generation. On the ENTSO-E transparency platform, energy transfers between BZNs are listed separately, with flows in each direction recorded as distinct series. We aggregate the flows for all BZNs directly connected to NO2. In this aggregation, exports from NO2 to neighbouring zones are marked as positive values, reflecting a net outflow of electricity. On the other hand, negative values indicate a net inflow or import of electricity to NO2.

The dataset spans a total of 7 years, covering the period from 2017 to 2023. This time frame was selected to encompass data from a variety of conditions, including seasonal changes and market fluctuations. The data is recorded with an hourly

resolution. Table 1 provides an overview of the dataset features, including the sources of the data and their respective resolutions.

Table 1: Overview of the dataset, including source information and original resolutions.

Data type	Hourly	Weekly	Other	Source
Electricity price	✓			ENTSO-E
Load forecast	✓			ENTSO-E
Generation forecast	✓			ENTSO-E
Hydro filling levels		✓		ENTSO-E
Residual load			✓	ENTSO-E
Net position	✓			ENTSO-E
Gas price			✓	Investing.com
Oil price			✓	FRED

4.2 Data Preprocessing

Creating a consistent dataset suitable for our ML models posed a considerable challenge in this thesis. Although the bulk of our data comes from a singular source, the data was of varying quality in terms of outliers and missing values. The data also contained different resolutions, that had to be accounted for.

The extent of missing values across our time series is detailed in Appendix A.1. Notably, none of our time series exhibits a high relative frequency of missing values. Furthermore, we did not encounter extended consecutive periods of missing values, except for initial phases in some series where data were not available. Given these characteristics, we determined that linear interpolation is an adequate method for addressing the missing values within our dataset. For initial and final missing values, i.e. at the start and end, we employ backward and forward filling respectively.

For hydro filling levels, we were constrained to using data with weekly resolution – less than ideal, yet the only publicly accessible information. To adjust the data to a hourly resolution, we opted for linear interpolation. This method, while straightforward, simplifies the complexity of real-world dynamics by assuming a constant rate of change throughout the week. Such an assumption does not hold up well against the unpredictable fluctuations in water levels caused by varying rates of precipitation and water usage within a single week. However, the data still carries some information about general trends and seasonal changes in hydro filling levels.

In handling our data, we addressed changes in the German BZN, which, before October 1, 2018, was combined with Austria and Luxembourg. From this date, the BZN split into two: a separate one for Austria and a new German BZN that included Luxembourg. Given Austria being its own control area, with separately available data on renewable generation and load forecasts before the split, we followed the

approach of Trebbien et al. [2]. This involved subtracting Austrian data from the previously joint BZN data and merging it with the data from the newly formed German BZN. Similarly, for the German price data, the post and pre-split prices were concatenated.

To detect outliers in our dataset, we apply a Hampel filter, a method based on a rolling window that calculates the Mean Absolute Deviation (MAD) within each window [39]. The reason for choosing this approach is to have a narrow window due to the strong non-stationary of our data. Using the MAD is advantageous as it is less influenced by extreme values, in comparison to other methods such as the standard deviation.

An observation is considered an outlier if it exceeds the MAD by a chosen threshold. This threshold is set conservatively to specifically target exceptionally abrupt variations in the dataset. This cautious approach is due to the occurrence of extreme yet genuine events, which complicates the differentiation between real events and actual outliers. We manually adjusted the threshold to ensure it captures only the most unrealistic data spikes.

We chose not to apply the aforementioned filter to the price data because the rapid and unpredictable nature of price changes makes it difficult to distinguish between true fluctuations and outliers. Moreover, we believe the price data to be more accurate than other time series in our dataset and thus argue in favour of its correctness.

Many of the covariates in our dataset are on different scales. It has been demonstrated that ML models, particularly neural networks, yield better results when features are scaled to a similar range [40]. This is especially true for algorithms that utilise gradient descent or other forms of sequential optimisation. When data spans widely varying scales, it can cause the loss function to 'stretch' in certain dimensions. This stretching can lead to unwanted oscillations during training, which might slow down convergence. To mitigate these effects, two common approaches are usually employed: *normalisation* and *standardisation*.

Standardisation involves centring the values around a zero mean with unit variance. This method makes a strong assumption that the variables adhere to a Gaussian distribution, an assumption that does not necessarily hold for our data. Therefore, we opted against standardisation in favour of normalisation.

Normalisation adjusts the features to a range between $[0, 1]$, without making any assumptions about the data's distribution. However, it is highly sensitive to outliers because it uses the maximum ($x_{\max}$) and minimum ($x_{\min}$) values as the scaling range's bounds. The normalisation process for each feature in the training set is conducted using the following equation,

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \tag{4.1}$$

5 Methods and Related Work

In this chapter, we delve into the ML models that have been employed in this thesis, outlining their development process in detail. We discuss the methods used for hyperparameter tuning and cross-validation, along with the recalibration strategies implemented for testing. Subsequently, we provide an overview of the evaluation metrics selected to assess the performance of these models. The chapter concludes with a literature review, which explores various approaches to EPF.

The first part of this chapter is divided into two distinct parts, a decision driven by the limitations of the SHAP python package. These limitations prevented us from obtaining satisfactory explanations from the LSTM model, which is our main model for EPF. Consequently, we were compelled to develop an additional methodology for the explanatory component of our thesis, which led us to the adoption of a tree-based model. This bifurcated approach allows us to address both the predictive and explanatory aims of our research.

5.1 Electricity Price Forecasting – LSTM

The primary aim of this segment of the thesis is to explore the potential of ML models for precise EPF in Norway, with a specific focus on the NO2 region. Our methodology is designed to align with the actual market timings, incorporating only the information that is available at the time of market closure to forecast a 24-hour series of prices, one for each hour of the following day. The LSTM model was chosen as our primary forecasting tool, due to its proven efficacy in EPF across various other markets.

To facilitate a robust evaluation of our model, the dataset was divided into a training set and a test set. The test set comprises data from the entire year of 2023. The rationale behind selecting a full year of data is to ensure that our model is thoroughly assessed across a wide range of conditions. This includes seasonal variations, which can significantly impact electricity demand and supply, and thus prices, as well as less frequent events that may lead to sudden price surges. This approach is in line with recommendations from Lago et al. [41] who advocate for a minimum of one year of test data to capture the full spectrum of dynamics and events that could influence electricity prices.

Due to us including forecasted and observed time series in our dataset, we shift the forecasted series backwards by 24 hours. This is so that our model incorporates the most recent available information at market closure into the training data.

For our LSTM model, the data shape had to be changed from its tabular format. That is because recurrent neural networks, such as LSTM models, require the input data to be in a sequential format to capture the temporal dependencies and patterns within the data. The amount of past time steps included in our model was selected as part of our hyperparameter optimisation. The resulting three-dimensional shape that is inputted to the LSTM reflects the batch size, time steps, and the number of

features.

During the training process, the Mean Absolute Error (MAE) was used as the loss function. MAE is one of the most popular evaluation metrics in EPF [41] and provides an intuitive metric measuring how far off the predictions are on average. Its linearity makes it well suited to quickly assess models' applicability in real-world applications. Particularly because trading profits or losses often depend linearly on the forecasting errors. The MAE is also less influenced by outliers than other metrics relying on squared errors, as it weighs them less. Ultimately using the MAE as our loss function, aligns the optimisation with the metric we deem most important.

The architectural design of our LSTM network was mainly shaped through hyperparameter optimisation. However, the choice of the optimiser (ADAM) was made based on proven effectiveness in existing literature [42]. The details of our hyperparameter optimisation approach are detailed in section 5.1.2, with the resulting LSTM architecture being shown in the Table 2.

Table 2: Resulting LSTM hyperparameters from optimisation.

Hyperparameter	Value
LSTM layer n_1	128
LSTM layer n_2	64
Dropout rate	0.0
Batch size	64
Time steps	96
Optimiser	ADAM

The final model architecture includes three stacked LSTM layers, followed by a final linear dense layer with 24 neurons, one for each hour of prediction. To combat overfitting, dropout layers are applied between each LSTM layer. Dropout regularises the network by randomly dropping units in the neural network. A simple intuition for why this works is that any single unit cannot rely too heavily on any of its inputs, because it might randomly get eliminated from the network. Consequently, the weights get spread out more between the different inputs, which tends to shrink the weights [43]. This is similar to L2 regularisation.

5.1.1 Cross Validation

Another measure taken to limit overfitting and ensure strong model generalisability is the application of cross-validation. The conventional approach of a static split into training, validation, and test segments does not adequately capture the model's performance across different time periods. This method falls short primarily due to the significant non-stationarity of electricity prices. Evaluating the model across a wide variety of time periods reduces the risk of our model overfitting to a particular subset of the data.

Our cross-validation strategy involves selecting consecutive eight-month periods at random from the training set, spanning from 2017 to 2022, to serve as our validation data. This approach is executed across 20 distinct validation splits, theoretically enabling us to assess our model against various segments of the dataset to achieve a better understanding of the models overall performance. However, while this method aims to provide a robust evaluation by exposing the model to diverse data conditions, it also introduces potential drawbacks, particularly concerning the disruption of the data's temporal continuity at the points where validation splits are made. Such disruptions might compromise the model's ability to accurately learn, given the importance of temporal continuity in time series analysis. Moreover, the random subsampling method does not guarantee a uniform evaluation across all data segments, raising concerns about the possibility of the model overfitting to specific periods. The selection of 20 sampling periods was mainly restricted by computational resources.

5.1.2 Hyperparameter Tuning

To optimise model hyperparameters, we utilise the Optuna Python package [44], a framework designed to streamline the optimisation process. While it supports various optimisation algorithms, including the commonly used grid and random search, these methods often fall short in terms of efficiency. Their main drawback is the lack of incorporation of information from past evaluations, leading to a significant amount of time being wasted on evaluating sub-optimal hyperparameters. Due to this fact, we try to utilise a smarter approach in search of optimal network parameters. That is, we use Sequential Model-Based Optimisation (SMBO), which is a type of Bayesian optimisation technique. More specifically, we use a Tree-structured Parzen Estimator (TPE) [45].

The core idea of SMBO is to utilise a surrogate model to approximate the objective function's behaviour. This surrogate model maps our hyperparameter search space to a probability score of achieving a specific objective function value. Different SMBO methods vary in their approach to constructing this surrogate model, $p(y|x)$, where y represents the actual value of the objective function when applying hyperparameters x. The TPE constructs this model using Bayes' rule:

$$p(y|x) = \frac{p(x|y)\,p(y)}{p(x)}. \tag{5.1}$$

Here $p(x|y)$ represents the probability of the hyperparameters given the value of the objective function, which can be expressed as

$$p(x|y) = \begin{cases} l(x), & \text{if } y < y^*, \\ g(x), & \text{if } y \geq y^*. \end{cases} \tag{5.2}$$

This equation forms two distributions based on the threshold value y^*. The 'good' distribution $l(x)$, and the 'bad' distribution $g(x)$. The expected improvement, used

for selecting the next set of hyperparameters, is chosen using the surrogate function, and given as

$$\mathrm{EI}_{y^*}(x) = \int_{-\infty}^{y*} (y^* - y)p(y|x)\mathrm{d}y. \tag{5.3}$$

By combining the aforementioned equations, it can be shown that the expected improvement is proportional to the ratio $l(x)/g(x)$ (for derivation, see [45]). This means, in order to maximise the expected improvement we need to maximise this ratio. Intuitively this makes sense, we pick values of hyperparameters that are more likely for $l(x)$ and less for $g(x)$.

The approach of modelling the good and bad distributions employs Parzen Estimators, from which part of the methodology's name is derived. The idea is to fit a distribution, in this case, a normal distribution, centred at each of the observations belonging to $l(x)$ or $g(x)$, and with a standard deviation equal to the maximum of its left and right neighbour. These individual distributions are combined to create the overall distribution in equation (5.2).

To evaluate the different hyperparameters, we combine the optimisation with our cross-validation approach. For each set of hyperparameters, the model is evaluated across all the validation splits. In order to obtain a model that generalises well and performs consistently not only on one part of the data but also on data with vastly different characteristics, we minimise two quantities. That is, the average MAE across the validation splits, along with the standard deviation. By minimising not only the error but also the standard deviation, we find a model that performs stably across various subsets. This entails that the model has learned underlying patterns in the data rather than adapting to specific instances, making it more likely to make accurate predictions on new, unseen data.

Addressing the challenge of optimising several objectives simultaneously is non-trivial. One method, scalarisation, combines multiple objectives into a single scalar value. However, this approach faces difficulties due to the different scales of the objectives and requires precise a priori knowledge to balance them effectively. An alternative is employing the concept of the Pareto front, a method in multi-objective optimisation that identifies a set of optimal solutions. The Pareto front represents a set of solutions in a multi-objective optimisation problem where no solution can be improved on one objective without worsening at least one other objective. This concept is illustrated in Figure 5.1, where the solutions forming the blue area represent the Pareto front. In our multi-objective scenario, rather than using an acquisition function to select hyperparameters based on their likelihood to minimise a scalar value, we adapt our strategy to maximise or minimise, according to our overall objective, the hypervolume defined by the Pareto solutions. Simply put, we modify the expected improvement function (EI_{y^*}) to the expected hypervolume improvement. The detailed mathematics are beyond the scope of this thesis, but are explained clearly in [47].

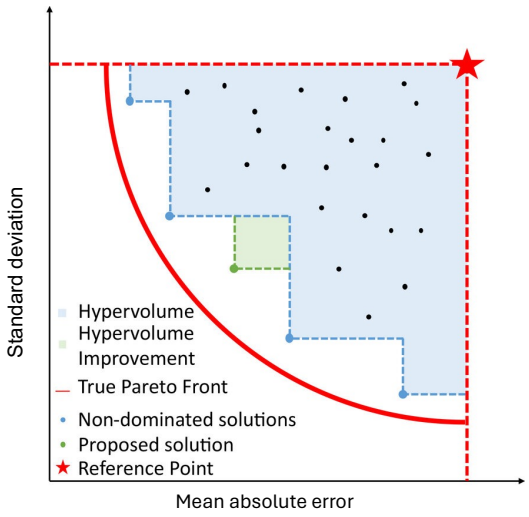


Figure 5.1: Illustration of a Pareto front optimisation problem focused on minimising two objectives, the mean absolute error and the standard deviation. The blue points and region indicate the current Pareto front and its respective hypervolume. The green point and area show the increase in hypervolume generated from a newly evaluated solution. The red line represents the ground-truth Pareto front, which is the best theoretically achievable set of solutions. Adapted from [46].

The use of the Pareto front for multi-objective optimisation is incorporated through the high-level API provided by Optuna. After executing the hyperparameter optimisation, we obtain a collection of Pareto-optimal solutions. Among these, we choose the solution with the lowest MAE, as we regard minimising the MAE to be our most critical objective.

5.1.3 Model Testing

In testing our models, we use a recalibration process that includes the latest data in our training set, aiming to better reflect real-world usage. This is achieved through an expanding window strategy, where the training set gradually incorporates data from the testing set, ensuring the model stays updated and relevant over time.

While a daily recalibration would be ideal for optimal accuracy, computational limitations necessitate a weekly recalibration approach. This means the model is retrained at the end of each week, utilising the updated model to forecast the subsequent week's electricity prices. The forecasting performance is then evaluated on a variation of evaluation metrics discussed in section 5.3. The expanding window approach is visualised in Figure 5.2.

In our testing procedure, we designate the second-to-last month of the training data as our validation set. This strategy ensures that the most recent month's data remains in the training set, which is crucial for the effectiveness of our model. To mitigate overfitting of our model, we incorporate early stopping on our validation data, setting the early stopping parameter to 20 epochs based on a heuristic approach.

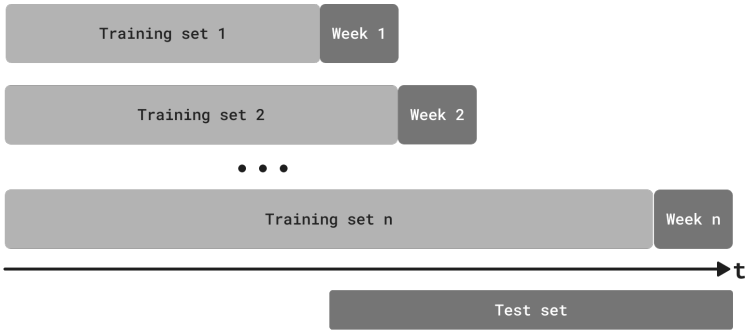


Figure 5.2: Visualisation of our recalibration approach using an expanding training set and a one week test set. The training and test sets are not to scale.

5.2 Explainable AI – XGBoost

In this part of the thesis, our goal is to explain the different market drivers before and after the energy crisis. To facilitate this objective, we employ a methodology where our model is trained on two distinct datasets: one comprising data from the period before the crisis, and another from the period following it. Given the ambiguous onset of the energy crisis, we opt to use the opening date of the Nord Link interconnector, 27 May 2021, as our demarcation point. This choice is due to the correlation with the observed surge in prices, as depicted in Figure 4.1. Our analysis is centred on computing SHAP values for models trained on these separate datasets, enabling us to dissect and understand the shifts in market drivers across the two periods.

Given that our primary focus is on explaining market dynamics rather than predicting future prices, we adjust our approach to simplify the forecasting challenge. Contrary to our earlier strategy with the LSTM model, which involved forecasting price batches, we now predict prices based on all the features in that point in time. This adjustment, while deviating from actual market timings, significantly enhances the accuracy and interpretability of our forecasts.

For the training process, we implemented a static division of our dataset into training (50%), validation (30%), and test (20%) sets. Given the pronounced non-stationarity of our data and the scarcity of similar data points across different segments, we opted to shuffle the data prior to splitting. This strategy was primarily adopted to mitigate the significant challenge of underfitting. Through experimentation with various shuffling 'splits', we discovered that a daily shuffle yielded the best performance, leading us to select this approach. Consistent with the rationale behind our LSTM methodology, we chose MAE as our loss function. Furthermore, to prevent our model from overfitting, we incorporated early stopping. This technique halts the training when there is no improvement in validation loss over a certain number of epochs, ensuring our model achieves better generalisability.

In modelling electricity prices for the NO2 area, we employ an XGBoost model, utilising the same dataset as employed in our LSTM model. However, we exclusively focus on 'exogenous' variables, excluding electricity prices from our data. Our

methodology refrains from incorporating lagged features, opting to include only data from the current hour. To evaluate our model's performance, we compare it against a simple Least Absolute Shrinkage and Selection Operator (LASSO) model. This comparative analysis is intended to verify the XGBoost model's ability to capture nonlinear effects, as well as to evaluate overall performance. Both models are implemented using the 'scikit-learn' package in Python [48].

To find the optimal hyperparameters we use the same hyperparameter optimisation approach as described in section 5.1.2. However, we do not implement cross-validation, and only evaluate the model's performance on a singular validation split, minimising only the MAE. The optimisation approach was carried out separately for both datasets. The hyperparameter search space is given in Appendix A.2, with the resulting optimal parameters given in Table 3.

Table 3: Hyperparameter optimisation results.

Hyperparameter	Pre-Crisis	Post-Crisis
Learning rate	0.01	0.08
Max depth	5	8
Subsample	0.80	0.75
Colsample bytree	0.70	0.57
Min child weight	10	3

5.3 Evaluation Metrics

In order to evaluate and quantify the forecasting performance of our ML models, we use various evaluation metrics with different characteristics. Among these is the previously mentioned MAE, and Root Mean Square Error (RMSE). Where the MAE weights all errors equally, the RMSE punishes larger deviations more severely as the difference between the predicted value $\hat{y}$ and the true value y is squared. The equations for calculating these metrics averaged across n observations, are given by,

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{5.4}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{5.5}$$

Since absolute error metrics can be challenging to compare across different datasets, we also incorporate a 'percentage' metric. Although the Mean Absolute Percentage Error (MAPE) is commonly utilised, its suitability is compromised in our case due to the frequent occurrence of prices near zero. This is due to MAPE's tendency to yield infinitely large errors or become undefined for values close to zero, which can skew the assessment of forecast accuracy disproportionately. In contrast, we utilise

the symmetric Mean Absolute Error (sMAPE), which ensures that its values are kept within a $[0\%, 200\%]$ range. This approach limits the disproportionate influence of smaller values. The equation of sMAPE is given by

$$\text{sMAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \frac{2|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|}. \tag{5.6}$$

Additionally, we include the relative Mean Absolute Error (rMAE), which was argued by Lago et al. [41] to be particularly suitable in the context of EPF. Relative measures such as rMAE are beneficial for their ease of interpretation, as they compare the performance of a model against a benchmark model, thus facilitating comparisons across datasets. The equation for rMAE is given as

$$\text{rMAE} = \frac{\frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i^{\text{naive}}|}. \tag{5.7}$$

It is important to note that the selection of the benchmark model, $\hat{y}_i^{\text{naive}}$, significantly impacts the rMAE. Although Lago et al. recommend using the previous week's prices for their simplicity and ability to capture weekly effects, we opted for a slightly more sophisticated benchmark model which we had already developed for performance comparison. This 'naive model' is defined as follows

$$y_{d,h}^{\text{naive}} = \begin{cases} y_{d-1,h}, & \text{if weekday,} \\ y_{d-7,h}, & \text{if weekend.} \end{cases} \tag{5.8}$$

An rMAE value lower than one indicates that the model on average performs better than the 'naive model'.

5.4 Literature Review

Research on EPF has been ongoing since the market liberation in the 1990s, and a variety of different modelling approaches have been suggested with varying degrees of success [4]. The different methodologies differ greatly not only in modelling approach but also in regard to forecast horizons and market specifics. Forecasting ranges span from short-term, such as minutes and days, to long-term looking months or even a year ahead, yet literature does not distinctly demarcate these horizons. This thesis focuses on forecasting prices in the DAM, hence our literature review is selectively focused on this area.

The models used for EPF can be divided into three main categories: statistical, ML, and hybrid methods [41]. Most statistical models rely on linear regression, which uses a linear combination of regressors/features to represent the predictive variable, electricity price.

In recent years there have been various developments in the area of statistical models, mainly linear regression models with a large amount of regressors. Whereas traditional regression models are optimised using ordinary least squares, for a large

number of regressors using the LASSO or elastic net as implicit feature selection methods have shown improved performance [41]. The strength of this approach is shown in a work from 2018 by Weron et al. [49], where the authors use auto-regressive (AR) models including various exogenous variables such as load and wind forecasts, to forecast electricity prices. This is done for two different markets, namely Nord Pool and PJM Interconnection (an American market), where they compare performance the performance of the ordinary AR models with the same models using LASSO. To similarly study the importance of the shrinkage property of LASSO, they develop models that only utilise LASSO for feature selection and not to optimise the model parameters. Their results show that providing a large amount of regressors is not a problem for the LASSO procedure and that increasing the number of variables seems to substantially improve predictive performance, even though LASSO only uses a small amount of the exploratory variables.

In addition to parameter-rich regression models, the field of statistical modelling has also seen other developments. Among these is the inclusion of variance stabilising transformations in the preprocessing step, as exemplified in the aforementioned research by Weron et al. [49]. Previously favoured logarithmic transformations have been substituted by other variance stabilising transformations, a shift necessitated by the occurrence of negative prices which render the former inapplicable [41].

The field ML for EPF has seen a substantial increase in interest since 2016. However, as highlighted by Lago et al. [41] in their 2021 work, much of the research within this domain suffers from limitations, including a lack of rigorous comparisons with state-of-the-art models and inadequately sized testing datasets. In their work, the authors endeavour to rectify these shortcomings by establishing a set of best practices for the EPF field.

One critical contribution of their work is the identification of state-of-the-art benchmark models through a comprehensive literature review. For statistical forecasting methods, they endorse the Lasso Estimated Auto-Regressive (LEAR) model as a highly accurate benchmark. When it comes to ML models, they recommend a simple two-layer DNN, noting that while complex models like LSTM have demonstrated potential, there was, at the time of their study, insufficient empirical evidence to advocate for their superiority.

To facilitate fair and consistent comparisons between new models, Lago et al. [41] provide a large open-access benchmark dataset encompassing a total span of 6 years. This dataset consists of five distinct subsets, each corresponding to a different electricity market, thus capturing a variety of market characteristics. Alongside historical electricity prices, each market dataset includes forecasts of two exogenous variables that are crucial for accurate EPF in the respective market. Lago et al. rigorously test various configurations of their benchmark models across these datasets. A noteworthy aspect of their methodology is that for the DNN models they use binary variables for feature selection. The choice of these variables is included in their hyperparameter optimisation process, which relies on TPE. This method allows for the

inclusion or exclusion of specific features, ensuring that the DNN models are tuned not just in their architecture but also in the choice of input data. The DNN model is made available through an open-access library [41], serving as a benchmark model for future research.

In 2022, Wagner et al. [50] introduced a novel method for forecasting electricity prices, utilising DNNs with an embedding layer for encoding calendar information, inspired by word embeddings. Their empirical analysis of the German electricity market showed that incorporating calendar information significantly enhances DNN model performance, with the embedding approach being competitive with, and occasionally outperforming, established benchmarks like those from Lago et al.. The paper notes that LSTM models underperformed, which might attributed to the lack of model architecture optimisation. The LSTM hyperparameters were selected solely based on existing literature.

A recent advancement in ML models for EPF is the integration of XAI. Although ML models have demonstrated exceptional predictive accuracy, their 'black-box' nature makes them difficult to interpret. In order to explain model predictions, and understand the underlying dynamics of electricity markets, XAI tools like SHAP and LIME have been employed.

The first work combining EPF with XAI was published in 2022 by Tschora et al. [51]. This work extends the benchmarking framework created by Lago et al. by including different unused predictive features and considering more recent data. It zeroes in on three markets previously examined in Lago and colleagues' work: Germany, Belgium, and France. The study explores a range of model architectures, such as DNN, Support Vector Regressor (SVR), Random Forest Regressor (RFR) and Convolutional Neural Network (CNN), applying them to both the original dataset from Lago et al. and an enriched dataset that incorporates additional features. These new features include renewable energy forecasts, prices from neighbouring countries, and gas prices, along with a cyclic encoding of calendar information to account for the models' lack of temporal awareness. The performance of the models is rigorously evaluated through the Diebold & Mariano test, with results demonstrating that an increase in predictive features consistently leads to better forecasting accuracy.

Finally, they try to identify the most important features by using SHAP values. They calculate SHAP values for all models, showing that the most discriminating feature is Swiss electricity prices, generation and load forecasts, and also that feature contributions vary greatly for the different markets. One important thing to notice is that Swiss prices are cleared before the other markets, and can therefore be used by market participants to form their order books. The different time lags are also studied, and results show that the features with no lag, one-day lag or seven-day lag are most important for the model decision-making process.

In 2023, Trebbien et al. [5] employed XAI techniques, specifically leveraging SHAP values, to uncover the determinants of electricity prices in the German market. Their methodology centred around a Gradient Boosted Tree (GBT) model, in-

corporating a rich dataset featuring exogenous variables like renewable energy and load forecasts, imports and exports, and the ramps of these variables–where 'ramps' refer to the change in values between two-time points. The dataset was further enhanced with information on fuel prices, specifically oil and natural gas prices. To evaluate the GBT model's performance, it was compared to a simpler model that relied solely on residual load, described through a third-order polynomial function. The results clearly showed that the GBT model outperformed the simpler alternative, indicating its ability to capture several market effects that are neglected in the single-feature model. To study these effects, the author make use of SHAP dependency and interaction plots, visualising the different features SHAP values alongside their most important interaction effects. Furthermore, they also calculated feature importance scores by utilising SHAP values. Their analysis revealed that renewable energy and load forecasts are the predominant factors influencing electricity prices in the German market. The analysis also sheds light on how specific features impact prices differently, for instance, it was observed that imports tend to decrease prices, while exports generally contribute to an increase in prices.

In 2023, Demloj [52] introduced a novel approach to the field of EPF, focusing on a methodology that exclusively incorporates electricity price data from an expansive set of European countries, precisely 39. This approach leverages an LSTM model for predicting the electricity prices in the NO1 area, with the model's performance being compared to that of a persistence model and a DNN. While the primary aim was the prediction of prices in NO1, the adaptability of Demloj's methodology was further demonstrated by retraining the same model for price forecasts across all the different BZNs present in the dataset.

To explore the coupling between NO1's prices with the other, Demloj applied Local Interpretable Model-agnostic Explanations (LIME). The analysis showed some strange coupling effects, like German prices negatively influencing prices in NO1, and a strong dependency on Serbian and Croatian prices. Additionally, this procedure was replicated for six other selected European markets, where the influence of Serbian and Croatian prices remained prominent. The analysis also highlighted the strong impact of British and German prices across these markets. However, it was found that the most crucial features were those with direct electrical connections to the respective region.

There are however some obvious limitations to Demloj's study. Firstly, and perhaps most obviously, the forecasts are not very good. This is not necessarily an issue with the modelling approach but rather with strong non-stationary of the data, and the challenge of forecasting the unusually high electricity prices observed in 2022. Interpreting weak forecasts risks drawing misleading conclusions about the factors driving the electricity prices, as the interpretation is inherently limited by the forecasting quality. If the model performs poorly, essentially, one is explaining meaningless predictions. Additionally, many of the features in the dataset are strongly correlated, which might cause the spreading of feature importance among

these highly correlated features. These issues, amongst others, underscore the need for careful interpretation of the results.

The application of XAI in power systems is widespread, but in the context of electricity markets, research is relatively sparse, with only a few publications to date. Many markets remain under-researched, presenting substantial opportunities for further study.

5.5 Applications of Artificial Intelligence Tools

Throughout this thesis, generative AI tools have been employed to enhance the quality of the work and streamline routine tasks. Primarily, two tools have been utilised: ChatGPT-4 and GitHub Copilot. ChatGPT-4 has been used to improve the writing, offering grammar checks, sentence paraphrasing, and assistance in refining paragraph structure. It has also provided debugging support for coding tasks. GitHub Copilot served as a coding aide, offering code suggestions and assisting with straightforward coding tasks, such as plotting figures. These tools have modestly contributed to the efficiency and quality of the work presented.

6 Results

In this chapter, we present the key findings of our research. We begin by examining the results from our predictive model, which has been precisely aligned with the actual timings of the DAM. Following, we try to disentangle the main drivers behind the electricity prices, focusing in detail on the effects of fuel prices on our predictions.

Firstly, we present our attempt at EPF using an LSTM model. We discuss the hyperparameter tuning process, evaluate the model's performance across various metrics, and compare it to other predictive models. Secondly, we examine the performance of our XGBoost models and delve into the importance of different features. Key features are selected for further analysis, employing both simple correlation analysis and SHAP dependency plots to study their impact.

6.1 Electricity Price Forecasting with LSTM

The LSTM forecasting model is the result of an extensive hyperparameter optimisation process. Interestingly, the hyperparameter tuning yields a rather simple LSTM network architecture, as depicted in Table 2. The model is able to forecast electricity prices with an average error of 15.48 EUR/MWh. This performance is considerably better than the 'naive' forecasting model, which predicts the previous day's prices if it is a weekday, and the previous week's price if it is a weekend, and forecasts with an MAE of 18.07 EUR/MWh.

Subsequently, we developed a two-layered DNN model to compare against our LSTM model. This DNN underwent the same hyperparameter optimisation steps as the LSTM, with the search space given in Appendix A.3. The data, initially formatted three-dimensionally for the LSTM, was flattened and inputted directly into the DNN without the addition of extra calendar features or employing a feature selection strategy. The DNN's performance closely mirrors that of the LSTM model in all metrics, with an average error of 15.62 EUR/MWh.

We also compare our model to the benchmark DNN model from Lago et al. [41], which they have recommended as a standard for future research. Their model, like our DNN, features a simple two-layered architecture. However, they enhance their model with TPE for feature selection and integrate additional calendar features into the dataset. Compared to the benchmark DNN model from Lago et al. our model underperforms in all metrics. However, this comparison may not be entirely fair, as the Lago model addresses a different forecasting problem. Specifically, it incorporates all data available from the previous day into its predictions, rather than limiting data to the information available at market closure at 12:00. This makes the forecasting problem simpler, and the comparison not valid. The LSTM model is however not too far off its performance, which builds confidence in our model predictions.

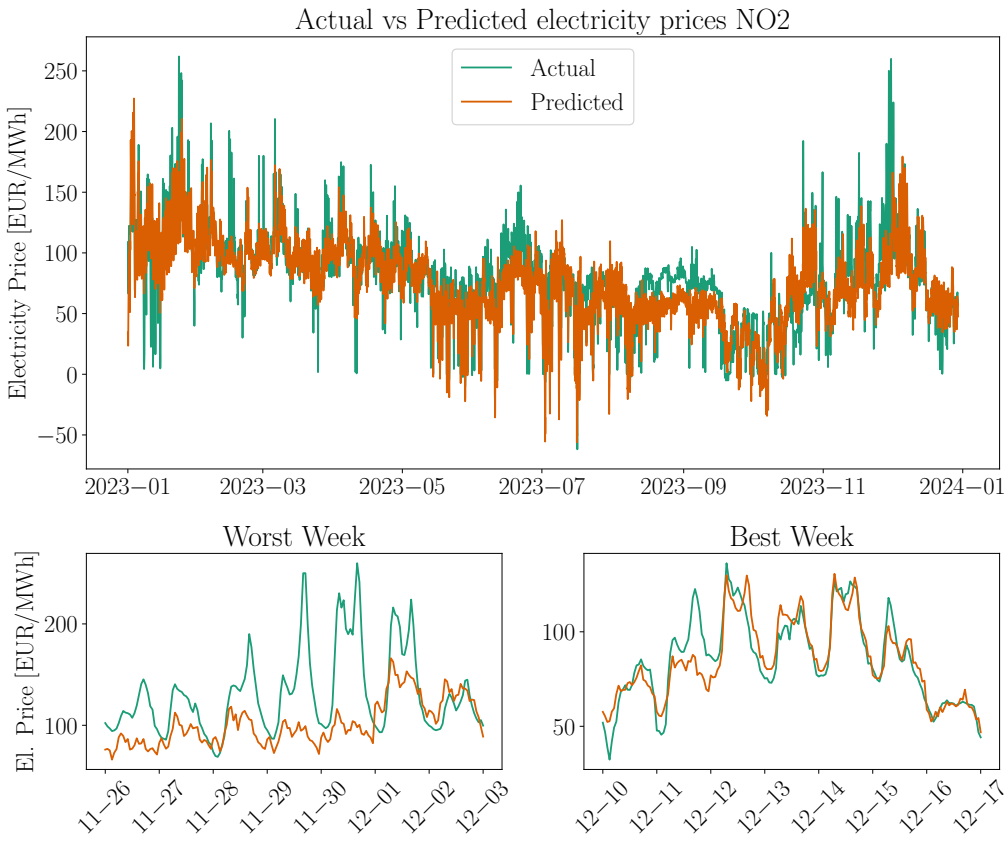


Figure 6.1: The top figure shows the LSTM model's annual predictions for 2023, with actual prices shown in green and predicted values in orange. The bottom two figures depict the week with the most accurate predictions, having a MAE of 6.9 EUR/MWh, and the week with the least accurate predictions, with a MAE of 34.9 EUR/MWh.

The performance of the LSTM, DNN, and naive model are all listed in Table 4, along with the Lago et al. model, which is added as a best practice more than for actual comparison. All models are evaluated with regard to the evaluation metrics presented in section 5.3.

Figure 6.1 presents the annual prediction results of our LSTM model on the test dataset, with our predictions in orange and the actual prices in green. This visualisation highlights the LSTM model's adeptness at tracking the general trends of electricity prices, though it struggles with accurately predicting sharp spikes or rapid changes in prices. The best and worst week is selected based on the average MAE. The plot of the worst week depicts the weakness of the model at predicting fluctuations when prices are at a high level, in this case, some of the spikes overshoot 200 EUR/MWh. In the best week, the overall price levels are lower and the model follows the actual prices much better while also predicting the daily fluctuations better. Our model captures the fundamental daily patterns in the data, exhibiting

lower electricity prices at night compared to daytime, and reflecting distinct price spikes in the morning and afternoon. Notably, the LSTM model also manages to predict negative price events, which are typically challenging to forecast.

Table 4: Performance metrics of forecasting models for the year 2023. Evaluated using an expanding window approach (see section 5.1.3).

Model	MAE [EUR/MWh]	RMSE [EUR/MWh]	rMAE [EUR/MWh]	sMAPE
LSTM	15.48	20.62	0.86	30.29%
DNN	15.62	20.93	0.86	29.81%
Naive	18.07	26.37	1.00	37.13%
DNN (Lago et al.) [41]	9.70	13.57	0.54	22.34%

6.2 Understanding Electricity Price Drivers using SHAP

In order to study the drivers of electricity prices before and during the energy crisis, we developed two XGBoost models, one for the pre-crisis data and one for the post-crisis period. After tuning the hyperparameters through our optimisation process, we assess the performance of our two XGBoost models using 100 random daily shuffles to ensure the models' generalisability. Our XGBoost models demonstrate strong forecasting accuracy, with pre-crisis predictions deviating by an average of just 1.96 EUR/MWh. Post-crisis, although the MAE increased due to generally higher price levels, the models continues to perform consistently in terms of the sMAPE.

Furthermore, the results highlight the superior capability of the XGBoost models in forecasting electricity prices, significantly surpassing our benchmark, the LASSO model, across all evaluation metrics (see Table 5). This superiority suggests that the XGBoost model captures important non-linear patterns in the data, which the LASSO model fails to detect.

Table 5: Summary of performance metrics for the two models, on dataset before (a) and after (b) the energy crisis.

(a) Pre-Crisis

Model	MAE	RMSE	sMAPE
XGBoost	1.96	4.22	11.99%
LASSO	4.64	6.91	41.35%

(b) Post-Crisis

Model	MAE	RMSE	sMAPE
XGBoost	16.15	29.90	12.09%
LASSO	35.53	51.77	29.20%

The XGBoost models take into account various different exogenous features. The overall impact of these features on the prediction is evaluated using SHAP, by aggregating local explanations to feature importance as detailed in section 3.6.2.

In Figure 6.2, the feature importance for the two models are shown for the top six features for each model. The feature importance indicates the degree to which a feature affects the model's predictions. The calculation of feature importance is based on the average magnitude of SHAP values, which reflects the overall impact of a feature on model predictions. This calculation disregards the direction of impact (whether positive or negative) and focuses solely on the strength of the relationship between each feature and the predictions. By normalising these values, we obtain values that range from 0 to 1, indicating the relative importance of each feature within the model. Higher normalised values indicate a larger impact of the feature on predictions, suggesting that the model relies more heavily on that particular feature for making decisions.

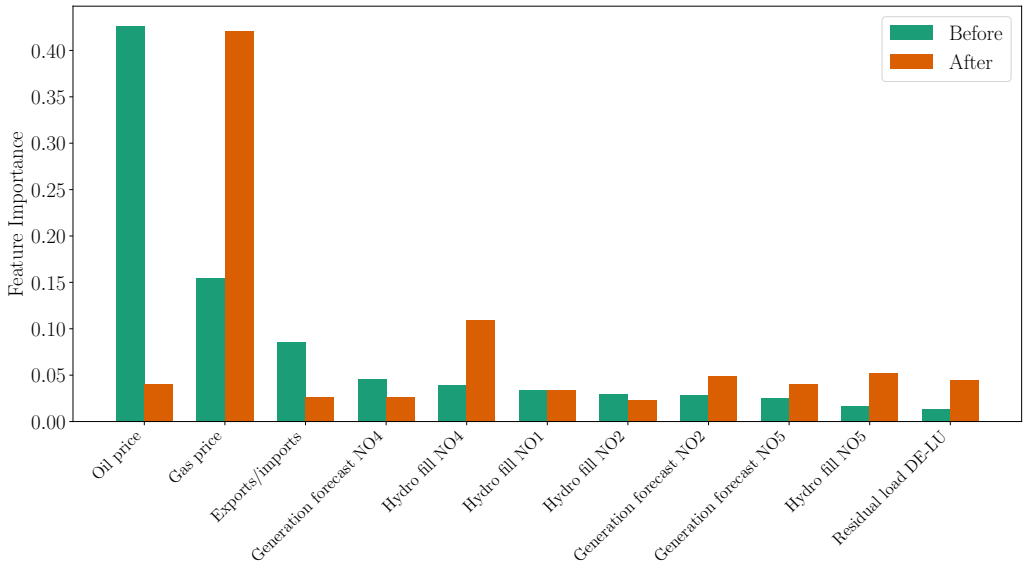


Figure 6.2: Comparison of the top six feature importances before and after the energy crisis for the two XGBoost models.

Figure 6.2 reveals that fuel prices are the most dominant features pre-crisis, with oil prices being the most important. After the crisis, gas prices overtake as the most important feature, whereas the importance of oil prices drops dramatically. The interchange in dominance between fuel price features is explored further in the discussion section. Additionally, the extent of power exchanged with other BZNs significantly influences model predictions prior to the crisis but sees a substantial decline in impact afterwards.

Another notable shift is observed in the reliance on the German residual load. Initially, this factor has minimal influence on the predictions, yet post-crisis, its significance nearly triples, indicating a heightened dependency. This makes some intuitive sense as the demarcation point is set to the opening of NordLink, where NO2 suddenly got direct transfer capacity to Germany via the underwater high-voltage DC cable. The overall dependence on German residual load is however small,

but one can notice that it is the only residual load feature that shows up in the figure, although residual loads from four other countries were included in our dataset.

Figure 6.2 yields some surprising results regarding the feature importance of generation, both within its own area, NO2, and across other Norwegian BZNs. Pre-crisis, the generation forecast for NO2 ranks as only the eighth most important feature, but it rises to the sixth position post-crisis. Interestingly, the load in NO2, which is typically pivotal in price determination by the EUPHEMIA algorithm due to its direct impact on market equilibrium between supply and demand, ranks near the bottom. This may suggest that the valuation of electricity production might change rapidly based on shifting market conditions, causing the actual load and demand balance to have little effect on prices. The same tendency is also seen for the remaining BZNs in Norway, where the generation and load seem to have little influence on prices in NO2.

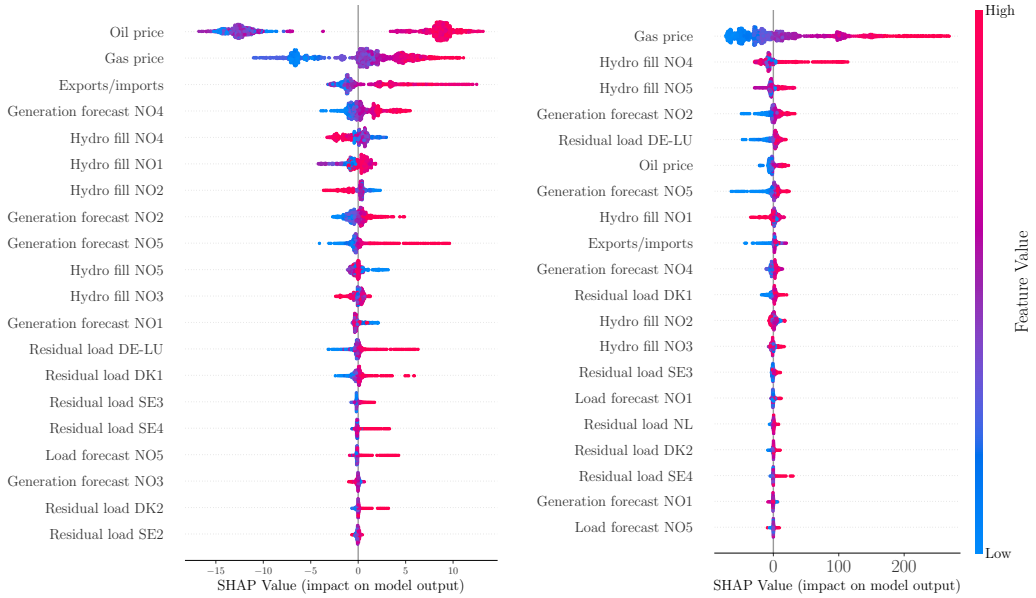
To further evaluate the effects of different features, we create a 'beeswarm' plot shown in Figure 6.3, where SHAP values for each individual instance and feature are displayed. Each dot on the figure represents one instance, with the colour indicating the relative value of the specific feature. The dispersion along the vertical axis highlights the density of predictions. With this beeswarm plot, we can discern how different feature values affect actual predictions.

A dramatic shift in the magnitude of SHAP values is evident across all features. Before the energy crisis, the maximum magnitude of the SHAP values was around 15 EUR/MWh, but this escalated to more than ten times after the crisis, reflecting significant price fluctuations and resulting in larger deviations from the baseline prediction (base SHAP value). Moreover, we observe distinct effects on gas and oil prices. In Figure 6.2 (a), the distribution of instances is clearly differentiated. Notably, oil prices seem to affect the model's predictions in almost a binary fashion, high oil prices significantly boost predictions, while low to medium prices tend to decrease them. The impact of gas prices displays a similar pattern, albeit split into three distinct categories: negative influence at low feature values, minimal impact at medium prices, and increased predictions at higher prices. Post-crisis, the same effects are no longer evident, with oil- and gas prices both having a more continuous distribution.

The data reveals a significant shift in how residual load values alter model predictions when comparing the periods before and after the crisis. Focusing on Germany, the residual load previously showed a strong positive correlation with price increases, a high residual load often coincided with higher electricity prices, while a low residual load did not have as marked an effect on decreasing prices. Interestingly, this relationship appears to reverse after the crisis, suggesting a change in market dynamics.

The underlying causes for this shift are not immediately clear, but it may be hypothesised that they are linked to the increased incorporation of renewable energy generation into the power grid. The substantial entry of renewables in recent

years, characterised by their negative price bids into the electricity market, could be influencing this relationship between residual load and pricing. Nonetheless, this explanation remains speculative.



(a) Pre-crisis

(b) Post-crisis

Figure 6.3: Beeswarm summary plot illustrating the impact of input features on the model predictions. Each dot represents an explained instance, with the colour indicating the relative feature value and the SHAP value depicted along the horizontal axis. The vertical dispersion reflects the density of instances.

Furthermore, we observe some odd effects with regards to the hydro filling levels in Norway. The effect of feature values seems almost random, altering greatly in both importance and in how filling levels influence prediction. Focusing on the hydro filling levels in NO4, which has a significant contribution after the crisis, high filling levels seem to increase the prices. This seems counter-intuitive as one could expect that power producers increase their water values when filling levels are low, and reduce them conversely. Pre-crisis however the effect seems to be the opposite, and more in line with our intuition.

For hydro filling levels in other Norwegian BZNs, there is no consistent trend in how these levels influence electricity prices, with varying effects observed both between different BZNs and in the pre- and post-crisis scenario. This inconsistency underscores the need for a cautious interpretation of the results, particularly given the low resolution of hydro-filling data, which includes only one measurement per week. With a test period spanning approximately five months, this frequency results in very few precise measurements, with the majority of data points being generated through interpolation.

To further disentangle the two most contributing features to electricity prices,

oil, and gas prices, we study them in more detail. In Figure 6.4, the electricity price in NO2, gas price, and oil price are plotted together. One can clearly see a strong similarity between gas prices and electricity prices, especially during the periods of high volatility. In 2021 and 2022, both gas and electricity prices show synchronous spikes, suggesting a strong correlation between them. For oil prices, the relationship with electricity prices is not as evident. While oil prices follow the same overall upward trend from 2020 to the middle of 2022, they do not contain the same spikes and characteristics of gas prices, and thus does not show the same degree of alignment.

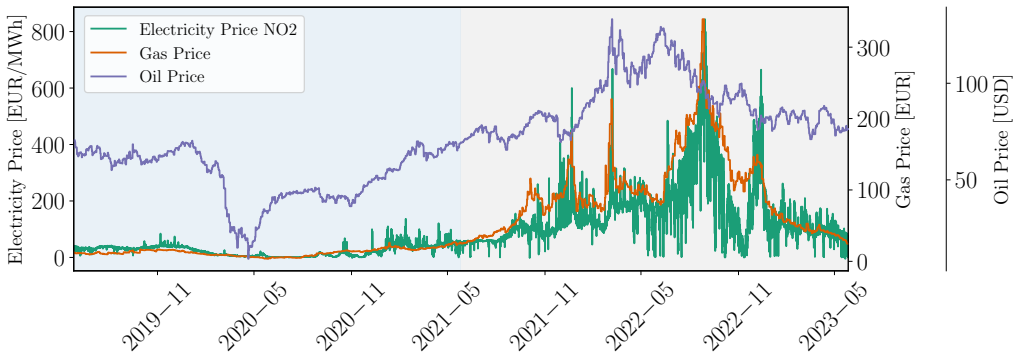


Figure 6.4: Triple vertical axis showcasing oil, gas, and electricity prices in NO2 from May 27, 2019, to May 27, 2023. The blue and grey background colours indicate the demarcation of our two datasets: pre- and post-crisis.

To quantify the relationship between oil and gas prices with electricity prices, we calculate the Pearson correlation coefficient, which measures the linear correlation between the variables. In Figure 6.5, the numeric value for the coefficient is given, along with a scatter plot. The scatter plots show gas and oil prices, pre-crisis in the top row and post-crisis in the bottom row. The orange line represents the best fitting linear regression line.

Pre-crisis, both gas and oil prices have a similar Pearson correlation coefficient of 0.74 with electricity prices, suggesting a strong linear relationship. The top row scatter plots reveal that as gas and oil prices increase, electricity prices tend to rise in a comparable fashion, as indicated by the upward trend of the orange best-fit regression lines.

In the post-crisis period, the correlation between gas prices and electricity prices strengthened, as evidenced by an increased Pearson correlation coefficient of 0.81. This is visible in the bottom left scatter plot, where the points are more closely aligned with the regression line. In contrast, the linear correlation between oil prices and electricity prices weakens substantially post-crisis, with the coefficient dropping to 0.35. The bottom right scatter plot illustrates this with a flatter regression line and

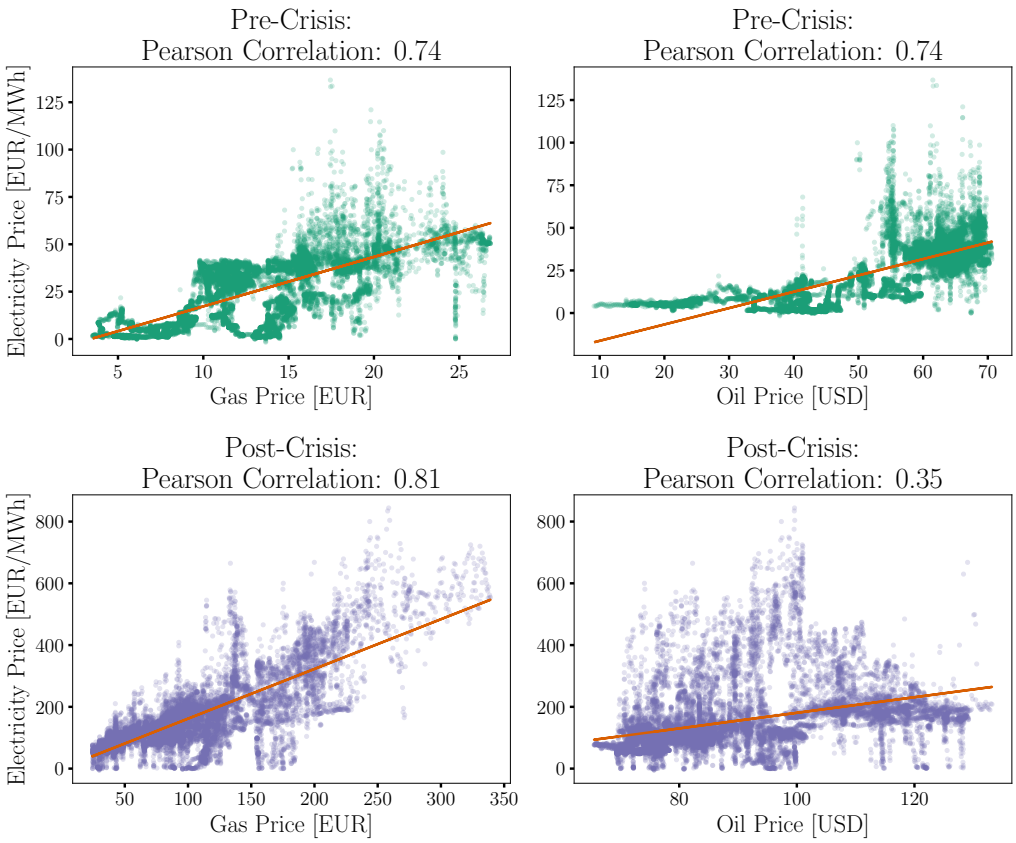


Figure 6.5: Correlation plots displaying the relationship of oil and gas prices with electricity prices. The orange line represents the best fitting regression line.

a broader distribution of points, implying that oil prices might have less predictive power after the crisis.

To provide a clearer visualisation of the relationship between oil and gas prices with electricity prices in our predictive models, we employ SHAP dependency plots. These are depicted in Figure 6.6, where the feature values are plotted along the horizontal axis with the corresponding SHAP values on the vertical axis. The top row, with green scatter points, represents the dependencies in our pre-crisis model, while the bottom displays the post-crisis model. Notably, in the pre-crisis oil price dependency plot, there is a pronounced non-linear relationship. It reveals that the impact of oil prices on electricity prices remains relatively stable above and below a certain threshold, approximately 55 USD. The two-state separation is somewhat contrary to what we would expect from our correlation analysis, where oil prices seemed to have a strong linear correlation with electricity prices. However, the ML model depicts a different relationship. The same tendency can be seen in the post-crisis model, but with a less pronounced separation and with a different threshold value at approximately 105 USD.

For gas prices, the relationship appears linear, with an increase in gas prices

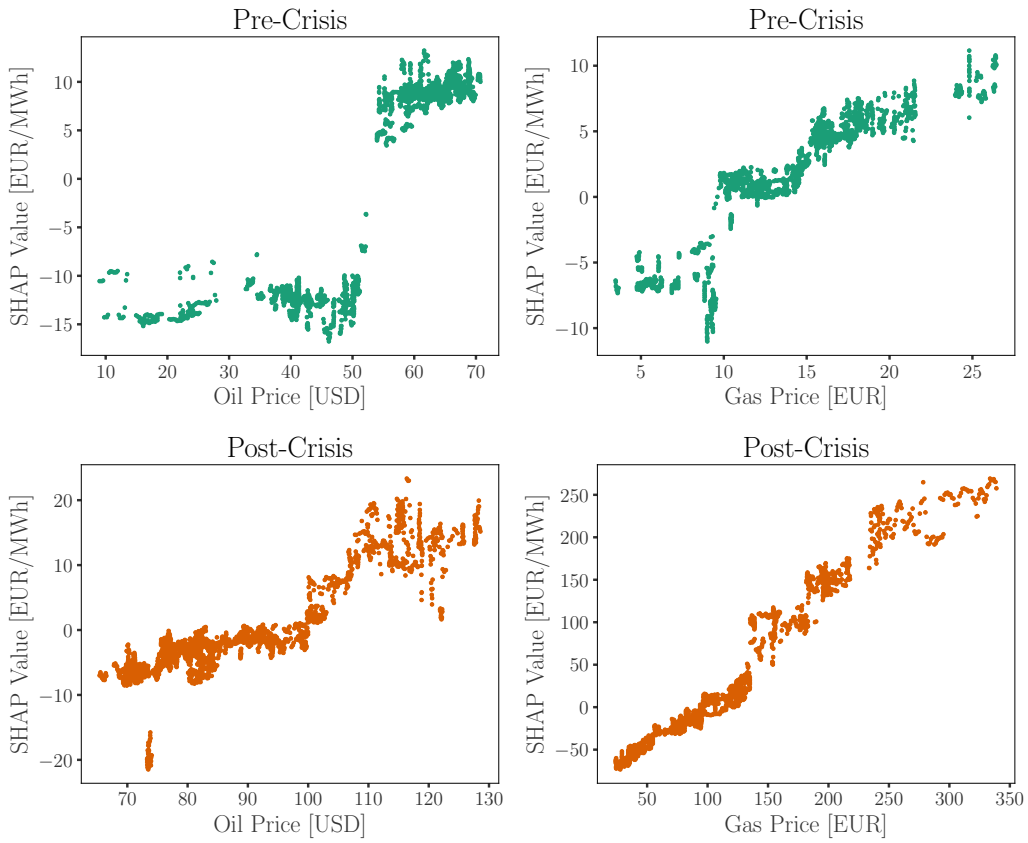
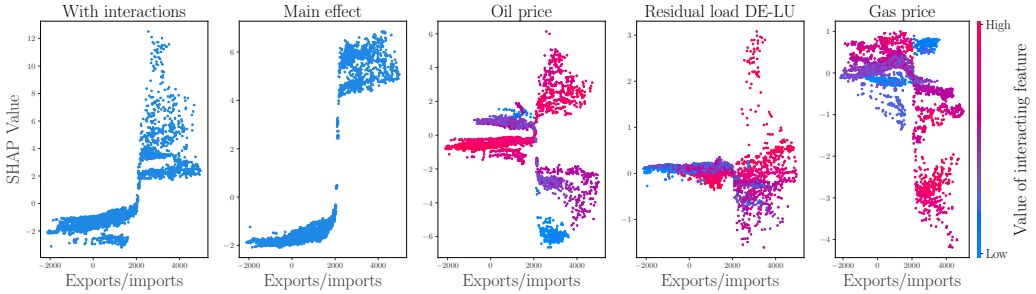


Figure 6.6: SHAP dependency plots for oil and gas prices before and after the energy crisis. Each point in the scatter plots represents an individual instance, with feature values on the horizontal axis and the corresponding SHAP values on the vertical axis.

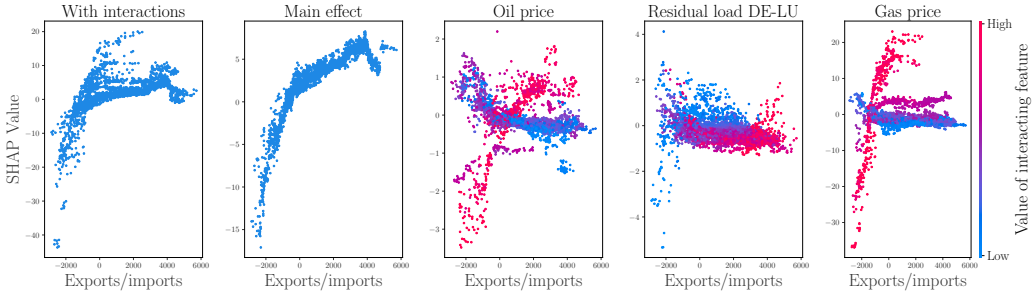
corresponding to a consistent change in the prediction of electricity prices. This trend holds true across both pre-crisis and post-crisis models, as seen in the SHAP dependency plots, although the linear relation is clearer in the post-crisis model.

Further analysis focuses on the impact of the net position, that is, exports and imports, on NO2 prices, which has been a contentious issue in the context of the energy crisis. The dependency plots exhibit some odd relationships and notable dispersion along the vertical axis, which suggests significant interaction effects with other features. Therefore, to better understand these relationships, we deconstruct the dependency plot into its main and interaction effects. Figure 6.7 displays this deconstruction, with the dependency plot inclusive of all interactions in the first column, the main effect of the exports/imports feature in the second, and the three most pronounced interactions in the remaining columns.

Pre-crisis, the main effect of the exports/imports feature shows a clear separation at exports above 2000 MW, where the dependency suddenly transitions from negative to positive. Beyond this threshold, there is also a pronounced increase in



(a) Pre-crisis



(b) Post-crisis

Figure 6.7: Dependence plot for exports/imports feature, i.e. the net position in NO2. The figure illustrates the net position's impact on the model predictions, decomposed into its main effect and the three most significant interaction effects.

scattering. The increased scattering seems to primarily stem from the interaction effects with oil and gas prices. In the oil price interaction plot, a similar shift is observed at high oil prices. However, for low and medium oil prices, the relationship reverses. The interaction effect is minimal for exports below the 2000 MW threshold, however, above this threshold, the oil price appears to strongly influence predictions in opposite directions.

For gas prices, the interaction effect is similar but opposite, where high gas prices influence the exports/imports feature negatively above the threshold. For medium prices, the effect is less pronounced, and for low prices, the interaction effect reverses. The interaction effects are also much stronger above the threshold, generating much stronger scattering in the dependency plot.

After the energy crisis, the dependency plot reveals two distinct quadratic trends, where the main effect clearly delineates one of them. For lower values, indicative of large imports, the negative influence on prices is more pronounced than what is captured by the main effect alone. This can be primarily attributed to the substantial interaction effect with gas prices. At higher gas prices, this interaction becomes particularly strong. Here, the influence of the export/import feature intensifies, negatively affecting prices when there are imports (values below zero), and positively when there are exports, thereby increasing prices.

Similarly, the interaction with oil prices follows the same trend, yet with a less pronounced SHAP value magnitude. Despite the interaction effect being weaker, these oil price interactions contribute to the overall dual quadratic trends observed in the dependency plot. As for the interaction with Germany's residual load, there appear to be no strong or distinct interaction effects present.

7 Discussion and conclusion

The final chapter of this thesis begins with a discussion of the main findings, comparing them with existing research and proposing conjectures for some of our observations. Next, it delves into a detailed examination of the limitations inherent in our methodology. The chapter wraps up with a conclusion and recommendations for future research.

7.1 Discussion of Results

In the first part of the thesis, we have shown the efficacy of using an LSTM for EPF in the Norwegian market, specifically NO2. The comparison was mainly made with a persistence model, a self-made DNN model, and the benchmark model provided by Lago et al. [41]. However, as previously argued, the comparison with the benchmark model is not one-to-one. Comparing the performance against other work is also not easy, due to little research on the Norwegian market. Most of the research mainly centres around central European markets where price dynamics differ from the Norwegian one, due to a widely different energy mix. Also, much of the published work on EPF has not tackled the problem of the energy crisis and the drastic increase in price levels and volatility. Either because the work was published before the energy crisis started, or because the authors of such works simply chose datasets predating the energy crisis. In addition, our approach of only including data until 12:00, i.e. the DAM closing time, seems to be quite novel as well. Previous work often includes all information of the previous day for forecasting, which simplifies the forecasting problem significantly, and fails to obey the real functioning of the DAM.

One of the most comparable studies is the master's thesis of Walid Demloj [52], which tackles the problem of surging prices head-on, and also validates his model on the NO2 market. Demloj focuses on forecasting for the year 2022, a period with price levels even higher than those we are evaluating in our forecasts for, 2023. The LSTM model developed by Demloj shows an MAE of 45.73 EUR/MWh, significantly worse than our forecasting model which has an MAE of 15.48 EUR/MWh. Important to point out is also that Demloj uses a different dataset than ours, only including price data in his model. The dataset also only includes data up until 2022, giving Demloj less useful historical data compared to us.

Research on other European markets appears to be within the same error range as our predictions. For example, the research of Trebbien et al. [42] reports an MAE of 10.32 EUR/MWh in 2021 and 29.85 EUR/MWh in 2022, and the study by Tschora et al. [51] achieves an MAE of 7.66 EUR/MWh, with a test set spanning from 2020 until 2022. It is important to note that the aim of this thesis was not to develop the most complex and highest-performing model for EPF, but rather to demonstrate that ML models are still viable under the volatile market conditions we now face. Our performance appears to be within a reasonable forecasting range when compared to the aforementioned studies.

In the second part of this thesis we looked into the inner workings of our XGBoost models by applying SHAP. From the feature importance plot, we observe fuel prices as the largest contributing factor to model predictions prior to the energy crisis. Interestingly, after the energy crisis, the importance of oil prices gets almost eliminated, whereas the importance of gas prices increases. This change is not immediately evident, but at least suggests that during the crisis gas prices become the primary driver of electricity prices. The strong dependence on oil prices prior might not be a true reflection of the market, but rather due to strong correlation between gas and oil prices (more on this in the limitation section). Moreover, although the influence of oil and gas prices on our model predictions is clear, it is crucial to note that these fuels do not directly impact Norways electricity production, as they are not used for power generation in Norway. Instead, their influence on electricity prices in Norway likely arises from the integration into the broader European electricity market, where such fuel costs can indirectly affect market prices through cross-border trading.

Our finding that gas prices are the main driver of electricity prices is further supported by the correlation analysis done as part of our study. Before the energy crisis, oil prices and gas prices demonstrated a strong correlation with electricity prices. However, following the crisis, the correlation with oil prices weakened significantly, whereas the correlation with gas prices increased, mirroring the changes observed in the feature importance analysis. This shift could potentially be attributed to the direct role that gas plays in electricity generation. Many power plants outside Norway rely on gas as their primary fuel source, which directly links changes in gas prices to electricity prices. The simultaneous increase in both gas prices and electricity prices might point to a causal effect during the energy crisis, where the increase in gas prices drives the electricity prices. However, with our current methodology, it is impossible to conclude definitively, other than to state the correlation.

Further studying the dependencies on oil and gas prices in our ML model, we see a clear non-linear dependence on oil prices, most evident pre-crisis, but also noticeable in the post-crisis model. From the strong linear correlation in before the crisis, one might expect the model to show a linear dependence, however this is not the case. This actually aligns well with previous findings from Trebbien et al. [2] which discovered a similar dependence on fuel prices in the German market. Important to point out, is that while using a similar methodology, applying SHAP to study market effects, Trebbien's model does not include any information about the Norwegian market.

The similarities in the dependence on fuel prices in both works might indicate a consistent market response to fuel price fluctuations in both countries. It is particularly intriguing to observe these similarities given that Norwegian energy production primarily relies on hydro-power, which has substantially lower operational costs compared to fuel-based power generation. This observation suggests that global fuel prices may influence market dynamics even in regions with different primary energy

sources, highlighting the interdependence in global energy markets.

The reason for this strong dependence on fuel prices in Norway is not immediately obvious. One might expect the filling levels in Norway's hydro reservoirs to be the predominant factor contributing to electricity prices, given that more than 80% of the electrical power generation in Norway comes from hydro power. It is important to point out, however, how the pricing of hydro power is managed. The cost of production from hydro power in Norway is almost negligible. To determine when to produce power from hydro reservoirs, they are assigned a water value that is used as their marginal cost. The determination of water values is not only determined by the amount of water in the reservoirs, but also significantly influenced by expected future electricity prices in other countries, which are themselves affected by broader energy market trends, including global fuel prices. Therefore, even though hydro power's direct production costs are minimal, the evaluation of the water might be strongly influenced by European market effects. This creates a somewhat indirect, but very real, linkage between global fuel prices and hydro power operations. A Master's thesis by Løfgren & Ingstad [53], highlights this effect. They show a strong correlation between the increasing gas prices and water values in southern Norwegian reservoirs in the late period of 2021 (the start of the energy crisis).

Our results also reveal a notable correlation between the net position and electricity prices, highlighting the impact of cross-zonal trading on electricity prices in NO2. In the pre-crisis model, we showed a peculiar non-linear relationship with a pronounced threshold effect at an export level of 2000 MW. The causality behind this threshold effect is not immediately obvious, indicating a need for further analysis to interpret this phenomenon. Additionally, the model exhibits intriguing interaction effects, where high oil prices seem to amplify the price increase for exports beyond the threshold, while lower oil prices tend to reduce them. Contrarily, for gas prices, this interaction effect is reversed, complicating any causal interpretation.

The dependency in the post-crisis model is widely different, and appears more conventional, aligning with what we would expect from a purely economical view. The dependency plot show that imports tend to decrease prices while exports increase them. From a supply and demand perspective, this makes sense, as exports indicate more accepted offers than what is needed to satisfy the local demand, and thus increases prices in the respective BZN. The opposite is the case for imports, where rather than utilising expensive offers in the local BZN, cheaper offers are excepted elsewhere.

The interaction plots post-crisis reveal consistent effects from high fuel prices, notably amplifying the dependence on exports and imports. This increase in dependency, particularly for exports, can be attributed to rising marginal costs, not only in markets reliant on gas power plants but also in the Norwegian market. In Norway, the valuation of water resources is adjusted upwards to reflect their true economic value. As a result, the heightened marginal costs in other markets enable Norway to generate higher profits from the same volume of exported electricity, as indicated

by the dependency plots.

While this scenario is beneficial for Norwegian power producers, who see increased profits, it poses significant ramifications for Norwegian consumers, who now face higher electricity prices. These prices do not necessarily reflect the local cost of producing electricity but are instead influenced by prices in other markets. This opens a range of new questions which are widely outside the scope of this thesis such as whether the economic benefits, namely the increased income outweigh the premium paid by Norwegian consumers.

It is important to note that within the current market structure, Norwegian power producers do not intentionally set high electricity prices to disadvantage Norwegian consumers. Instead, their pricing strategy aims to maximise revenue. If Norwegian electricity were priced below its true market value, it would likely lead to increased exports. This, in turn, could deplete reservoir levels, potentially leading to even more significant long-term adverse effects for Norwegian consumers.

7.2 Limitations

One interesting observation from our research is that the LSTM model does not outperform the DNN model, despite literature suggesting that LSTMs are typically more effective at capturing temporal dependencies in sequential data [54]. A potential factor contributing to this outcome could be the optimisation strategy we employed using the TPE. Although TPE is generally more efficient than random search by focusing on promising regions of the hyperparameter search space, a low number of trials may cause them to work similarly. With a rather large hyperparameter search space, the relatively low number of optimisation trials may have hindered our model's ability to adequately converge. The stochasticity of this process might cause it to find more optimal hyperparameters for the DNN than for the LSTM model. Ideally, to mitigate this uncertainty, we would increase the number of trials and repeat the optimisation process multiple times to validate the results. However, computational constraints made this more rigorous testing unfeasible.

Another potential issue with our optimisation process is the use of an inadequate number of validation splits, which might not sufficiently represent the entire training dataset. This limitation can lead the model to overfit to specific regions of our training data, potentially skewing the performance results. A more conventional approach, such as k-fold cross-validation, which divides the dataset into equal segments, might be preferable, when the number of validation splits are low.

With our SHAP analysis, there are two main limitations. Firstly, there is a problem with highly correlated features in our dataset. This is not just a problem with SHAP but any feature attribution method. When input features are highly correlated, it is difficult to fairly distribute importance values among features. The model might rely heavily on one of the features and thus get assigned a disproportionately high SHAP value, while the other equally informative feature, could receive a much

lower SHAP value. Simply put, correlation can dilute the attribution among correlated features. This might explain why there is an extremely high dependence on oil prices before our demarcation, as it could be 'stealing' attribution from gas prices, which might actually be the real driving factor. However, the different dependencies in the model, i.e. linear and non-linear, might point to this not being the case. High correlation might also influence other features, such as residual loads and exports/imports, which makes it hard to determine if a change in feature importance before and after the energy crisis is a true change or just that the model distributes it differently.

The second limitation concerns the use of SHAP for causal inference. As applied in this thesis, SHAP lacks the capacity to perform causal analysis, that is, to identify the true causes of events. It allows us to discern correlations between predictive variables and the target (electricity price), but these correlations should not be conflated with causation. While SHAP provides valuable insights into which variables significantly influence electricity price predictions, it does not enable us to pinpoint which variables actually cause changes in electricity prices. Moreover, the model may not accurately represent the actual market, and the selected features might merely act as proxies for other, unaccounted causal factors. Although some efforts have been made using 'causal SHAP' [55], these require a well-defined understanding of the causal structure, which is not apparent in complex markets like the electricity market.

Lastly, a significant limitation in our dataset concerns the varying resolutions of some of our features compared to the hourly resolution of our electricity price target variable. For fuel prices, we only have daily data, which we address by employing forward filling to generate new data points. This method assumes constant prices throughout the day, potentially overlooking intraday fluctuations. In the case of hydro filling, the data resolution is even sparser, with just weekly measurements available. To account for this, we use linear interpolation for resampling, applying a steady rate of change across each week. This approach seem insufficient, as wee see some anomalous relationships in our analysis of hydro filling data. Applying more sophisticated techniques to generate new data points, such as time statistical or ML models, could be considered, however, when the discrepancy in resolution necessitates the creation of a large number of data points, the efficacy of such methods becomes questionable.

7.3 Conclusion and Future Work

Drawing firm conclusions from our results is challenging due to limitations in our dataset and the inherent constraints of SHAP analysis. The SHAP analysis alone does not allow us to determine causal effects. Instead, we can only hypothesise causal relationships based on domain knowledge. However, with some degree of confidence we can state that gas prices plays a crucial role in the determination of electricity

prices in Norway, as it shows strong dependency both in our correlation analysis and in our ML models. Based on our dual dataset approach, we also find a strong increase in dependency after the energy crisis.

Furthermore we have discovered a strong reliance on oil prices, which before the energy crisis is the most important feature for our predictions. The relationship between oil prices and electricity prices shows a strong non-linear effect, which aligns with findings from previous works on other markets. The causal reason for this relationship is however not known, and requires further analysis.

We have demonstrated a notable reliance on exports and imports in our models. In the pre-crisis model, this feature was ranked as the third most significant, dropping to ninth post-crisis. Caution is advised when interpreting these rankings due to the high correlation with other variables, such as the generation forecast in NO2. This feature gained prominence post-crisis, potentially diminishing the perceived importance of the export/import feature.

A clear shift in dependency on exports and imports is evident when comparing the two models. Pre-crisis, the model exhibited unusual results and a distinct separation effect. Post-crisis, the reliance on the export/import feature appears more logical. Additionally, there is a marked interaction with fuel prices, underscoring a significant interaction between these features.

In the predictive segment of our thesis, the model demonstrates average results, which are consistent with the performance of similar models in other markets. However, direct comparisons with other research are challenging due to variations in datasets, forecasting methodologies, and market conditions. Our approach to hyperparameter optimisation, which utilises SMBO along with a multi-objective optimisation technique, is relatively experimental and appears to be constrained by computational limitations. For future research, a more rigorous validation of this methodology would be beneficial to better assess its effectiveness and potential limitations.

Future research could also explore more sophisticated ML models to enhance the accuracy of EPF. Integrating LSTM networks with 1-D convolutional networks could leverage the strengths of both techniques, potentially leading to better forecasts. Additionally, focusing solely on convolutional operations through Temporal Convolutional Networks might yield improvements. Another model worth investigating is the Temporal Fusion Transformer, which has shown promising results in various forecasting domains, but to my knowledge have not been extensively applied to EPF.

Improving the dataset by incorporating diverse and more effective features is also crucial. Specifically, features related to precipitation could be more informative than hydro filling levels for capturing hydrological impacts. Furthermore, considering the inclusion of British spot prices could be beneficial, particularly since including prices cleared before the predictive market has shown good results in other markets.

For future work, it would be intriguing to explore the application of SHAP di-

rectly to the LSTM model, given that this model more accurately reflects the complexities of the actual market dynamics. This exploration could be facilitated by employing SHAP's DeepExplainer, which is specifically designed for interpreting neural networks. However, there are inherent challenges associated with this approach, such as the absence of dependency plots, interaction values, and the necessity to average SHAP values over time steps.

The applications of ML models combined with XAI tools far surpass the applications in this thesis. Despite the methodological constraints, XAI is a crucial tool for interpreting model decisions. As we witness a significant transition in energy systems, not only within electricity markets but across other areas due to the shift towards renewable sources following the energy crisis, the role of XAI becomes increasingly valuable. Effective ML models, when accurately representing the systems they are designed to model, enable XAI to clarify and make transparent the underpinnings of complex phenomena. This is vital, whether it involves understanding the causes of grid instability, grid faults, electricity prices or other relevant areas.

References

[1] I. R. Gran, H. Taule, P. V. Hansen, L. Hagen, J. Hilland, M. H. Lie, K. Bjella, S. Sundsbø, N. Lillelien, and N. Hansen. Balansekunst. Regjeringen (2023). URL: https://www.regjeringen.no/contentassets/95acb3e7cbe54e62b82173d9885935ed/rapport-fra-stromprisutvalget-12.-oktober-2023-balansekunst.pdf.

[2] J. Trebbien, S. Pütz, B. Schäfer, H. S. Nygård, L. Rydin Gorjão, and D. Witthaut. Probabilistic Forecasting of Day-Ahead Electricity Prices and their Volatility with LSTMs. *2023 IEEE PES Innovative Smart Grid Technologies Europe (ISGT EUROPE)* 1–5, 2023. DOI: 10.1109/ISGTEUROPE56780.2023.10407112.

[3] T. T. Pedersen, E. K. Gøtske, A. Dvorak, G. B. Andresen, and M. Victoria. Long-term implications of reduced gas imports on the decarbonization of the European energy system. *Joule* **6**(7), 1566–1580, 2022. DOI: 10.1016/J.JOULE.2022.06.023.

[4] R. Weron. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting* **30**(4), 1030–1081, 2014. DOI: 10.1016/J.IJFORECAST.2014.08.008.

[5] J. Trebbien, L. Rydin Gorjão, A. Praktiknjo, B. Schäfer, and D. Witthaut. Understanding electricity prices beyond the merit order principle using explainable AI. *Energy and AI* **13**, 100250, 2023. DOI: 10.1016/J.EGYAI.2023.100250.

[6] T. Bye and E. Hope. Deregulation of Electricity Markets: The Norwegian Experience. *Economic and Political Weekly* **40**(50), 5269–5278, 2005. URL: http://www.jstor.org/stable/4417519.

[7] Norwegian Ministry of Energy (2023). The power market. https://energifaktanorge.no/en/norsk-energiforsyning/kraftmarkedet/. Accessed: 2024-30-04.

[8] A. Jdrzejewski, J. Lago, G. Marcjasz, and R. Weron. Electricity Price Forecasting: The Dawn of Machine Learning. *IEEE Power and Energy Magazine* **20**(3), 24–31, 2022. DOI: 10.1109/MPE.2022.3150809.

[9] H. Hewamalage, C. Bergmeir, and K. Bandara. Recurrent Neural Networks for Time Series Forecasting: Current status and future directions. *International Journal of Forecasting* **37**(1), 388–427, 2021. DOI: 10.1016/j.ijforecast.2020.06.008.

[10] R. Machlev, L. Heistrene, M. Perl, K. Y. Levy, J. Belikov, S. Mannor, and Y. Levron. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI* **9**, 100169, 2022. DOI: 10.1016/J.EGYAI.2022.100169.

[11] Wikipedia contributors. Electricity price area (2015). https://en.wikipedia.org/wiki/Electricity_price_area. Accessed: 2024-02-05.

[12] N. (2010). NASDAQ OMX Acquires Nord Pool ASA. https://www.nasdaq.com/about/press-center/nasdaq-omx-acquires-nord-pool-asa-0. Accessed: 2024-01-05.

[13] J. K. Kolberg and K. Waage. *Articial Intelligence and Nord Pool's Intraday Electricity Market Elbas: A Demonstration and Pragmatic Evaluation of Employing Deep Learning for Price Prediction*. Master's thesis. Norwegian School of Economics, 2018. http://hdl.handle.net/11250/2560898.

[14] T. Bye, M. Bjørndal, G. Doorman, G. Kjølle, and C. Riis. Flere og riktigere priser - Et mer effektivt kraftsystem. https://www.regjeringen.no/globalassets/upload/oed/rapporter/2010_1130_flere_og_riktigere_priser_et_mer_effektivt_kraftsystem.pdf?id=2200911. Accessed: 2024-01-05.

[15] Nord Pool. Price calculation. https://www.nordpoolgroup.com/en/trading/Day-ahead-trading/Price-calculation. Accessed: 2024-30-04. 2023.

[16] NEMO Committee. Single Day-ahead Coupling (2019). https://www.nemo-committee.eu/sdac. Accessed: 2024-30-04.

[17] NEMO Committee. EUPHEMIA Public Description (2020). https://www.nordpoolgroup.com/globalassets/download-center/single-day-ahead-coupling/euphemia-public-description.pdf. Accessed: 2024-05-04.

[18] Regjeringen. The electricity market (2008). https://www.regjeringen.no/globalassets/upload/oed/pdf_filer/faktaheftet/evfakta08/evfacts08_kap07_eng.pdf. Accessed: 2024-01-05.

[19] Eurostat. Renewable energy statistics (2023). https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Renewable_energy_statistics#Almost_one_quarter_of_energy_used_for_heating_and_cooling_from_renewable_sources. Accessed: 2024-04-30.

[20] Norwegian Ministry of Energy. The electricity grid (2024). https://energifaktanorge.no/en/norsk-energiforsyning/kraftnett. Accessed: 2024-30-04.

[21] L. Tucci. What is machine learning and how does it work. https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML. Accessed: 2024-01-05. 2023.

[22] S. Mc Loone and G. Irwin. Improving neural network training solutions using regularisation. *Neurocomputing* **37**(1), 71–90, 2001. DOI: 10.1016/S0925-2312(00)00314-3.

[23] D. Svozil, V. Kvasnicka, and J. Pospichal. Introduction to multi-layer feed-forward neural networks. *Chemometrics and Intelligent Laboratory Systems* **39**(1), 43–62, 1997. DOI: 10.1016/S0169-7439(97)00061-0.

[24] I. Goodfellow, Y. Bengio, and A. Courville. Deep Learning. URL: http://www.deeplearningbook.org, Accessed: 2024-05-04. MIT Press, 2016. ISBN: 978-0262035613.

[25] P. Velikovi. TikZ (2016). https://github.com/PetarV-/TikZ/tree/master. Accessed: 2024-01-05.

[26] M. A. Nielsen. Neural Networks and Deep Learning. Determination Press, 2015. http://neuralnetworksanddeeplearning.com/.

[27] F. Chollet. Deep Learning with Python. 1st ed. USA: Manning Publications Co., 2017. ISBN: 9781617294433.

[28] S. Raschka and V. Mirjalili. Python Machine Learning. 3rd ed. Birmingham, UK: Packt Publishing, 2019. ISBN: 978-1789955750.

[29] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation* **9**(8), 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735.

[30] R. Jozefowicz, W. Zaremba, and I. Sutskever. An empirical exploration of recurrent network architectures. *International Conference on Machine Learning* **37**, 2342–2350, 2015. URL: https://proceedings.mlr.press/v37/jozefowicz15.html.

[31] R. Mitchell and E. Frank. Accelerating the XGBoost algorithm using GPU computing. *PeerJ Computer Science* **3**, e127, 2017. DOI: 10.7717/peerj-cs.127.

[32] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* **30**, 4768–4777, 2017. DOI: 10.48550/arXiv.1705.07874.

[33] L. S. Shapley. 17. A Value for n-Person Games. *Contributions to the Theory of Games (AM-28), Volume II*. Princeton: Princeton University Press, 307–318, 1953. ISBN: 9781400881970. DOI: 10.1515/9781400881970-018.

[34] C. Molnar. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. 2nd ed., 2022. URL: https://christophm.github.io/interpretable-ml-book.

[35] S. M. Lundberg, G. G. Erion, and S. Lee. Consistent Individualized Feature Attribution for Tree Ensembles. *ArXiv preprint*, 2018. DOI: 10.48550/arXiv.1802.03888.

[36] ENTSO-E. ENTSO-E transparency platform. https://transparency.entsoe.eu/. Accessed: 2024-01-05.

[37] Investing.com. Dutch TTF Natural Gas Futures Historical Data. https://www.investing.com/commodities/dutch-ttf-gas-c1-futures-historical-data. Accessed: 2024-01-05.

[38] Federal Reserve Economic Data (FRED). Crude Oil Prices: Brent - Europe. https://fred.stlouisfed.org/series/DCOILBRENTEU. Accessed: 2024-01-05.

[39] R. K. Pearson, Y. Neuvo, J. Astola, and M. Gabbouj. The class of generalized Hampel filters. *2015 23rd European Signal Processing Conference (EUSIPCO)* 2501–2505, 2015. DOI: 10.1109/EUSIPCO.2015.7362835.

[40] X. Wan. Influence of feature scaling on convergence of gradient iterative algorithm. *Journal of Physics: Conference Series* **1213**, 032021, 2019. DOI: 10.1088/1742-6596/1213/3/032021.

[41] 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. DOI: 10.1016/j.apenergy.2021.116983.

[42] Z. Chang, Y. Zhang, and W. Chen. Electricity price prediction based on hybrid model of adam optimized LSTM neural network and wavelet transform. *Energy* **187**, 115804, 2019. DOI: 10.1016/j.energy.2019.07.134.

[43] P. Baldi and P. J. Sadowski. Understanding Dropout. *Advances in Neural Information Processing Systems* **26**. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/71f6278d140af599e06ad9bf1ba03cb0-Paper.pdf.

[44] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. *ArXiv preprint*, 2019. DOI: 10.48550/arXiv.1907.10902.

[45] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for Hyper-Parameter Optimization. *Advances in Neural Information Processing Systems* **24**. Curran Associates, Inc., 2011. URL: https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf.

[46] A. Low, E. Vissol-Gaudin, Y.-F. Lim, and K. Hippalgaonkar. Mapping pareto fronts for efficient multi-objective materials discovery. *Journal of Materials Informatics* **3**, 11, 2023. DOI: 10.20517/jmi.2023.02.

[47] S. Daulton, M. Balandat, and E. Bakshy. Differentiable Expected Hypervolume Improvement for Parallel Multi-Objective Bayesian Optimization. *ArXiv preprint*, 2020. DOI: 10.48550/arXiv.2006.05078.

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830, 2011. DOI: 10.48550/arXiv.1201.0490.

[49] B. Uniejewski and R. Weron. Efficient Forecasting of Electricity Spot Prices with Expert and LASSO Models. *Energies* **11**(8), 2039, 2018. DOI: 10.3390/en11082039.

[50] A. Wagner, E. Ramentol, F. Schirra, and H. Michaeli. Short-and long-term forecasting of electricity prices using embedding of calendar information in neural networks. *Journal of Commodity Markets* **28**, 100246, 2022. DOI: 10.1016/j.jcomm.2022.100246.

[51] L. Tschora, E. Pierre, M. Plantevit, and C. Robardet. Electricity price forecasting on the day-ahead market using machine learning. *Applied Energy* **313**, 118752, 2022. DOI: 10.1016/J.APENERGY.2022.118752.

[52] W. Demloj. *Electricity Price Forecasting using Multivariate Price Time Series*. Master's thesis. Oslo Metropolitan University, 2023. https://hdl.handle.net/11250/3100586.

[53] T. A. Løfgren and H. Ingstad. *Norwegian Hydropower Producers' Response to the 2021 Energy Price Shock: An Analysis of the Development in the Water*

Values. Master's thesis. Norwegian University of Life Sciences, 2023. https://hdl.handle.net/11250/3078976.

[54] G. Van Houdt, C. Mosquera, and G. Nápoles. A review on the long short-term memory model. *Artificial Intelligence Review* **53**(8), 5929–5955, 2020. DOI: 10.1007/s10462-020-09838-1.

[55] T. Heskes, E. Sijben, I. G. Bucur, and T. Claassen. Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models. *ArXiv preprint*, 2020. DOI: 10.48550/arXiv.2011.01625.

A Appendix

Table A.1: Hyperparameter optimisation search space LSTM model.

Hyperparameter	**Type**	**Options**
Number of LSTM Layers	Integer	1 – 3
Nodes per Layer	Categorical	16, 32, 64, 128, 256
Dropout Rate	Categorical	0.0, 0.1, 0.2, 0.3, 0.4, 0.5
Batch Size	Categorical	16, 32, 64, 128, 256
Sequence Length	Categorical	24, 48, 96, 168

Table A.2: Hyperparameter search space XGBoost model.

Hyperparameter	**Type**	**Options**
Learning rate	Float, log	10^{-3} – 0.1
Max depth	Integer	1 – 10
Subsample	Float	0.05 – 1.0
Colsample bytree	Float	0.05 – 1.0
Min child weight	Integer	1 – 20

Table A.3: Hyperparameter search space DNN model

Hyperparameter	**Type**	**Options**
Nodes per Layer	Categorical	16, 32, 64, 128, 256
Dropout Rate	Categorical	0.0, 0.1, 0.2, 0.3, 0.4, 0.5
Batch Size	Categorical	16, 32, 64, 128, 256
Sequence Length	Categorical	24, 48, 96, 168

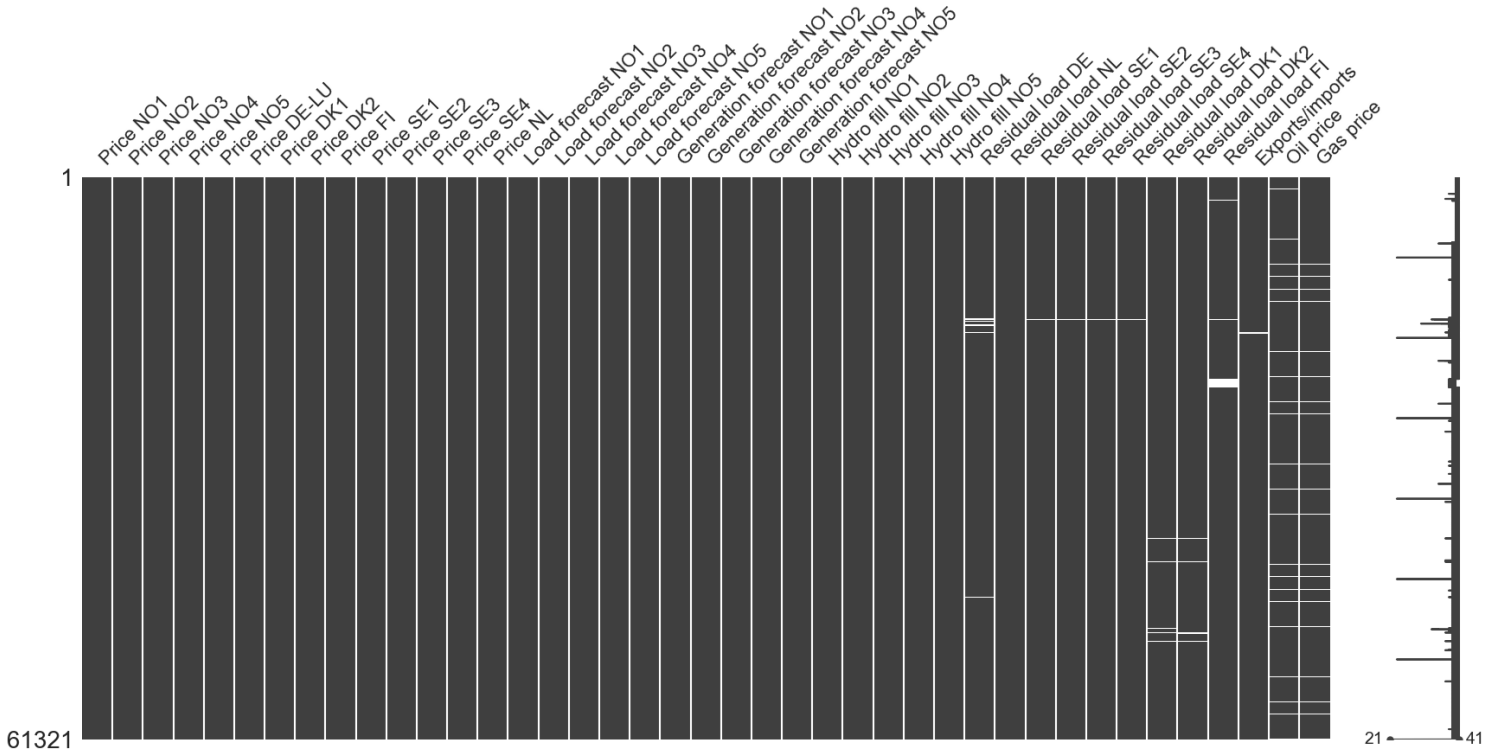


Figure A.1: Visualisation of missing values in the dataset. For composite features, i.e. features that are combinations of other features, missing values have been consolidated for a clearer visual representation. This approach is also applied to features with resolutions different from the predictive feature.

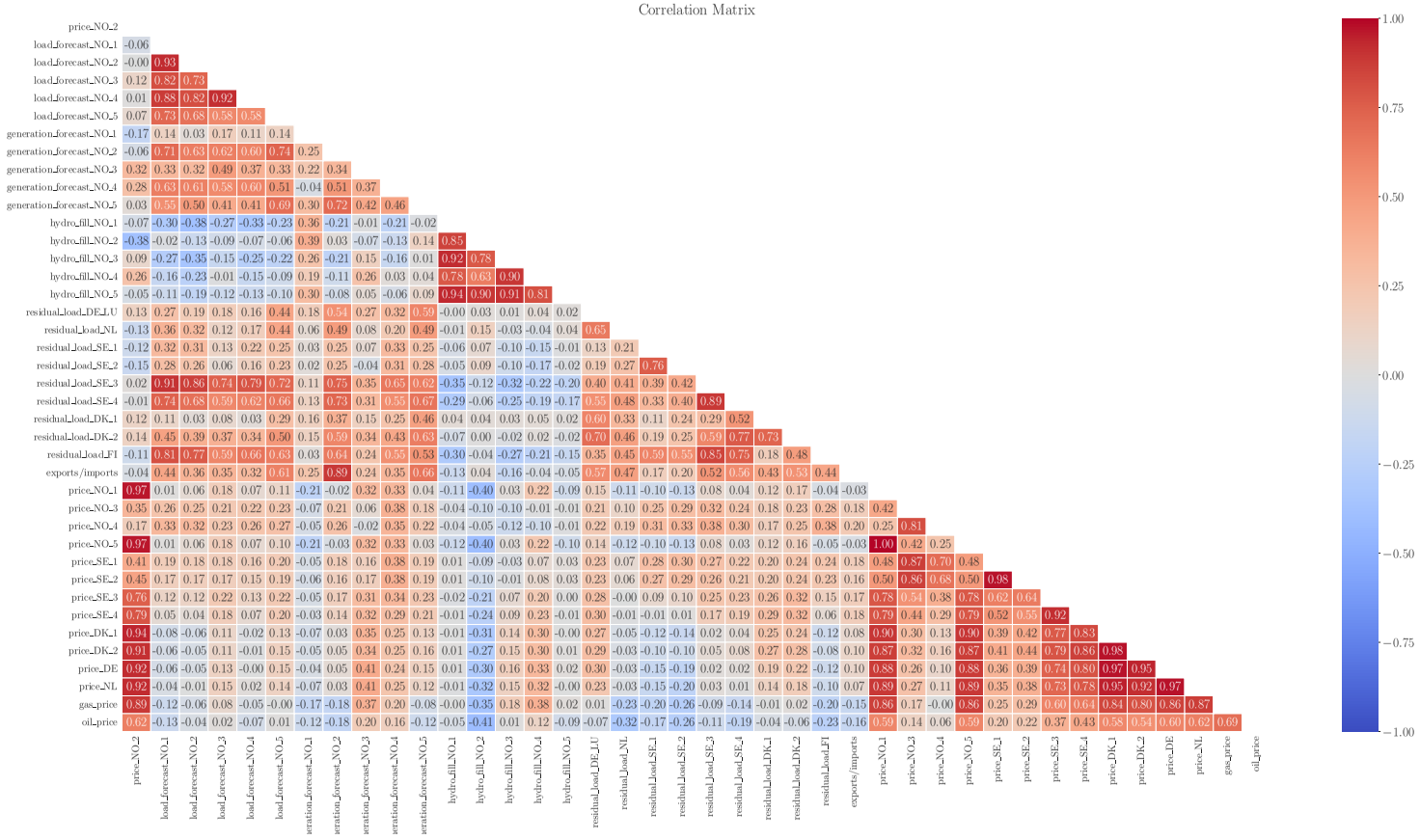
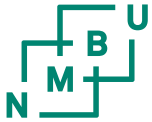


Figure A.2: Correlation heatmap displaying Pearson correlation coefficients for features from 2017 to 2023.



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskaplege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway