Norwegian School of Economics

Bergen, Spring 2022

Buy on Intraday Market or not: A Deep Learning Approach

A decision tool for buyers in the Norwegian electricity markets to decide optimal market to purchase electricity

Sondre Eide & Olai Viken

Supervisor: Jonas Andersson

Master thesis, Economics and Business Administration,
Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Abstract

As the share of variable renewable energy sources increases, so does the need for near-delivery offloading of surplus electricity. The availability of potentially cheap energy sources in intraday markets begs warrants the reconsideration of a potentially overlooked market. From a power buying perspective, this thesis has applied promising deep neural network techniques to produce accurate electricity price forecasts before day-ahead market closure. Architectures tested in this thesis include *long short-term memory (LSTM)*, *gated recurrent units (GRU)*, *deep autoregressive models (DeepAR)* and *temporal fusion transformers (TFT)*. Using *nested cross-validation scheme*, we seek to better approximate the generalization error of our models. LSTM and GRU models are found to be the best performing, in day-ahead and intraday markets, beating the benchmark measured in MAE by 30.6 % and 29 %, respectively. The increase in performance achieved by deep neural architectures are found to be particularly prominent in periods of high price volatility.

Our overall goal has been the creation of decision tool, to be used by an electricity buyer to determine optimal electricity market for a given set of delivery hours. The results presented in this thesis are based on the NO2 power region (South Norway) as a result of its relative intraday liquidity. We implement the decision tool by means of a a probabilistic classifier trained specifically on the forecasts of the optimal deep neural architectures. We find that the use of a probabilistic classifier increase classification performance when compared to using sign-difference of the forecasts directly.

Despite numerous potential error sources, our decision tool is shown to increase expected marginal profits when compared to a day-ahead-only trading strategy by testing in a out-of-sample simulated "production" environment. We model a decision tool to fit the needs of various risk profiles, and find that higher risk tolerance warrants higher profits. Though beyond the scope of this thesis, the general outline of this decision tool can be modified and extended to fit the needs of power producers.

Keywords: *Deep Neural Networks, Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Temporal Fusion Transformer (TFT), Autoregressive Recurrent Networks (DeepAR), Probalistic classification, Energy Quantified, Nord Pool, NO2 (South Norway), Intraday market, Elbas, Day-ahead market, Elspot, Regulating market, Electricity price forecasting, Nested Cross-Validation*

Acknowledgements

We would like to thank our supervisor Professor Jonas Andersson at NHH for guidance and good discussions about potential avenues of improving our thesis. We would like to thank Nord Pool for providing historical historical market data. We thank Hugo Birkelund the CEO of Energy Quantified who has helped us greatly by giving access to their rich archive of meteorological, capacity and production data. We attribute the results presented in this thesis especially to the high-quality data provided to us by Energy Quantified. We thank our near and dear ones Amalie Myrmel Gjengedal and Silje Eikrem Slettebakk for love and support.

Contents

Figures

Tables

1 Introduction

Nord Pool was one of the the first electricity exchanges to be opened after a wave of deregulation in the Nordic countries (Bye and Hope, 2005). Shortly after its debut in 1991, Nord Pool expanded its offering by allowing for continuous trading. Initially conceived as a supplement for the *balancing markets*, *intraday markets*[1] have seen somewhat of a reconsidering as *intermittent energy sources* are increasing its share of total electricity production in the Nordic countries. An increasing share of wind power in particular, is likely to lead to an increased dependence on intraday trading Mauritzen (2015). Perez-Arriaga and Batlle (2012) define intermittency as a component of non-controllable variability and partial unpredictability.

Price determination hours ahead of delivery poses problems, when the availability of power fluctuates. Price discovery in *day-ahead markets*[2] assumes that no fundamental price-drivers have changed in the period between price settlement and delivery. As the share of intermittent power sources increases, the validity of this assumption may diminish. This is particularly relevant for *South Norway's (NO2) power region* (see figure 1), where the Cross-Skagerrak subsea power line significantly increases inter-dependency of energy sources between Denmark and Norway by adding 1,700 MWh worth of capacity ("The Skagerrak 4-interconnector - cable contracts signed", 2013). Intermittent energy producers will be increasingly faced with the decision to trade on intraday markets or to face balancing costs for any deviation from the promised production. From a power buyer's perspective, this development enables wholesale electricity buy available wind power not priced in the day-ahead market. Participating in intraday trading will however involve inherent risks. Balancing costs may incur for any market participant that are not able to meet their predetermined buying or selling volume.

Recent advances in forecasting techniques may offset this risk by allowing for more accurate forecasts. Incorporating multitude of factors affecting electricity prices can be difficult using traditional econometric forecasting techniques. *Deep neural networks* have been successfully applied to time-series, often as *convolutional neural networks (CNN)* or *recurrent neural networks (RNN).* Iterations aimed to improve upon issues with RNN, such as the *long short-term models (LSTM)* originally proposed by J. Hochreiter (1991) have been shown outperform traditional such as *autoregressive integrated moving average (ARIMA) forecasting models* (Lago et al., 2018). More accurate forecasts may allow buyers to compare *electricity prices* on day-ahead and intraday at a critical closing decision point. Day-ahead markets required market partici-

[1] Commonly known as Elbas market

[2] Commonly known as Elspot market

pants to submit their bids at noon before the day of delivery (Mauritzen, 2015). Ahead of the bid-submission deadline, it may be feasible for electricity buyers to decide, for each and every upcoming delivery hours the next day, which market to participate on. In other words, bidding on the day-ahead market or wait until next day and buy on the intraday market, or perform a combined trading strategy.

In this thesis we will explore this possibility by training and validating a range of neural network architecture to forecast day-ahead and intraday electricity prices. We opt to take a buying perspective of the market participants on the Nord Pool markets, which determines particular choices in the creation of the decision tool. We find that our best performing neural networks outperform benchmarks in terms of *mean absolute error (MAE)* by 30.6 % and 29 % in day-ahead and intraday markets, respectively. Applying a probabilistic classifier trained on day-ahead and intraday forecasts, improves forecasts modestly. The application of the decision tool in a simulated use-case is found to increase expected profits when compared to buying electricity exclusively in day-ahead markets.

1.1 Problem Definition

A central prerequisite of determining the optimal electricity market for an electricity buyer is well-performing forecasts. We follow a *k-fold cross-validation* scheme as described in Hastie et al. (2008). To ensure more efficient use of data, while having adequate time dispersed samples, we perform a rolling cross validation within in each fold. The end result is a *nested k-fold cross-validation* technique, similar to previous work by Varma and Simon (2006). Simpler validation techniques such as *holdout-validation* may not give an accurate estimate of a model's generalization error. Our approach uses a probabilistic classification model to rectify forecast errors of the implemented deep learning models. The end result, simply referred to as the *decision tool*, may be used by electricity buyer to optimally allocate buy volume for the delivery hours of the following day. The decision tool itself is transferable to power producers, but it is beyond the scope of this thesis to demonstrate the efficacy of its application from a producers perspective. Throughout this thesis we seek to answer the following research question:

RQ *Through the use of deep learning forecasts, can an optimal market for buying electricity be determined ahead of the day-ahead market submission deadline?*

We will place a particular emphasis on using variables, machine learning techniques, performance validation, and forecasting horizons that would be available to a potential electricity buyer.

2 Literature review

Electricity market forecasts are usually reserved to predict either day-ahead or intraday markets. Day-ahead market forecasts garners the most attention probably due its dominant share of total trade. Examples of forecasting day-ahead prices using neural networks include Lago et al. (2018) and Sprangers et al. (2022). Intraday forecasts are mostly reserved to the immediate hours before delivery, as seen in Narajewski and Ziel (2020) and in Kolberg and Waage (2018). These forecasts may be useful for intermittent power producers determining whether surplus power should be left to balancing market or sold on the intraday markets. Power producers incur balancing costs when power promised for any delivery hour deviates from power produced. The *Transmission System Operator (TSO)*[3] keeps reserves available, levying the increased balance costs on any power producers failing to comply with promised supply. Holttinen (2005) estimates that wind power producers could reduce their balancing costs significantly by participating in intraday trading rather than selling on the day-ahead market. Similar results are found by Bourry and Kariniotakis (2009), proposing a combined approach of participating in both day-ahead and intraday markets to lower balancing costs.

Faria and Fleten (2011) take the perspective of a price-taking hydro-producer when considering the value of participating in intraday trade. They find that given the lack of liquidity and uncertainty when simulating prices in intraday markets, taking intraday prices into account, does not make a meaningful impact on profits. It should be noted that the decision to trade in intraday markets is made in a two-stage stochastic model, where intraday trade is made after deciding the optimal sell volume in the day-ahead markets. A central point of interest for our thesis is whether more accurate forecasting tools could provide foresight into at a critical decision point of any discrepancy between intraday and day-ahead prices before day-ahead price settlement. This entails building forecasting tools solely with the data that would be available at the moment of decision. Maciejowska et al. (2019) forecasts intraday and day-ahead prices to classify the sign of the price difference between intraday and day-ahead markets in Poland and Germany. Employing econometric techniques they achieve a 57.3 % accuracy on classifying day-ahead and intraday price differences in the Polish electricity markets. The top performing *autoregressive model (ARX)*, as seen in Maciejowska et al. (2019) utilises exogenous market variables, but were not able to efficiently use long series of data.

The extent to which the price difference can be accurately classified depends on the accuracy of

[3] Statnett (Norway), EnergiNet (Denmark) and Svenska kraftnät (Sweden) are examples of Transmission System Operators

the forecasts. Nord Pool opens for trade at noon, and any forecasts has to be ready for relevant decision makers to act before the day-ahead market submission deadline. By the time of delivery multiple factors affecting the price may have influenced the price to the point where the forecast no longer is sufficiently accurate.

Neural network based forecasting architectures have been successfully applied to electricity price forecasting. As forecasting models they have been demonstrated to be flexible and able to model non-linear characteristics (X. Chen et al., 2012). *Traditional multilayered perceptrons (MLP)* have been refined to capture sequential data as the families of RNN and CNN are examples of. Advances in RNN, such as LSTM networks originally proposed by S. Hochreiter and Schmidhuber (1997), have been shown to outperform econometric forecasting techniques. Lago et al. (2021) presents several *deep*[4] *neural architectures* for forecasting electricity prices in Belgian, French, German and Nordic markets. They experiment with adding additional linear layers atop of LSTM, RNN and CNN architectures. They find that deep neural networks outperformed their benchmark model *lasso estimated autoregressive (LEAR).*

Gated recurrent unit (GRU) networks originally proposed by Cho et al. (2014), which features prominently in this thesis, simplify the forgetful properties of the LSTM networks. GRU models are found to have a comparable performance to that of LSTM models, at a computational discount (Chung et al., 2014). Forecasting electricity load, Wu et al. (2019) show that GRU networks perform better and have lower computational complexity than LSTM networks.

Considering the aforementioned advances in forecasting techniques, a key goal of this thesis is to research the possibility of using accurate forecasting methods to evaluate intraday and day-ahead markets before market submission deadline. We aim to add to existing literature by doing for Nord Pool markets with neural network forecasting techniques, what Maciejowska et al. (2019) did for Polish and German markets with econometric techniques.

[4] In the context of neural networks, deep neural networks usually refer to models with at least one hidden layer

3 The Power Market

Nord Pool is a European power market primarily owned by Euronext and TSO Holding and offers trading in both day-ahead and intraday markets across 16 European countries ("About us", n.d.). 360 companies in 20 countries actively trade on Nord Pool, and Nord Pool is committed to allowing for all kinds of traders regardless of company size and location. Since Nord Pool is the counterparty for all participants their key role is to handle payments and ensure delivery. The magnitude of the total trade volume across its markets is significant, during 2021 the total buy and sell volume exceeded 450 Terra Watt hours (TWh) each ("Nord Pool Announces 2021 Trading Figures", 2022).

The pretext of Nord Pool's establishment was the deregulation of governmental ownership in the 1990s in order to create an efficient market across countries. As a result, the price for a given trade hour depends on the market balance between supply and demand which contributes to preventing monopolistic prices ("The power market", n.d.). Additionally, this social-economic price model does not only serve the purpose of serving competitive prices but eases the identification of production or capacity shortcoming by for instance observing higher demand than power production supply.

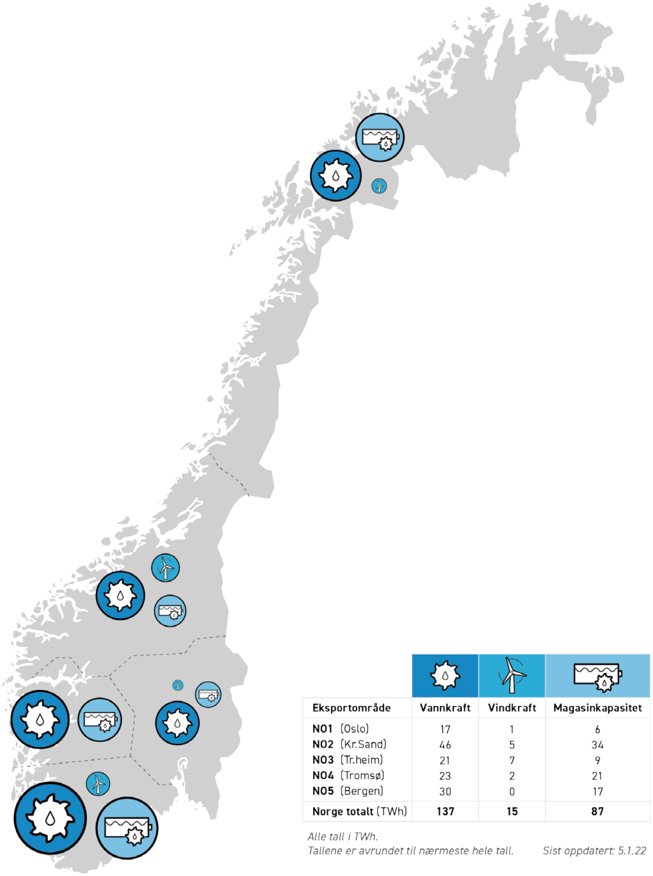
3.1 The Scandinavian Power Market

The Scandinavian power market comprises many power-intensive industries, delivering a mixture of e.g. hydro, nuclear and wind power ("An overview of the Nordic Electricity Market", 2019). Power consumption changes as a consequence of the seasonal weather conditions e.g. change in temperature (see section 4.2). Due to large seasonal variations in weather conditions, Nordic household stands for a large share of electricity heated houses which distinguishes from the rest of European power consumption. In addition, the Scandinavian power market holds a large share of *renewable power generating sources* especially hydro, but wind power production is also increasing. In a span of 10 years wind power's share of total produced electricity in Denmark has increased from 13 % in 2006 to 42 % in 2016 (Unger et al., 2018). Which in turn affects the imported volume and power price in i.e. Norway, since wind power cannot be stored on large scale. Due to its intermittent property (see definition in section 4.1), an overproduction of power for windy situations will decrease the power prices in Denmark, as a consequence of high supply and low demand. Additionally, since wind power heavily relies on the weather conditions, and weather is hard to forecast, it will cause deviation in the planned capacity available on the day-ahead market, which is partly why the intraday market exist. The same affect occurs for run-of-river power production when the precipitation is high, which is among the important power sources in Norway. The intermittent power generating sources are considered as important price-drivers on predicting the intraday and day-ahead prices, which also the literature reveals in section 4.1.

3.2 The Norwegian Power Market

The Norwegian power market is separated into five different areas: *NO1 (South East Norway), NO2 (South Norway), NO3 (Middle Norway), NO4 (North Norway)* and *NO5 (West Norway)* (see figure 2). Figure 1 made by Norwegian Ministry of Petroleum and Energy shows the different mixture of power delivery where 93 % of the renewable power production in Norway, given a normal production year, are given by hydro power production and 7 % by wind power production ("Kraftmarkedet", n.d.). In general, the hydro production sources consist mainly of hydro reservoir and run-of-river production. Hydro reservoirs regulate power production in order to meet demand, while run-of-river and wind power production are variable sources suffering intermittency as mentioned in literature review. Both run-of-river and wind power generation shares the disadvantage of heavily relying on the weather conditions, i.e. precipitation or wind, as these sources generates electricity independent of demand. In contrast, the storage reservoirs can produce electricity in periods with inadequate precipitation and inflow. Considering the important power generating sources give an indication on what type of power source that may drive the day-ahead and intraday power prices we want to forecast.

The Norwegian Power Market



Eksportområde	Vannkraft	Vindkraft	Magasinkapasitet
NO1 (Oslo)	17	1	6
NO2 (Kr.Sand)	46	5	34
NO3 (Tr.heim)	21	7	9
NO4 (Tromsø)	23	2	21
NO5 (Bergen)	30	0	17
Norge totalt (TWh)	**137**	**15**	**87**

Alle tall i TWh.
Tallene er avrundet til nærmeste hele tall. *Sist oppdatert: 5.1.22*

Figure 1: The figure shows the distribution of renewable power sources in Norway for each power areas given normal production year. All the number are in Terra Watt hours rounded to closest whole number. Ekspertområde = Power Market Area in Norway, Vannkraft = Hydro Power (e.g. run-of-river), Vindkraft = Wind Power, Magasinkapasitet = Hydro Reservoir Capcity Power. Reprinted from THE POWER MARKET, in Energi fakta Norge, n.d., retrieved May 27, 2022, from https://energifaktanorge.no/norsk-energiforsyning/kraftmarkedet/

3.2.1 Transmission connections

Norway's renewable power production often leads to power surplus, making it possible to export large amounts of power to foreign countries ("Mest kraftutveksling med Norden", 2021). At the time of writing, Norway has transmission connections to Sweden, Denmark, Germany, Netherlands and Great Britain. The most recent interconnection opened in October 2021, named the North Sea Link, connecting Great Britain with Norway and was initially transmitting half of its capacity and gradually increased to full capacity by March 2022 ("UK and Norway", n.d.). In addition, Sweden and Denmark account for most of the power exportation share in Norway. Norway purchases power when the power price in the foreign country is lower than the internal power prices ("The power market", n.d.). An overview of the different transmission connections in-between Norwegian areas and to other countries from Norway, is shown in figure 2. The main reason for the partition of area in Norway is due to the transmission capacity constraints ("Bidding areas", n.d.). Additionally, the capacity can differ given the direction of exchange.

Transmission Connections

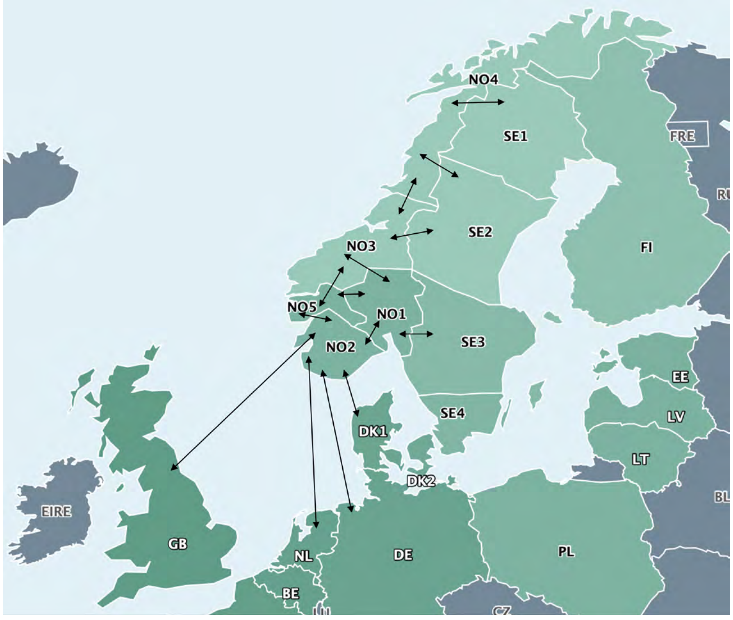


Figure 2: The figure shows the internal and external transmission connections in Norway. Reprinted and adapted from Nord Pool, Nord Pool, n.d., retrieved February 1, 2022, from https://www.nordpoolgroup.com/en/Market-data1//nordic/map. Reprinted with permission.

3.3 Trade markets in the Norwegian areas

3.3.1 Day-ahead markets

The day-ahead market provides the opportunity to trade hourly power one day in advance ("The power market", n.d.). It is also the most liquid market because the majority of trades occur in this market (see section 5.5). Market participants can start their bidding and provide offers between 8 a.m. and 12 p.m. Ahead of 10 a.m. the TSO publishes the available capacities in each bidding area (more about TSO in section 3.3.3). The decision tool presented in this thesis, has to provide forecast of day-ahead prices between 10 a.m. and 12 p.m. such that the end-user have time to decide a bid placement, or not. This is due to the fact that when the auction closes at 12 p.m. the market prices will be calculated for the next day (Mauritzen, 2015). The day-ahead prices are calculated as a function of buy and sell orders, involving price and volume, and the available transmission capacity. Nord Pool uses a common European algorithm to calculate the prices for the different areas. The market prices in different areas are revealed at 2 p.m. and showing if the trade for the market participant did pass through or not. The period between 12 p.m. and 2 p.m. is commonly called clearance period. This decision time-window constraints the rest of this thesis where also the forecast prices of the intraday market is performed in the same period.

3.3.2 Intraday markets

Nord Pool offers continuous intraday trading, opening two hours after price settlement is completed on the day-ahead market (Tangerås and Mauritzen, 2014). The market exists as a consequence that weather conditions are unreliable such that the market balance in the day-ahead market may be interrupted, because of change in actual production or consumption ("The power market", n.d.). The intraday market may compensate for this imbalance providing continuous trades between the periods of market clearance and one hour before operation. A more exhaustive study on intraday market is provided in section 5.5, which shows that the prices in, i.e. NO2, is lower on the intraday market relative to the day-ahead market on median. It is then interesting to consider this market in order to exploit cost reducing opportunity, i.e. buy in the market where the price is lowest.

3.3.3 Balancing markets

In order to understand the entirety of the energy markets there are, in addition to day-ahead and intraday, *market balances* between demand and supply. Balancing markets coexisting to resolve sudden events that can disturb the aforementioned market balance ("The power market", n.d.). Households, service industries, especially involving human life and health, and manufacturing industries heavily rely on continuous power supply, and any imbalances can cause fatal consequences ("Security of electricity supply", n.d.). Service interruptions may occur such as faults in power lines, substations and control systems, but the most common cause of failure is weather-related incidents. Operational security is then crucial to secure end-users power consumption with reliable power supply and the Transmission System Operator functions as a regulator to ensure balance in the power markets ("The power market", n.d.). In Norway, Statnett is assigned this task and generally uses the flexible hydro power plants to maintain instantaneous market balance. In more details, the TSO has to ensure a system frequency of between 49.9 and 50.1 Hz. If deviation from this frequency occurs, depending on how long the imbalance lasts, the TSO can act by using different remedy. See section A.1.1 in the appendix for more details on the different reserves.

3.4 Merit order

Figure 3 shows the merit order curve on the Nord Pool power exchanges. The merit order curve illustrates that different power sources supply electricity depending on demand in the market and available capacities. At the left most of figure 3, beige and light grey, the renewable energy produces a large share of power, to a relative low marginal cost. Due to the intermittency of wind production, and the high variance in operating capacity, has a wide implication on Nord Pool exchanges as these capacity sizes increases (reduces), and due to its low marginal cost drive the power prices down (up). On the other hand, as the demand for power increases the capacity for each power sources will reach their production limit, resulting in other dispatchable power sources to step in to meet the demand. Thermal power plant uses fuel to produce energy, and as we will see in section 4.3, the carbon price shown in dark grey in figure 3, drives the marginal cost of these dispatchable power sources further up. In short, the power is produced where the marginal cost is lowest, and increases as a function of demand and available capacities.

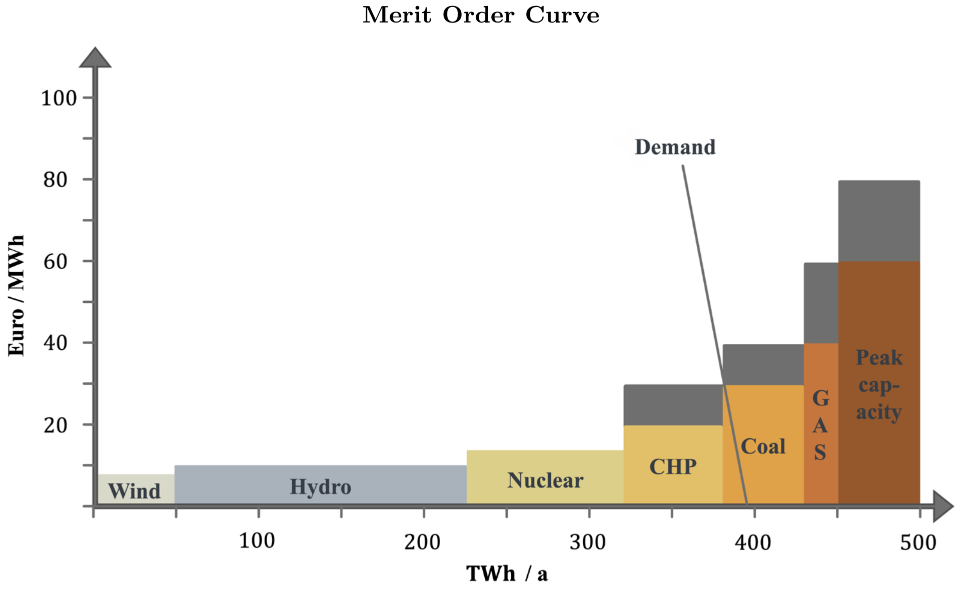


Figure 3: Merit order curve of Nord Pool Exchanges 2009, where CHP is short for *combined heating and power production*. Adapted from “System and market integration of wind power in denmark”, by Pöyry from Lund et al., 2017, *Energy Strategy Reviews*, 1 (3), p. 143-156. Copyright 2012 by the Elsevier Ltd.

4 Price-Drivers

We opt to use literature as a guide to find relevant variables that can explain the day-ahead and intraday prices, but well aware that there is a possibility that historical findings may not yield similar relationships in recent period of time. In addition, we opt to include temporal features and binary structural change features.

4.1 Intermittent Renewable Power Generation

The intermittent energy sources have over the years increased its share of energy generating sources in the Nordic electricity markets (see section 3) and are considered important price-drivers both on intraday and day-ahead markets. As mentioned in the introduction, Perez-Arriaga and Batlle (2012) defines intermittency as a component of non-controllable variability and partial unpredictability. Non-controllable variability meaning that renewable power plants are unavailable, due to low solar intensity, especially at night, for solar photovoltaic power generators or insufficient wind in wind production, when there is demand for it or providing significant amount of power production when demand is low (Kyritsis et al., 2017). The latter, partial unpredictability is referring to the limitation of knowing future power production of renewable energy due to the fact that the production depends on the stochastic nature of weather conditions.

The importance of intermittent renewable power generating sources on intraday and day-ahead electricity prices are also revealed in the literature as there exist empirical evidence that renewable energy production, such as wind and solar photovoltaic production, are significant and have an price-reducing effect on the day-ahead power prices when the production increases, especially in the Spanish, Austrian, German and Italian power market (Gelabert et al., 2011; Würzburg et al., 2013; Cludius et al., 2014; Ketterer, 2014; Clo et al., 2015). As intraday and day-ahead markets are highly correlated, with a correlation coefficient of 94.26 % when calculating it for price area NO2 in the period of 2019 and 2022, it is a reasonable assumption that price-drivers in the day-ahead market have a significant explanatory power on intraday markets.

Only considering the price-effect of increased renewable generation may not explain volatility in day-ahead and intraday prices. In the paper of Kyritsis et al. (2017), on intermittent solar and wind power effect on electricity prices in Germany, finds that solar power generation reduces the volatility and the probability of day-ahead electricity price spikes. The opposite effect with wind power generation where the volatility increases and introduces price spikes on day-ahead electricity price. Similar findings when Rintamäki et al. (2017) studied both the danish and

German power market. These effects occurs during all hours, peak hours and off-peak hours (Kyritsis et al., 2017).

Another renewable power generating source is hydroelectric plants. Since Norway is the largest hydro power producer in Europe and the Norwegian power market is highly dependent on hydroelectric generation, it is considered as an important price-driver in the power market ("Energy and marine resources", n.d.; Mosquera-López et al., 2018). Geissmann and Obrist (2018) confirms this in their findings that hydro power generation, i.e. storage and pump-storage generation, has even larger impact on power prices relative to other renewable sources. They argue that it is due to water reservoirs availability, that supply and demand intersects each other in the steeper region of the merit order curve, which yields higher potential price reductions.

4.2 Weather Conditions

Intermittent renewable energy generation are highly dependent on weather conditions, but each power source is affected in different ways. Mosquera-López et al. (2018) finds negative relation between temperature and power prices when freezing event occurs, where freezing event is defined as temperature below zero degrees Celsius. This is due to the fact that when water freezes hydro power generations are stopped, triggering other power production sources such as thermal plants to turn on to fulfill demand requirements. Additionally, hydro power generation does not only depend on temperature, precipitations effects the water reservoir level as well when the temperature is not cold. Water reservoir might be less dependent in rapid changes in weather conditions, because it needs constant precipitations for a long period in order to increase the water-filling level.

Solar photovoltaic power depends on cloudiness, temperature and possible precipitation (Russo et al., 2022). Russo et al. (2022) explains, in the context of future climate change, that increasing average temperature might reduce the solar panels efficiency and increased cloud coverage ambient humidity. They also writes that clouds are one of the hardest meteorological features to simulate, indicating that it might be a unreliable feature to include in order to forecast power prices.

Wind speed affects the wind power production, where extreme weather events and high variability in wind speed, can disrupt power production (Russo et al., 2022). On the other hand, insignificant amount of wind speed when demand is high will likely lead to increased use of hydro power generation, according to the merit order curve (see section 3.4).

Besides, electricity demand is also affected by temperature (Mosquera-López et al., 2018). Temperature is an important variable because of the Nordic climate, where low temperature leads to increased power consumption in order to e.g. heat houses. In a study of factors affecting electricity demand in Athens and London by Psiloglou et al. (2009) finds that both cities peak in demand at winter due to low temperature and peak in demand at summer only for Athens due to hot temperature and increased use of air-conditioning.

4.3 Dispatchable Power Generation

We see that renewable generation sources are highly dependent on weather conditions, which means that in situations with production shortcoming or outages, e.g. freezing water due to temperature below zero degrees Celsius, windless conditions or low solar intensity, there is a need to use other power sources that has more reliable supply on-demand, i.e. dispatchable power generating sources. Solar, wind or hydro production are unlikely to produce at full capacity due their dependence on natural weather factors (Geissmann and Obrist, 2018). As expected, and which is also is written in the paper by Mosquera-López et al. (2018), when hydroelectric power plants stop generating power thermal power plants must be turned on, with its higher marginal cost observed in the merit order graph (see section 3.4), yields an increase in the price of electricity. This applies also to other intermittent generating sources such as wind and solar power production.

In contrast to renewable power generation, which are dependent on weather conditions and highly independent from the electricity demand, dispatchable power generating sources uses fuel in order to produce electricity, which are nuclear, coal or natural gas (“Understanding The Term 'Dispatchable' Regarding Electricity Generation”, 2021). The main advantage of using these power sources are the reliability, as fuel is constantly supplied.

In a study of fundamental price-driver on continental European day-ahead power market Geissmann and Obrist (2018) find that gas prices have a significant impact on power prices where an increase in gas prices lower the power prices. ACER (2021) presented a insights of the recent high energy prices, which they state to be mainly driven by the global gas price surge (“High Energy Prices”, 2021). They also writes that the high gas prices are driven by tight supply and high demand from North-East Asian and South America (liquefied natural gas), leaving less gas available for Europe. In contrast, Rintamäki et al. (2017) finds that natural gas prices are insignificant on the danish daily price volatility, stating that the reason behind this is due to small daily changes on gas spot prices which is unlikely to affect short-term bidding behaviours.

Additionally, Gelabert et al. (2011) argues that the insignificance of gas prices is caused by the fact that producers secures the gas supply through long-term contracts, and this isolates them from the variation in gas prices to a great extent.

Other factors that contributes to increased power prices in Europe are the increased coal and carbon prices ("High Energy Prices", 2021). Carbon prices was instituted in Europe as a consequence of Kyoto Protocol in 2005 in order to reduce greenhouse gas emission ("What is the Kyoto Protocol?", n.d.). The increase in coal and carbon prices are mainly driven by economic recovery, yielding higher demand, and change in weather condition pattern, such as cold winter and unusually hot summer. Geissmann and Obrist (2018) finds that carbon price had a insignificant marginal effect on day-ahead electricity prices. Same argument by Gelabert et al. (2011) about long-term contracts for gas supply applies to power plants using coal resource as well.

4.4 Custom features

We opt to include basic *temporal information features*, i.e. hours, day, week and month as addition features. This can be seen i table 11 in the appendix. The purpose of using such date or time components is to potentially find which time or seasons are important for the dependent variables of day-ahead and intraday price forecast. Besides, the date time in the selected time series itself uses different time zones. See section B.1.2 in the appendix for further details.

Binary structural change features may be relevant in order to capture sudden changes or price dynamic change of behaviour, in case that other covariates may fail to explain. As these variables are already known, we opt to given occurrence such as Covid-19 pandemic an activation between an approximate time interval, "Yes" in period and nothing else. The important transmission connection between Norway and Great Britain, and between Norway and Germany will have implications for exchanges with NO2, and we opt to include these as well. Nord Link and North Sea Link was active from March 31, 2021 and October 1, 2021, respectively ("NordLink", n.d.;"Trial operation at NSL starts on 1 October", 2021).

5 Data

The purpose with this section is to inform reader how the data is used, forecasting models are gathered, which pre-processing is needed such that the inputted data have an acceptable quality with the purpose of outputting decent results. Then use the insights from data exploration to explain why this thesis has scoped the decision tool to price area NO2 in Norway. As neural networks requires great amount of data records to perform decently, we intend to gather as much data as possible. Nord Pool has provided data as far back as 1999 and Energy Quantified provided us a student license with 3 years data where the start period is March 28, 2019. Because of the limited period of data from Energy Quantified, and due to important forecasts, synthetics and historical data, we decided to restrict our data period from March 28, 2019 until March 17, 2022.

5.1 Nord Pool

The Nord Pool data is consolidated from Nord Pool's File Transfer Protocol server (FTP-server) comprising historical price, volume and capacity data[5]. Price and volume/capacity are measured in Euro and megawatt hour (MWh), respectively. The day-ahead data is structured hourly and provides information about market prices, total power trade volume and transmission capacity for each area in Scandinavia, Balticum, Germany and Great Britain. As intraday trading involves continuous bidding on price and volume within specific hours or several hours (i.e. block orders) the data is represented as ongoing transaction data. Both day-ahead and intraday prices will be used as dependent variables in the different forecasting approaches, due to our architecture of predicting each market and afterwards make probabilistic classification of day-ahead exceeding intraday prices.

Winter and summer time cause a special feature in the data by increasing an extra hour record, or reducing an hour record, which Nord Pool has handled by including an extra hour on the 3rd hour ante meridian, i.e. hour 3a and 3b. In order to overcome this extra hour feature, as it only occurs twice a year, we opt to average both the price and volume data for these particular hours.

The resolution of our data should match the lowest common resolution, e.g. 15 minutes, hour or day et cetera. As day-ahead prices are submitted in hourly resolution, we opt to increase the ticker data resolution from intraday trades into hourly resolution. As the intraday data is

[5] The open market data can be gathered from https://www.nordpoolgroup.com/en/Market-data1/#/nordic/table or from the FTP-server if reader contacts Nord Pool directly

continuous, it can be transformed into hourly data by calculating an hourly volume-weighted average price, similarly done by Knapik (2017):

$$\text{VWAP}_h = \frac{\sum_{i=0}^{I} Price_{h,i} \cdot Volume_{h,i}}{\sum_{i=0}^{I} Volume_{h,i}} \tag{1}$$

where h indicates a given hour and I is the total number of price and volume i bid observations per hour. Block bids, on the other hand, extend for several hours and large block bids can affect the prices for each of these hours. It is then reasonable to populate its price and volume equally for each hour[6].

5.2 Energy Quantified

The Energy Quantified data is gathered from its API through a Python Client[7] and comprises a wide portfolio of determinants i.e. *weather*, *hydro*, *wind* and *regulation on wholesale energy market integrity and transparency data (REMIT)*[8]. REMIT contains information about abrupt messages e.g. production stop. Energy Quantified provides forecasts, backcasts, actuals and synthetic actuals data for each of the aforementioned determinants with 15 minutes, hourly and daily resolutions. Synthetic data are mainly actual time series corrected for missing or erroneous values by Energy Quantified. The Energy Quantified covariates may contribute to explain the two dependent target variables i.e. day-ahead and intraday prices. We will now dive into particularities of the Energy Quantified data and briefly explain the relevant covariates we are going to use in the neural network models.

5.2.1 Weather forecast models

Energy Quantified either borrows already existing forecasts, indicated by a tag for external issuer, or creates their own forecasts. Especially *weather forecasts* are issued by the ("Weather forecast models and schedule", n.d.):

- *European Centre for Medium-Range Weather Forecasts (ECMWF)*
- *Global Forecast System (GFS)*
- *Icosahedral Nonhydrostatic (ICON)*

[6] E.g. a block bid of 10 MWh volume for 2 hours yields 10 MWh at the first hour and 10 MWh on the second hour

[7] Python Client tutorial: https://energyquantified-python.readthedocs.io/en/latest/. Be aware that only freemium data is available and that we are provided additional data for this thesis

[8] Commonly known as UMM - urgent market messages

- *Arome (Nordic)*
- *UK Met Office (UKMO)*

The Global Forecast System is a American global weather model, that has been popular in the industry due to the public availability, but since the model is primarily not focusing on Europe, the ECMWF is preferred due its higher quality and resolution on topology and geography ("Weather forecast models and schedule", n.d.). Arome and UK Met Office are French- and UK-developed weather models, respectively. Due to their poor results, in contrast to the ICON model, which is the German weather model, these are discontinued after December 31, 2021, and replaced by the ICON model. In this research we opt to not use ICON, because its data only contains forecast from roughly 2022, and we are interested to use data from 2019 to 2022. Summarized, the preferred forecast model, in this thesis, is the ECMWF.

5.2.2 Forecasts

We are particularly interested in models that forecast well on 38 hours horizon, due to the decision tool forecasting horizon from 10 a.m. to next day 12 a.m. in a production setting (see subsection 3.3.1). Since the cross-validation of models is built for this particularity, it is then important to avoid using the latest forecast available as this introduces the problem of training on known future values. It is reasonable to use the latest forecast that is older than 38 hours. Energy Quantified[9] offers the opportunity by allowing for data gathering on day ahead forecast with certain date interval, which we chose to be the latest 2 days (48 hours) ahead forecasts.

Figure 4 illustrates this by an example for multiple forecast series for a specific forecast *ensemble short-term NO2 Wind Power Production MWh/h forecast* in the winter season. This forecast is issued two times a day, 8 a.m. and 8 p.m. shown as grey, and our rolling 38 hours forecast in yellow. It shows which samples are retrieved from Energy Quantified in order to make 38 hours rolling forecast possible in color of beige. The first 38 hours forecast, in figure 4, shows that only first and third forecast series are used, we could not have used the fourth forecast series because it is issued 4 hours after the 38 hours forecast starts. The second 38 hours forecast is even a better example showing that it uses three different forecast series which uses the latest sample from the latest possible forecast series number seven. Same procedure for every multiple forecast series from Energy Quantified.

[9] An in-depth description on how this can be done is provided on this link: https://energyquantified-python.readthedocs.io/en/latest/userguide/instances.htmlnext-steps

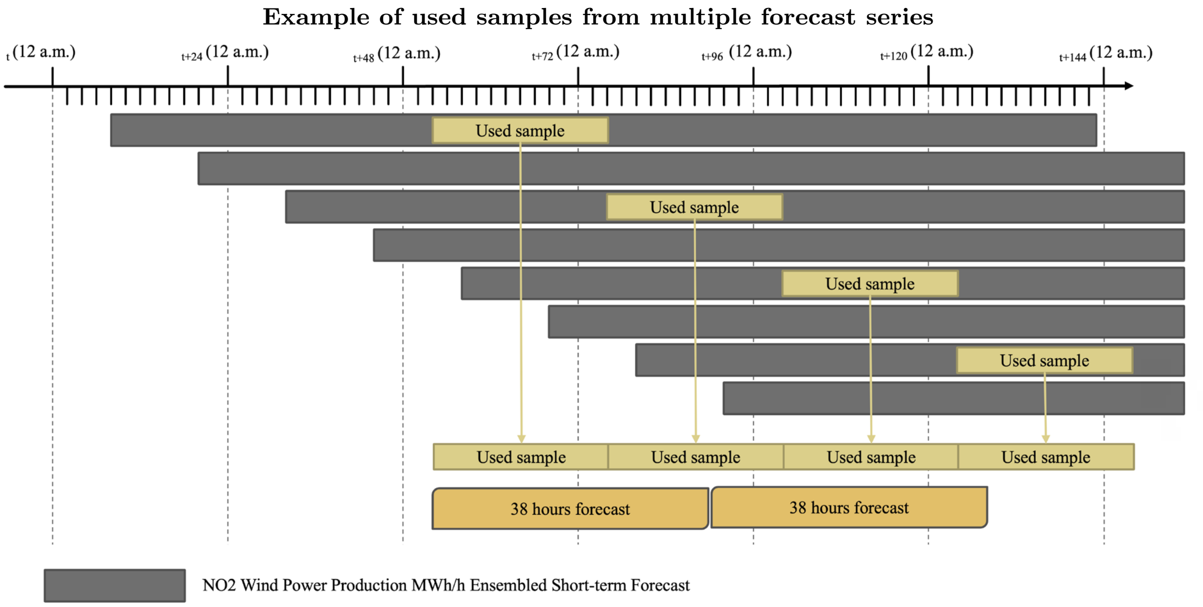


Figure 4: This figure shows how multiple forecast series of ensemble short-term NO2 Wind Power Production MWh/h forecast is retrieved in order to avoid overlapping future forecast when performing 38 hours forecast in our decision tool. *Be aware that this is only an example for a certain forecast series in a certain period of time.

Energy Quantified provides range of different forecast from in-house forecast to forecast issued by third-parties ("Data types for curves", n.d.). In general, they offers deterministic and ensemble forecasts, where deterministic forecast is a single forecast and ensemble forecast is a combination of multiple forecasts. Ensemble forecast is one of the forecasting methods that has been around since 1990s in numerical weather forecasting (Zhu, 2005). Zhu (2005) writes that the mean of ensemble forecasts often yields better performance on short term (3-5 days) in contrast to deterministic forecast, and one could argue that this is a better option. Even though ensemble forecasting yields better results on a longer time horizon than couple of days Boucher et al. (2011) finds in their research on comparison of ensemble and deterministic hydrological forecasts on short term, that ensemble forecast is more beneficial than deterministic forecasting. That ensemble forecast yields better results than deterministic forecast is also discovered by Zhao et al. (2021) in their research on precipitation forecasts.

For spot and exchange forecasts there are mainly two forecasts which is marked with *prefix* and *postfix*. Prefix forecasts are forecast issued before the day-ahead auction closes including forecast for the day-ahead, while postfix forecasts are forecast that are issued after spot auction is closed and actual spot price is published ("Instance tags", n.d.). As our main goal is to forecast before the day-head auction closes at 12 p.m., we opt to gather the prefix forecasts.

5.3 Selected time series

The selected features used in the neural network models are shown in table 1, and mainly four types of time series are used: forecast, actual, synthetic and remit. All except from forecasts are actual data, where the synthetics are actual time series corrected for missing values by Energy Quantified, and REMITs are actual urgent market messages (see section 5.2). There is one time series per area, which means that the number of total features used are the time series multiplied by area name or exchange direction.

As the reader can see in table 1, the selected covariates are of different types and units, and not on a shared resolution. Most of the time series is provided in Mega Watt hours (MWh). *Residual load* is *wind power production* and *solar photovoltaic production* subtracted from *consumption.* In other words, load on the power market that is not affected by wind and solar production, which indicates how much the dispatchable sources have to produce in order to meet the demand in the market. *Residual production* is *scheduled exchange net import on day-ahead* and *nuclear production* subtracted from *residual load.* In other words, how much the power productions there are in the market that are not imported from other areas, produced by wind or solar power sources, or nuclear production.

Power exchange occurs in both directions as import and export. As the *scheduled exchange day-ahead* is a net, the sign in the time series determines which direction the power flows. On the other hand, *actual exchange day-ahead capacity* on the transmission line between areas and countries it may occur, for some connections, that there are different capacity on the import and export cables. It is then necessary to include the exchange capacity in both directions.

Instead of representing the chilling, cloudiness and heating in form of e.g. wind direction or temperature, Energy Quantified represent these in form of consumption index in percentages ("Weather indexes", n.d.). Cooling index is defined as percentage of running installed cooling capacity in countries or areas, heating index is percentage of running installed heating capacity in countries and areas and wind chill index is measured as additional heating capacity needed by a combination of low temperature and wind in countries or areas. Energy Quantified explains it with an example that in the coldest areas in Finland the heating index reaches 100 % at (minus) -25 degrees Celsius, while the heating index reaches 100 % in Italy below zero degrees Celsius. The power consumption does not increase above 100 % as all the heating units are in use.

Selected covariates

Time series	Unit	Type	Resolution	Area
Consumption index chilling	%	Forecast	15 minutes	NO2, DK1
Consumption index cloudiness	%	Forecast	15 minutes	NO2
Consumption index heating	%	Forecast	15 minutes	NO2
Consumption	MWh	Forecast	15 minutes	NO1, NO2, NO5, DK1
Consumption Temperature	°C	Forecast	15 minutes	NO1, NO2, NO5, DK1
Hydro precipitation energy	MWh	Forecast	Hourly	NO2
Hydro reservoir water filling	%	Forecast	Daily	NO2
Hydro run-of-river production	MWh	Forecast	15 minutes	NO2
Solar photovoltaic production	MWh	Forecast	15 minutes	DK1
Spot price	EUR/MWh	Forecast	Hourly	NO2
Spot price short-term	EUR/MWh	Forecast	Hourly	NO2
Residual load	MWh	Forecast	15 minutes	NO1, NO2, NO5, DK1
Residual production day-ahead	MWh	Forecast	Hourly	NO1, NO2, NO5, DK1
Wind power production	MWh	Forecast	15 minutes	NO2, DK1
Schedule exchange day-ahead	MWh	Forecast	Hourly	[DE→NO2, DK1→NO2, GB→NO2, NL→NO2, NO1→NO2, NO5→NO2, SE3→NO1]
Dispatchable power production[a]	MWh	Actual	Hourly	DK1
CHP power production	MWh	Actual	Hourly	NO1, NO2, NO5
Hydro power production	MWh	Actual	Hourly	NO1, NO5
Price imbalance consumption	EUR/MWh	Actual	Hourly	NO1, NO5, DK1
Price regulation down	EUR/MWh	Actual	Hourly	NO2
Price regulation up	EUR/MWh	Actual	Hourly	NO2
Volume regulation netto	MWh	Actual	Hourly	NO2, NO5, DK1
Exchange day-ahead capacity	MWh	Actual	Hourly	[DE→NO2, DK1→NO2, GB→NO2, NL→NO2, NO1→NO2, NO5→NO2, SE3→NO1, DE←NO2, DK1←NO2, GB←NO2, NL←NO2, NO1←NO2, NO5←NO2]
Hydro reservoir production	MWh	Synthetic	Hourly	NO1, NO2, NO5
Hydro precipitation energy	MWh	Synthetic	Hourly	NO1, NO2, NO5
Hydro reservoir available capacity	MWh	REMIT	Hourly	NO2
Hydro run-of-river available capacity	MWh	REMIT	Hourly	NO2
Netto transfer exchange capacity	MW	REMIT	15 minutes	[DE→NO2, DK1→NO2, GB→NO2, NO1→NO2, NO5→NO2, DE→NO2]

Table 1: Showing all the selected independent variables used in the forecast models. Note that this is the raw-format given by Energy Quantified, but the used time series after pre-processing or chosen download resolution from Energy Quantified is given in hours. MWh = Mega Watt hours.

[a] See table 13 in the appendix for more information about the consolidated feature

5.4 Pre-processing

5.4.1 Non-trade hours and missing data in time series

Table 2 lists a summary of non-trade hour records for all time series involved with intraday or balancing market, an exhaustive summary can be found in the appendix. One can observe that the intraday and NO2 Price regulation data hold between 59.9 % and 72.5 % of non-trade hour records as a share of total number of observations. Intraday price and volume data contains 32.3 % of non-trade hours. We want to emphasise that among the non-traded hours there may exist missing values, which is fairly hard to detect as it is not possible to distinguish it from non-traded hours. Regardless of the cause of the missing record we treat it in the same manner and perform linear interpolation.

Summary of non-traded hours for given time series

Series	Non-trade hour	Share of total
NO2 Price Regulation Up EUR/MWh H Actual	17628	72.46%
NO2 Price Regulation Down EUR/MWh H Actual	14565	59.87%
Intraday trade buy volume	7853	32.28%
Intraday Price Difference	7853	32.28%
Intraday Price	7853	32.28%
NO5 Volume Regulation Net MWh H Actual	87	0.36%
NO2 Volume Regulation Net MWh H Actual	87	0.36%
DK1 Volume Regulation Net MWh H Actual	87	0.36%

Table 2: Summary of non-trade hours in time series. Note that there may exist missing values among the non-traded hours, which is not possible to detect.

Missing data records is particularly problematic for time series data, since the records has a temporal evolutionary effect (Bergmeir and Benítez, 2012). Figure 14 in the appendix shows all the missing data records of the selected time series. The share of missing values of total number of records are beneath 0.16 %, which indicates that the amount of missing values are relative low.

Figure 5 shows an example of *consecutive non-trade hours frequency* of the particular custom-engineered *intraday NO2* time series (see section 5.1). One can also find similar tendency in the other time series containing missing records, which is not provided. One can observe that most of the non-existing records occurs with single hour interval, and that it gradually decreases with increasing hour interval. In order to get a similar format as the continuous day-ahead time series adding empty records for the missing hours is necessary, as these are not included in the gathered time series. Hourly non-trade volume for the intraday data can be replaced with zero value, as these are actually non-trades occurring in the market, but where replacement of non-existing prices need more attention. The simplest technique is to replace the non existing prices with the average of all prices, but this is a very biased method that leads to great deviation in the time series due to the increased fluctuation in prices in the last years.

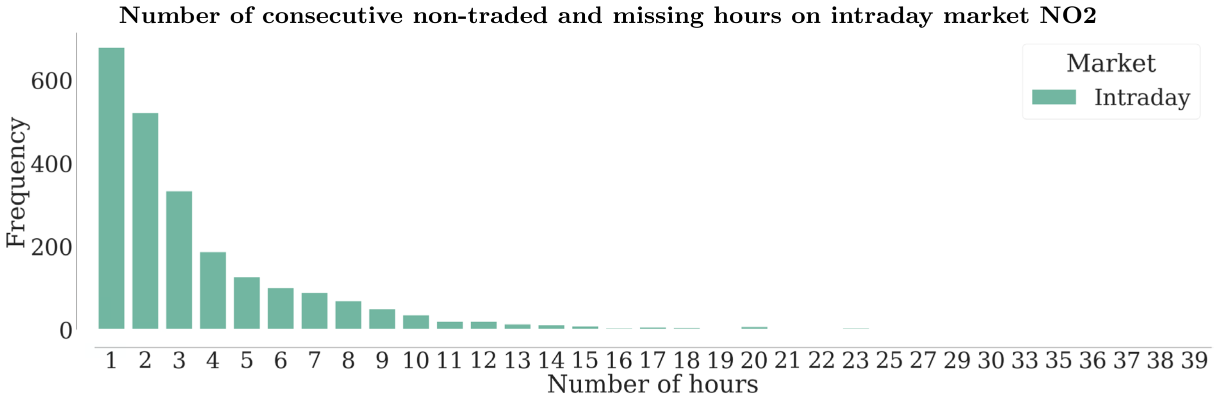


Figure 5: Number of consecutive non-trade hours between traded hours on intraday NO2 in 2019-2022

Junninen et al. (2004) stated that replacing missing records with the average values disrupted the inherent structure of the data and decreased model performance. Another technique is to impute the missing values in-between observable prices (neighbor prices) using linear interpolation, which Mohamed Noor et al. (2014) confirms to be better than the averaging method. The equation for this method is as follows (Canale and Chapra, 1998):

$$f(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0) \tag{2}$$

where x is the covariate, x_1 and x_0 are know values from ahead and previous values of the covariate and the function $f(x)$ is the transformation from missing value to a linear interpolated value.

Other more advanced methods such as nearest neighbor, neural network, multiple imputations or more hybrid model combining different techniques are considered. Studying the performance, speed and reliability in the paper of Junninen et al. (2004) reveals that linear interpolation performs well on short time gaps, is reliable on short time gaps and little less on long time gaps, but is the fastest method to compared to other methods researched. They also states that missing data is lost in entirety and a good approach will remedy the problem as far as possible. Considering time, effort and quality of choosing a imputation technique we opt to replace the missing price data using linear interpolation, even though Junninen et al. (2004) reveals that the more advanced methods performs slightly better. A minor critique of this technique is that the precision will be reduced and may lead to some distorted results in the forecast, but the alternative is to reduce the amount of data by removing actual observations of day-ahead data for the corresponding missing hours in the intraday time series.

As we are conducting cross-validation technique, in order to decide model hyper-parameters and perform model evaluation, using train, validation and test-sets it is necessary to be aware of information leakage. Since using pre-processing before performing cross-validation, in this case, using linear interpolation leads to information leakage between the cross-validated folds. In order to overcome this problem, the imputation technique can be performed on each fold after the fold split are determined. Performing it this way, introduces a new problem that each fold can contain missing values in the start and at end of the splitted time series. Missing values in the end of a time series fold is not an issue as these missing values are imputed with last known value observed. Missing values at the start of a series cannot be replaced by a linear interpolation, as the equation 2 does not have any initial value, because linear interpolation depends on previous and ahead observed data records. For simplicity we chose to impute these values with the first non-missing value as long as the entire series does not contain only missing values, since these are replaced with zero. For *categorical features* the missing values are replaced using forward filling of existing values.

5.4.2 Feature scaling

The selected covariates in this paper comes in different scales, e.g. Mega Watt hour or percentages. Machine learning models that is trained on scaled input feature usually yields higher performance compared to models that are not, which means that feature scaling is an important step in pre-processing of data (Cao et al., 2016). Haykin (2009) states that scaling of numerical features in deep neural network models is required in order to get a faster and more stable learning process. As there exist a variety of scaling methods it is required to find a proper scaling method that suits the electricity data which we have gathered. Ahsan et al. (2021) finds that model performance fluctuates with different scaling methods, but also reveals that there does not exist a single best scaling method. As we opt to use the same feature scaling method among all the features, and as not all of the features are either normally distributed or free for outliers, we choose to perform *Robust Scaling (RS)*[10]. Robust Scaling is a method that removes the median and performs the scaling in the *interquantile range (IQR)*, meaning that the observations centering and scaling statistics are not affected by outliers (Baijayanta, 2020). The Robust Scaling equation is presented as:

$$RS(x_i) = \frac{x_i - x_{median}}{x_{3rd-quartile} - x_{1st-quartile}} \tag{3}$$

where i is the ith observation in a feature minus the feature's median divided by IQR which is the 75th percentile minus 25th percentile of the feature. This type of scaling takes outliers into account by not include it in the scaling process. The resulting scaled observation has then zero mean and zero median with a corresponding standard deviation of one.

5.5 Data Exploration

In order to justify the relevance of our decision tool, choosing to buy on intraday or day-ahead market, we need to study if there exist enough trade-volume and that power prices on the next day intraday are significantly lower than day-ahead market prices. We have scoped the problem at hand to the power region NO2 (South Norway), due to the fact that it is the most liquid market of the Norwegian Nord Pool regions.

[10] See figure 30, 31, 32, 33 and section B.1.1 in the appendix

5.5.1 Buy-volume on intraday markets

Figure 6 shows median buy volume on the different intraday markets in Norway. It reveals that NO2 (orange line) is the only price region that by median has continuous trade-volume greater than zero MWh volume. It is noticeable that the other power areas, NO1 and NO3-NO5, normally (median) do not have any buy volume between 5 a.m and up until 11 a.m. Observing the plot, figure 6, in combination with missing trade-hours in table 7 shows that NO2 has least number of missing trade-hours, 32 % of total number of traded hours, in the intraday market. This supports our decision of choosing NO2 as the region to exploit the opportunity to buy power on the intraday market in this case.

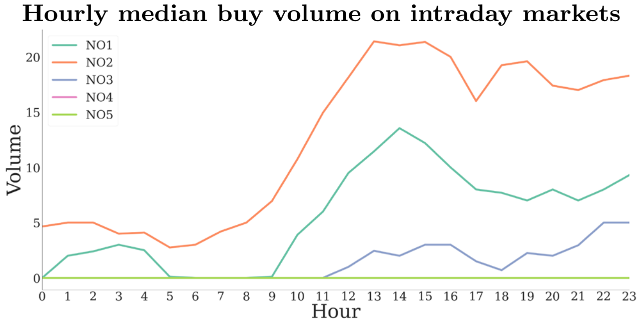


Figure 6: Median buy volume on intraday markets per hour in 2019-2022

Missing trade-hours

Area	Frequency	Share
NO2	8540	32.0%
NO1	11289	42.3%
NO3	13668	51.3%
NO4	18828	70.6%
NO5	18974	71.2%

Figure 7: Number of non-traded hours and share of total number of hours

Studying table 3 one can see that the traded buy-volumes are significantly lower on the intraday markets in contrast to the day-ahead markets, with median MWh between 0 and 10.8 from 2019 to 2022. It is noticeable that a power companies act of buying on the intraday market yields smaller portfolio position, because the volume is significantly lower than day-ahead, with a ratio of 0.29 % for NO2. But, it is still interesting for all power companies to exploit a buy-opportunity after all. As NO2 still reveals the highest median MWh volume, it is far more attractive to pursue the buying strategy on intraday market in this area. For more details on buy-volume spread in the different areas a descriptive statistics on intraday and day-ahead is provided in table 15 in the appendix.

Intraday median volume share of day-ahead median volume

Market	Day-ahead	Intraday	Share
NO2	3721.3	10.8	0.290%
NO1	3675.4	5.0	0.136%
NO3	2912.2	0.0	0.000%
NO4	1662.2	0.0	0.000%
NO5	1653.4	0.0	0.000%

Table 3: Intraday median volume share of day-ahead median volume

5.5.2 NO2 markets

In the previous section about buy-volume in intraday markets, we restricted the case to yield power area NO2. This section will then only focus descriptive statistics on price and volume for power area NO2. The line plots in figure 8 and 9 shows the evolution of median buy-volume, in mega watt, weekly from 2019 to 2022 for day-ahead and intraday, respectively. On the day-ahead market the median volume cycles approximately between 2,500 and 5,000 MW, where the volume is lowest in summer season and highest in winter periods. One can see that the volume on day-ahead market for NO2 has not changed volume level in this period. The weekly volume on the intraday market on the other hand reveals a upward trend from 2019 to 2022, which indicates that trades on this market has increased. This is attractive for the papers problem because it indicates that either more buyers or increased volume of power companies' portfolio is increasing.

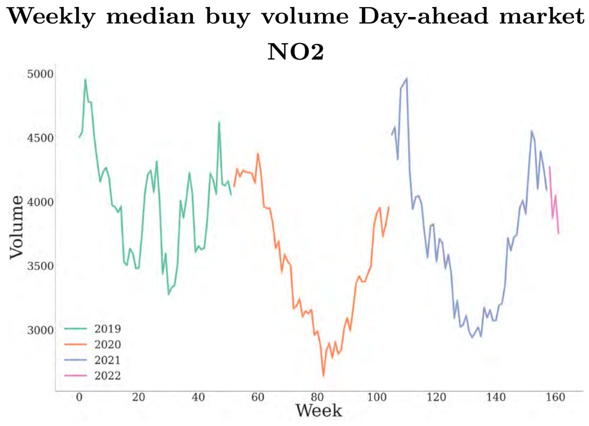


Figure 8: Weekly median buy volume development on day-ahead market in 2019-2022

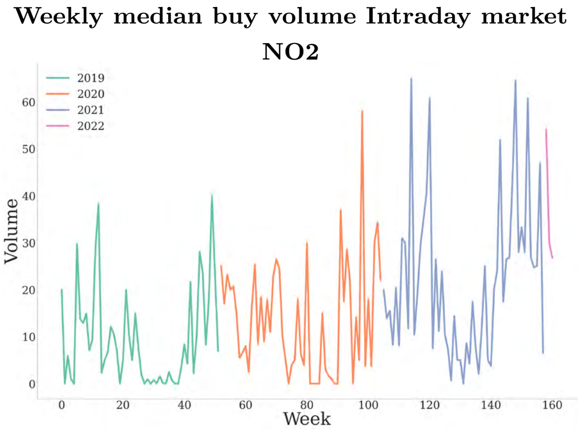


Figure 9: Weekly median buy volume development on intraday market in 2019-2022

Not only is volume an important factor for making power purchases attractive on NO2's intraday market, but we also need to consider the prices. Observing figure 10 it clearly indicates that the intraday prices for NO2 on median is lower than the day-ahead market prices. Studying the confidence interval the prices for the markets can overlap between 6 a.m. to 9 p.m. This is fundamental for making our decision tool usable in practice. Power trading occurs daily, as aforementioned in section 3, and figure 37 in the appendix shows that power prices is relative stable from Monday to Sunday on median both for day-ahead and intraday market NO2, but where the spread in price is broader in weekdays in contrast to weekends.

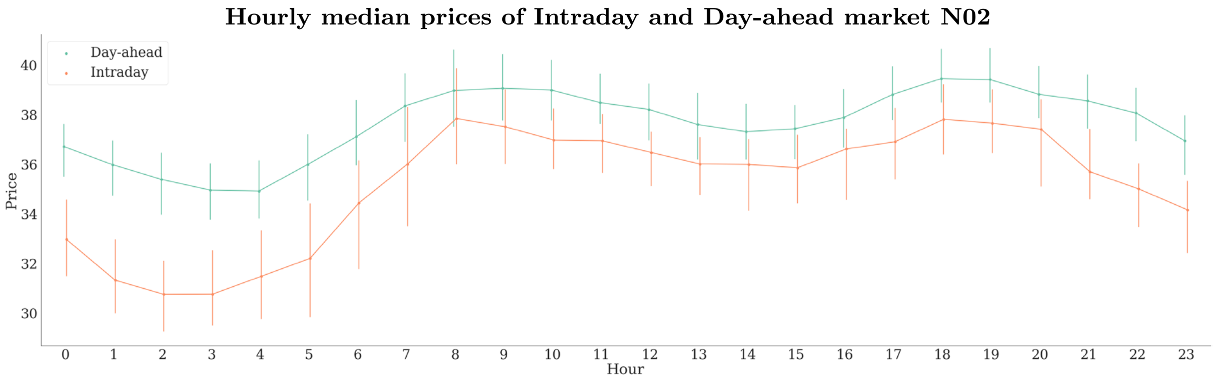


Figure 10: Median price of intraday and day-ahead NO2 market per hour in 2019-2022

In order to make forecasts on the hourly time series of prices it is important to understand how the price has evolved over time. Observing figure 11 it shows that power prices both in day-ahead and intraday market have been relative stable from 2019 to early 2021, where it thereafter becomes more volatile. Roughly speaking the time series has become non-stationary in the later period in time.

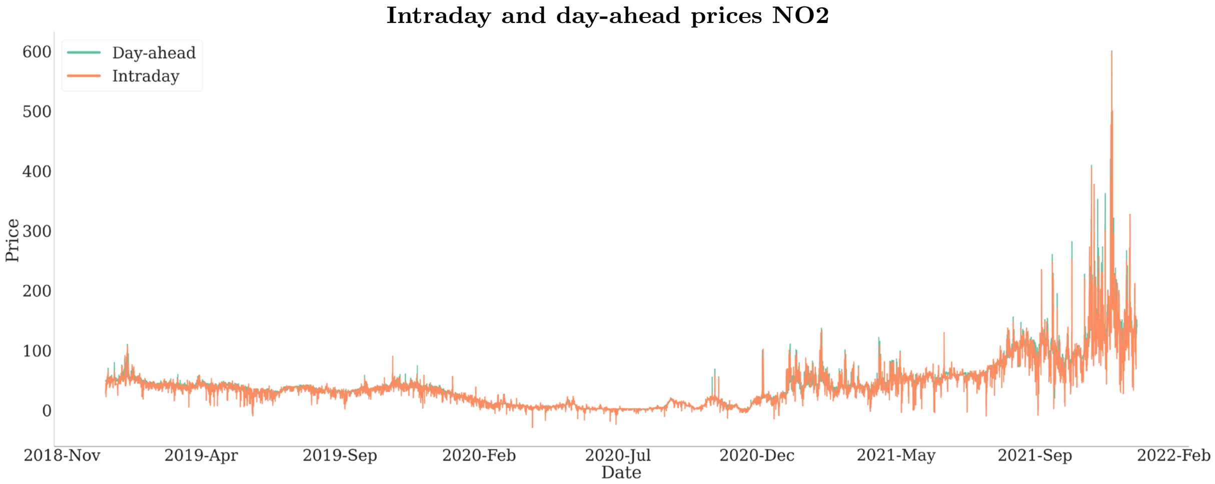


Figure 11: Intraday and day-ahead prices for area NO2 in 2019-2022

6 Methodology

In this thesis we seek to use deep learning architectures to provide accurate forecasts for Nord Pool day-ahead and intraday markets. These forecast will act as the foundation for a decision tool for electricity buyers. A probabilistic classifier trained on deep learning forecasts, and validated on historical price differences could plausibly provide useful information for electricity buyers at a critical decision point.

In this section we will provide explain the general outline of our decision tool, a basic introduction into neural networks in general, and the specific architectures applied in this thesis. To evaluate the implemented neural network models, a selection of benchmark models will be validated in an exactly similar manner to that of the neural models. The cross validation techniques applied in this work is a response to the peculiarities of dealing with non-stationary time series data, and will be explained in that context. We start off by elaborating on our end goal, a simulated production of a decision tool for electricity buyers.

6.1 Decision tool outline

A central goal of this thesis is the creation of decision tool, enabling power buyers to make an informed decision about which market to participate in for any delivery hour. To achieve this, we need accurate forecasts for the intraday and day-ahead markets.

The general outline of this decision tool can be seen in figure 12, where the difference between day-ahead and intraday market price is used as training data along with a small selection of predetermined important features. The end goal of the classifier network is to correctly classify day-ahead-intraday price based on available forecasts. Training and weight adjustment is done by calculating the binary cross entropy loss of the logits of the model and observed price differences of day-ahead and intraday markets. The networks shown in the figure is a multilayer feedforward network, but we also tested architectures with LSTM layers. We have viewed this classification problem as a point-point forecasts, and as a result, data is not treated sequentially. Figure 12 shows an outline of how data forecast data is passed to the probabilistic classifier and classifications generated.

Design of concept

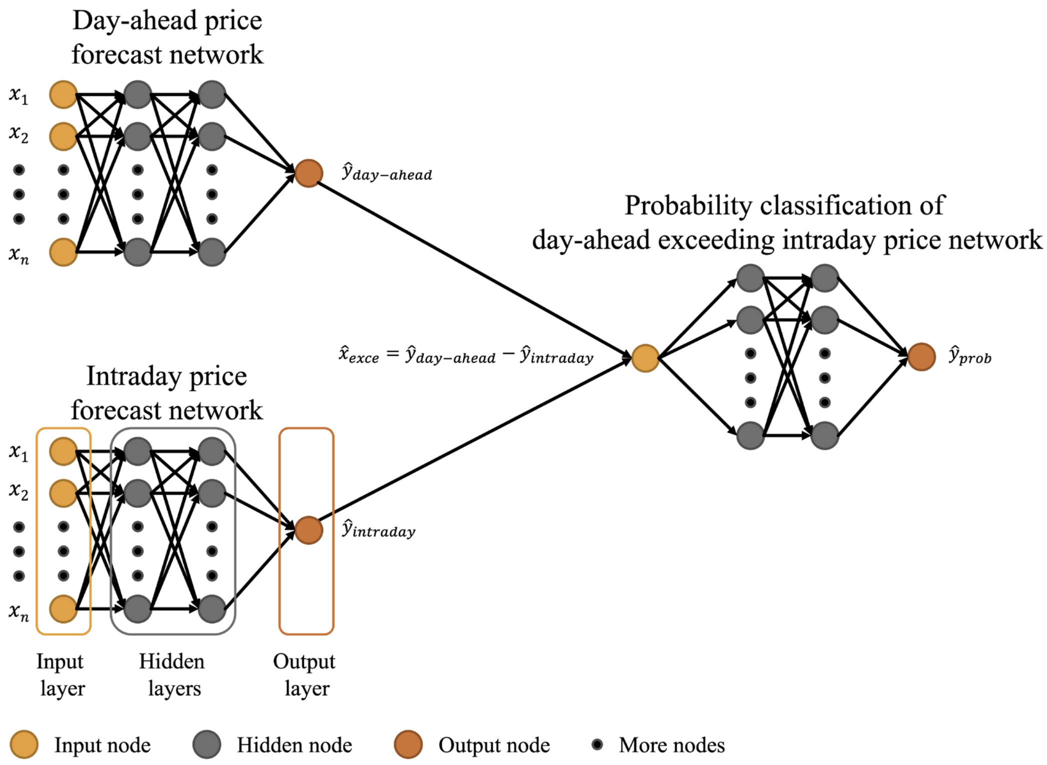

Day-ahead price forecast network
x_1
x_2
x_n
$\hat{y}_{day-ahead}$
Probability classification of day-ahead exceeding intraday price network
$\hat{x}_{exce} = \hat{y}_{day-ahead} - \hat{y}_{intraday}$
$\hat{y}_{prob}$
Intraday price forecast network
x_1
x_2
x_n
$\hat{y}_{intraday}$
Input layer
Hidden layers
Output layer
Input node
Hidden node
Output node
More nodes

Figure 12: Whole neural network process from price forecast to probability classification of day-ahead exceeding intraday price

An electricity buyer would have to make decision for every delivery hour before day-ahead market submission deadline. Any information available a later time, such as updated wind-forecasts will not be available at the critical decision point. We emphasize that forecasts for both markets are made using solely information that would be available for an electricity buyer at the time of decision. In practical terms this means that the forecasts and resulting probabilities of intraday discount has to be ready between 10 a.m (as day-ahead capacities are released) and before closing time for bids at the day-ahead market at 12 p.m.

Figure 13 shows a simulated production use-case , where the start point of forecasting every day at 10 a.m. (see section 3.3.1) with a forecast horizon of 38 hours in order to capture the whole day-ahead market next day. As forecast accuracy degrades with increasing horizons Kyritsis et al. (2017), increasing forecast length beyond 38 hours would only hinder forecast accuracy as new information has been made available in the meantime.

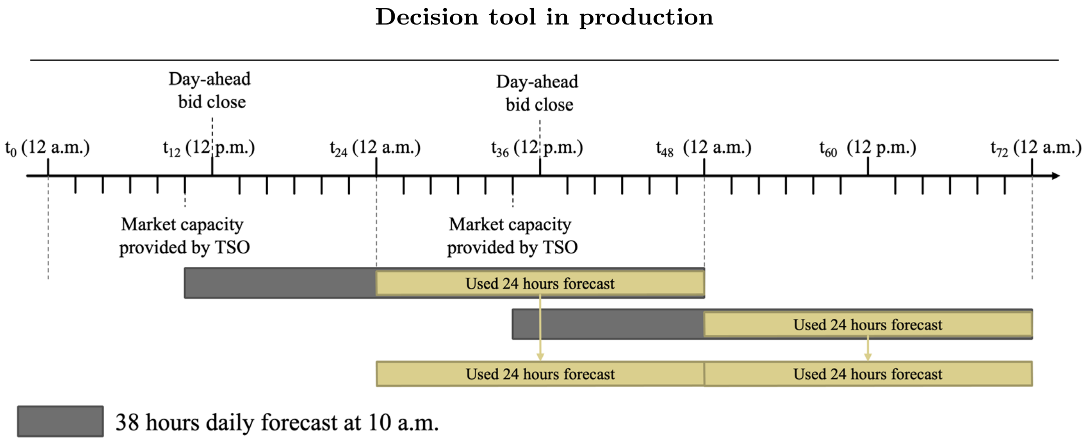


Figure 13: Shows how the decision tool will be used in production

The training needed to produce the needed day-ahead and intraday forecasts requires considerable computational resources. This puts constraints on how far an electricity buyer can postpone training the models used for the next day's forecasts. Assuming that necessary computational resources are available for the electricity buyer, a decision window of two hours should be sufficient to train and decide optimal market for every delivery hour.

6.2 Simulated use-case of decision tool

The goal of the decision tool outlined above is to provide probabilistic classifications of likely day-ahead-intraday price difference, where a probability above 50 % indicates likely intraday discount. Depending on the required, liquidity in the intraday market may not be sufficient to cover all or any part of the bid. This presents a considerable risk for a potential intraday electricity buyer.

By participating in intraday trading, the electricity buyer incurs a risk of not fulfilling part or the entire bid volume. Grid stability has to be ensured, and the mismatch between demand and fulfilled intraday and day-ahead volume has to be bought in the balancing market. The discrepancy between the buyer's demand and the volume fulfilled in intraday markets will have to be covered in the regulating markets. For consumers the regulating market follows a one-price system, where the price depends on the direction of the overall system imbalance ("Regulation information per area", n.d.). The system balance can be described in the equation below, where the added volume by leaving a bid to the regulating market is given by V_b.

$$system\ \widehat{imbalance} = system\ imbalance - V_b \tag{4}$$

In reality, as demand fluctuates, an electricity buyer may have to decrease *and* increase bid volume passed to the regulating market. We should emphasize that we have made three crucial assumptions when simulating the use of the decision tool.

1. The total bid volume of the electricity buyer is fixed and predetermined before day-ahead market submission deadline

2. A trade is made in its entirety in day-ahead, intraday or regulating market.

3. When participating in the intraday markets, the observed price is paid and volume is adequate to fulfill the bid in its entirety.

With the above assumptions in place we can boil down the scenarios facing an electricity buyer, and illustrate them in a decision tree:

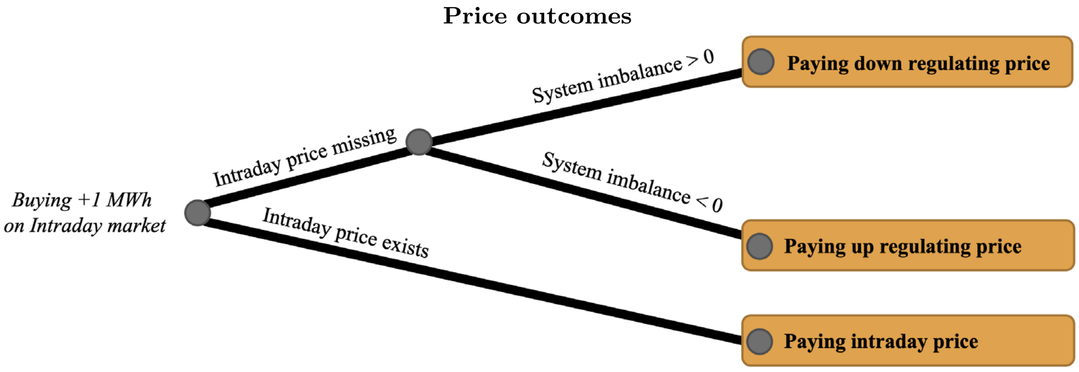


Figure 14: Different price outcome if deciding to buy on intraday market

Figure 14 shows the prices encountered by an electricity trader that has decided to participate in intraday trade with the entirety of its bid volume. The bottom branch of the tree shows the scenario in which price formation in intraday market is successful and the bid of the buyer is accepted in its entirety. The resulting price is assumed to be the observed intraday price for the delivery hour. We should again emphasize that using the observed intraday price as a metric for intraday price is problematic. Weber (2010) finds that liquidity in intraday markets often are insufficient for a buyer to remain a price-taker, and that large bids will affect intraday prices. It is thus unlikely that any large bids would end up with ex-post intraday price.

If no liquidity exists for a given delivery hour, as determined ex-post by a missing intraday price, the bid is processed in the regulating market. Depending on the regulation direction, the buyer can either pay the up or down regulating price following the one-price system for consumption imbalances (“Regulation information per area”, n.d.)[11].

[11] For a brief description of the one-price and two-price systems see A.1.1

Based on the price-scenarios outlined above, a profit function for an electricity buyer can be formulated. When demonstrating a simulated production use-case of this decision tool we solve the decision-process using various risk-thresholds μ.

$$\Delta\hat{P}_{h,d} = \hat{P}_{h,d}^{day-ahead} - \hat{P}_{h,d}^{intraday}, \forall \ \{h \mid 1 \leq h \leq 24, \ h \in N\}, \ d \in N \tag{5}$$

$$\text{System imbalance} \equiv \vartheta \tag{6}$$

$$\pi_{h,d} = \begin{cases} P_{h,d}^{day-ahead} - P_{h,d}^{intraday}, & \text{if } Prob(\Delta\hat{P}_{h,d}) \geq \mu \ \& \ \exists P_{h,d}^{intraday} \\ P_{h,d}^{day-ahead} - P_{h,d}^{\text{up regulation price}}, & \text{if } Prob(\Delta\hat{P}_{h,d}) \geq \mu \ \& \ P_{h,d}^{intraday} \ \& \ \vartheta < 0 \\ P_{h,d}^{day-ahead} - P_{h,d}^{\text{down regulation price}}, & \text{if } Prob(\Delta\hat{P}_{h,d}) \geq \mu \ \& \ P_{h,d}^{intraday} \ \& \ \vartheta > 0 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

$$\text{Profit} = \sum_{d=1}^{D} \sum_{h=1}^{H} \pi_{h,d} \tag{8}$$

Equation 5 $\Delta\hat{P}_{h,d}$ shows the probability of day-ahead prices exceeding intraday prices, where $\hat{P}_{h,d}^{day-ahead}$ and $\hat{P}_{h,d}^{intraday}$ are forecasts of day-ahead and intraday prices for delivery hour h in day d, respectively. In the first condition of equation 7 $\Delta\hat{P}_{h,d}$ exceeds our risk threshold μ, and the profit is given by the difference between observed day-ahead price $P_{h,d}^{day-ahead}$ and intraday price $P_{h,d}^{intraday}$ ex-post. In the second and third equations we still decide to participate in intraday markets as the probability is above our risk threshold μ, but in this intraday price and thus volume is non-existent and we pay either downward or upward regulation price. The profit function compares intraday trade with a baseline day-ahead trading approach, where all trade is done exclusively in day-ahead markets. This implies that paying the day-ahead price gives a profit $\pi_{h,d}$ of 0 as seen in the fourth condition in equation 7. This occurs when production and consumption is a state of balance, and no downward or upward regulation price exists. System balance will only persist as long as the bid volume is not significant enough to cause a consumption excess, which is plausible only in the case of small volumes. The total profit in our production period is the sum of all marginal profits for every day and delivery hour shown in equation 8.

6.3 Neural Networks

Neural networks have gained much attention in the recent decades and years much to its inherent flexibility. The name "neural network" is derived from the analogous similarity with the way our brain works trough its neurons connected by synapses, referred to as weights. Theoretical proofs have shown that neural networks can mimic any functional form given an adequate amount of neurons, making them universal approximators (Hornik et al., 1989). Similar proofs exist for networks of arbitrarily large depths (Lu et al., 2017). Without the assumption of infinite neurons however, width or depth becomes a trade-off. In recent years deeper neural networks been shown to have greater performance given the same overall network size. Eldan and Shamir (2016) prove that to approximate a given function f, an exponentially larger number of neurons would have to be added to a network composed of 1-layer fewer.

This relatively recent focus on network depth rather than width births the sub-field of *deep learning*. Deep learning is often ambiguously used, but a common definition is any neural network with at least one hidden layer (Goodfellow et al., 2016). Good complexity-performance trade-offs makes deep neural networks attractive for a wide set of problems. In our paper, all neural networks architecture are to a certain extent regarded as "deep".

In this thesis we focus primarily on LSTM, GRU, TFT and DeepAR models. Before going into details on these neural networks' architecture, we are now going to provide insights from the most basic neural network models, which is fundamental in order to understand the more complex models.

6.3.1 Feedforward and Recurrent Neural Networks

In its most classical sense, a *feedforward neural network* considers an array of input data $X = [x_1, ..., x_n]$ and maps them to an output prediction $\hat{y}$. Feedforward networks distinguish from other networks which considers groups of data in connection e.g. RNN, which considers data points in sequence, and CNN. A feedforward network attempts to fit a function $f(x;\theta)$ that most closely maps x to y, actual observations, by altering model weights θ (Goodfellow et al., 2016). In a multilayer neural network the value of each neuron is determined by a weighted sum of connected neurons. In a two-layer network composed of n neurons the value of neuron h, in the second layer, can be expressed as $h_i = g(x_1 \cdot \theta_1 + x_2 \cdot \theta_2 + ... + x_n \cdot \theta_n + b)$, where x_i is input data, θ_i the connected neuron weight and b denotes the bias, i.e the threshold in absolute terms needed for the neuron to become active. To introduce non-linearity into the neural network the input of every neuron is passed through an activation function g. Historically *sigmoid* σ activation functions have been used, but recently *rectified linear unit (ReLU)* has become popular as seen in Glorot et al. (2011), expressed as $\sigma(a_i) = \frac{1}{1+e^{-a_i}}$. The ReLU activation function is defined as the maximum of 0 and the neuron's input, $f(a_i) = max(0, a_i)$. It quickly becomes tedious to fully express a large neural network in the aforementioned fashion, and as result we commonly use vectors to express neural networks. The value of neurons h in vector-form can be expressed as $h_i = g(x_1 \cdot \theta_1 + x_2 \cdot \theta_2 + ... + x_n \cdot \theta_n + b) = g(\mathbf{W}^T\mathbf{x} + \mathbf{b})$. In the remainder of this thesis we will use vector notation as often as possible.

A feedforward neural network can be seen as a graph composed of chained functions, which can be seen in figure 15, where an additional layer denotes an added function. In the case of a network with three layers, constituting a *multi-layer perceptron (MLP)*, the output of the network can be described as a recursive application of f^i on x $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$. The depth of this recursion determines the number of *hidden layers*. In a feedforward network the connections are without cycles, and the outputs of the model is only passed a single way, forward from input to output, hence it represents a directed acyclic graph. In a recurrent network, feedback of model output is allowed, making it a cyclic graph.

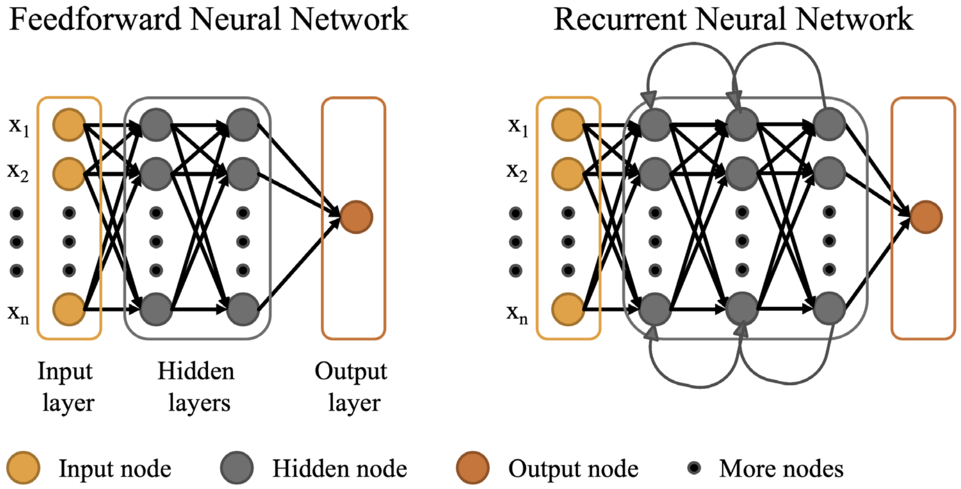


Figure 15: Feedforward vs. Recurrent Neural Network.

Starting off with a predetermined number of neurons, the weights θ are initialised to a small, often random, values. The initial weights may also follow a density function. As the initialisation values determine our starting point for further searches for local and global optimums, they may influence whether the network training converges at all (Goodfellow et al., 2016). Learning and thus altering of the initial weights are done by backward-propagation. In short this involves calculating model mismatch between target data, and our model predictions. In a regression setting, the mean absolute error can serve this purpose, as given by: $L(\theta) = \frac{1}{D}\sum_{(x,y)\in D} |y -$

$f(x|\theta)|$ (Goodfellow et al., 2016). By adjusting model weights the loss function can be minimized. As the loss for function is differentiable, we can describe the gradient of the loss function denotes the magnitude of weight change, given by $\theta_{t+1} = \theta_t - \gamma \nabla_\theta L(\theta_t)$, where θ denotes the vector of model weights, $\nabla = \frac{\partial L}{\partial \theta}$ and γ the *learning rate*. This approach is called *gradient descent* as we gradually move towards a local minimum. This local minimum may or may not be a global minimum, and gradient descent does not guarantee global optimality.

The loss function $L(\theta)$ is calculated after the model has seen a single subset or *batch* of data. If the batch or subset constitutes the entirety of the dataset it is called *vanilla gradient descent* (Ruder, 2017). Performing weight updates after seeing potentially numerous examples of the same data is inefficient, and we could greater utilize our data by weight updates after every new occurrence of data. The resulting fluctuating descent means that the convergence will be sporadic and with greater variance, allowing us to potentially skip over unwanted local minimums. This approach is called *stochastic gradient descent*.

In practical applications it is more efficient to split the dataset into many smaller batches, updating model weights after each pass on the mini-batch of n examples expressed as $\theta = \theta - \gamma \cdot \nabla_\theta L(\theta; x^{(i;i+n)}; y^{(i;i+n)})$ (Ruder, 2017). This approach is often referred to as *mini-batch gradient descent*, although the terminology of gradient descent methods is often used interchangeably and inconsistently. Several optimization methods combines the approaches of stochastic and mini-batch gradient descent such as *ADAM* (Kingma and Ba, 2014) and *Ranger* (Wright and Demeure, 2021). An in-detail discussion of optimization techniques comes in section 6.5.3. The key-takeaway of these approaches is the alteration of the learning rate γ as model training progresses, i.e applying an *adaptive learning rate*.

The training process in a feedforward network can be described algorithmically, loosely based on Chollet (2018) and Goodfellow et al. (2016) shown in algorithm 1. To more closely fit the probability network later described in this thesis we will use a sigmoid σ function on the output layer. The activation function ReLU is applied to the hidden units. In the network described below the initial weights values where randomized, but often follow distributions such as *Lecun* or *He initialisation* (Boulila et al., 2021).

Algorithm 1: Training in a multilayer feedforward neural network

```
Initialize weights $\forall\theta \leftarrow k \ \in \ [-0.5, 0.5]$ ;
foreach $s \in \mathcal{S}$ where $S$ is a set of all mini − batches do
    Forward pass:
        1. Receive input from all connected preceding neurons. The input in neuron $h_i$ is given by
           $z_i = W^T x + b$
        2. Apply activation function to the hidden units $h_i = ReLU(z_i)$

    Backward pass:
        1. From the mismatch between $y$ and $h$, store the loss $L(\theta)$ from the current batch. Using MAE
           $L(\theta) = \frac{1}{D}\sum_{(x,y)\in D}|y - f(x|\theta)|$
        2 . Update weights according to loss function $\theta_{t+1} = \theta_t - \gamma\nabla_\theta L(\theta_t)$ The updated weights are
            skewed in the direction of the gradient, reducing expected loss.
end
A single epoch of training is complete
Final predictions are made after passing through the output layer $\hat{y} = \sigma(W^T h + b)$
```

6.3.2 Incorporating time into Neural Network models

In the feedforward model each data point is considered separately, not allowing for relationships between data points of the same series. A pragmatic approach could incorporate relevant variable lags of the target variables. This methodology allows for some simple time-based relationships by adding time-steps as features, and complexity is increased dramatically as we increase our time-steps. Recurrent neural network allows for inherent modelling of sequential data.

Unfortunately handling long-term and short-term relationships in the same neural network presents problems. Originally described in J. Hochreiter (1991) as a problem faced when attempting to use backward propagation with sequential data. The problem can be shown with the last part of the gradient descent expression, the gradient of loss function L, $\delta L(\theta) = \frac{\partial \mathbf{L}}{\partial \theta}$ (see subsection 6.1.1) which can be expressed as $\frac{\partial \mathbf{L}}{\partial \theta} = \sum_{t=0}^{T} \frac{\partial L_i}{\partial \theta} \propto \sum_{t=0}^{T}(\prod_{t=k+1}^{y} \frac{\partial h_t}{\partial h_{i-1}})\frac{\partial h_k}{\partial \theta}$, where h_t is a hidden state, one can observe that vanishing gradient occurs when $|\frac{\partial h_t}{\partial h_{t-1}}| < 1$ and exploding gradient when $|\frac{\partial h_t}{\partial h_{t-1}}| > 1$ given a large value of T (Or, 2020).

6.3.3 Partially solving the vanishing gradient problem

Long short-term memory model solve the problem of vanishing gradients, by gating connections, allowing time dependent information to enter the network without altering previous information. New input data is fed to input gate, and existing temporal relations are kept unaltered by outputs from the previous timestep h_{t-1}. This means that backward propagation can occur without increasing the magnitude of the learned gradients, allowing for larger but not unlimited time spans. The memory cell's forget gate as presented first in S. Hochreiter and Schmidhuber (1997), uses a sigmoid function σ forcing the output to constrict to the interval $[0, 1]$, where 0 represents a complete reset and 1 a copy of the last state C_{t-1}. The forget gate computes this degree of forget-fullness by taking in the network's last state with a bias b_f, given by $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$ (Olah, 2015). This can be seen in the sigmoid σ layer leftmost in the figure below.

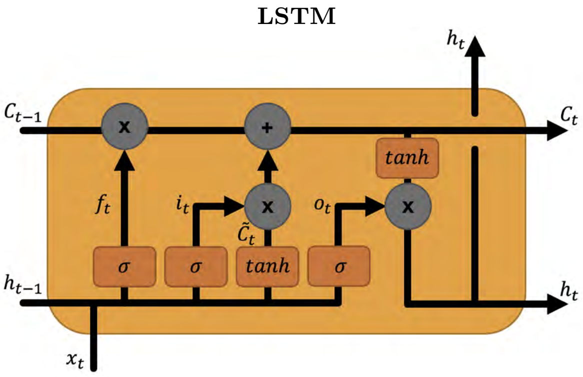


Figure 16: LSTM. Adapted from Understand LSTM Network, in *Colah's blog*, by Olah, 2015, retrieved May 1, 2022, from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

The current contents of the memory cell c_t is updated according to the new input values i_t, given by $C_t = f_t \cdot C_{t-1} + i_t$.

Although introducing added complexity, LSTM has become a common way of at least partially dealing with the problem of vanishing gradients and allowing for long-term dependencies in our neural networks. The partial solution to the vanishing gradient problems allows for more stable predictions, and have been shown to outperform traditional MLP networks as seen in Shah et al. (2018).

The added complexity of long short-term models presents problems, and numerous simplifications have since been made. Gated recurrent unit first introduced in Cho et al. (2014) builds on LSTM networks, combining the LSTM forget and input gate into a single update gate z_t as seen in the figure below (Olah, 2015). By removing the forget gate a significant number of parameters have been pulled from the network, but has lost its ability to control to degree of exposure to previous states C_{t-n} (Chung et al., 2014).

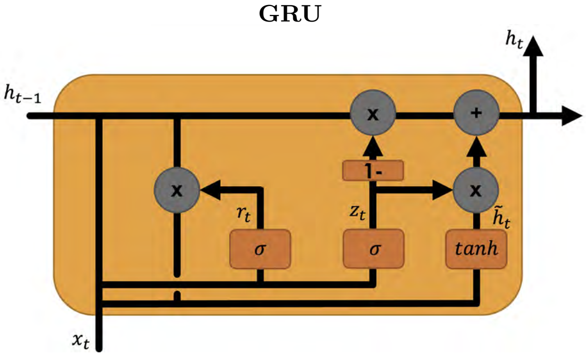


Figure 17: GRU. Adapted from Understand LSTM Network, in *Colah's blog*, by Olah, 2015, retrieved May 1, 2022, from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Chung et al. (2014) finds that GRU offers similar performance, at a computational discount. The computational complexity of forecasting using large recurrent neural network is already high. Any levitation in computational requirements are welcomed. In that regard we view GRU models as a promising avenue for electricity price forecasts.

6.3.4 DeepAR

DeepAR is a forrecast model based on *autoregressive recurrent networks*, other words LSTM RNN, and has been proven to improve forecast performance relative to state-of-the-art forecasting methods on a wide variety of datasets (Salinas et al., 2020). Challenges according to using multiple time series as independent variable is that magnitudes differ widely and distributions is strongly skewed in practical applications. Four main advantages of using DeepAR models in contrast to state-of-art models are:

1. Minimal manual handling of provided independent variables is needed, as the model learns seasonal behaviors and complex dependencies with minimal tuning.

2. The benefit of learning from other similar covariates makes forecasts possible even though there may be little historical data on the dependent variable, which traditional forecast methods are not able to perform as they truly use the dependent variable's historical information.

3. Can produce probabilistic forecasts generated from Monte Carlo simulation samples, which can be used to create forecast with quantiles.

4. Can incorporate a wide range of likelihood functions, such that user choose the suitable likelihood functions for the statistical property of the data.

DeepAR solves two of the LSTM shortcomings which are: the problem of fitting outliers due to the uniform distribution sampling, and handling of temporal scaling (Berk, 2021). Both being solved by aggregating information from covariates, but for a uniform distribution case the outliers are smoothed such that forecast values yield less extreme values which may reduce the forecast performance. DeepAR is beneficial because of its hyperparameter for user defined scale factor. Hence, Salinas et al. (2020) propose a heuristic approach, which works well in practice, of scaling each of the autoregressive input $z_{i,t}$ at each LSTM cell by dividing it on average values $v_i = 1 + \frac{1}{t_0}\sum_{t=1}^{t_0} z_{i,t}$. Meaning that a high (low) mean increases (decreases) the probability of sampling outliers.

The model architecture can control for how long in the past the network can see and how long in to the future it is going to predict. It is suggested to use over hundred related time series with at least 300 observations across all training time series in order to outperform traditional models such as ARIMA and ETS ("DeepAR Forecasting Algorithm", n.d.). Which is promising as we already have a large number covariates and fulfilling the least observation requirement.

6.3.5 Temporal Fusion Transformers

A relatively recent development in sequential neural networks is *Temporal Fusion Transformers (TFT)* (Lim et al., 2021). Temporal fusion transformers is a network of networks comprised of encoder-decoder LSTM networks, gating networks and more. In contrast to most neural network architecture handling sequential data, the TFT architecture distinguish between static and unknown variables by use of static covariate encoders (Lim et al., 2021). Examples of static variables may include days of the week or seasonal binary variables and other time-dependent variables which can be known at forehand. Static variables may increase or decrease the importance of temporal features, and help in variable selection.

In recurrent neural networks, and its variants, the significance of past observation may vary according to order. Transformer networks employ self-attention to rank the significance of parts of the sequential data used in the encoders. The encoders map a sequence of data into a vector format where the significance or weighting of the input data can be weighted by self-attention (Bahdanau et al., 2016). Self-attention functions maps a set of queries and a key-value pair to an output, where the output is a weighted combination of values. In the context of forecasting, transformer networks have used self-attention to highlight certain parts of a sequential data (Li et al., 2019).

Attention-based transformer networks constitute a valuable part of the TFT architecture, interpretability of model features, a key property which less complex LSTM and RNN architectures lack. Being able to identify the most important aspects of input data offers value for the end-users of the model, as well as allowing the TFT networks to weight input variables according to their significance at each time-step and reducing the impact of any noisy variable (Lim et al., 2021).

As a whole Temporal Fusion Transformers are complex and a highly specialized forecasting architecture, but offers great promise in terms of performance, outperforming both traditional forecasting methods ARIMA and ETS, but also complex deep learning architectures such as DeepAR (Arik and Pfister, n.d.). Of particular interest to this work, is the architecture's ability to distinguish between static and time-dependent variables.

6.4 Probabilistic neural network classifier

In the simplest sense, the likelihood of day-ahead prices exceeding intraday price for a given delivery hour can be deduced by subtracting intraday from day-ahead forecasts. However, this approach potentially neglects the inherent uncertainty in forecasts. Small changes in the level of intraday and day-ahead forecasts may result in exaggerated classifications of price differences. As an example, imagine if the day-ahead forecasts systematically overestimates day-ahead prices, i.e an expected positive forecast error. If intraday forecasts have a negative forecast error, the resulting predicted probabilities will be biased and skewed in favor of predicting a positive day-ahead intraday price difference. By training a neural network classifier we seek to close the gap between predicted and true probability of day-ahead prices exceeding intraday prices.

Using a probit model Maciejowska et al. (2019) calculates the probabilities of day-ahead discount, i.e intraday price exceeding day-ahead price. We will present as neural network classifier as an alternative to a benchmark logit model.

6.5 Training Neural Network models

In training and validating neural networks a large number of hyperparameters needs to be carefully tuned. The sheer size of the search space makes it at best a challenging process. In practice neural network training often involves narrowing the search space by starting off with the impactful hyperaparameters. We will only mention them briefly, and refer the reader to the appendix section for a more in-detail overview.

6.5.1 Learning rate

The learning rate γ controls how much the neural network model should respond to the estimated error when weights are updated (Brownlee, 2020). A high learning rate typically leads to faster learning, but risks missing local optimum. On the other hand, a small learning rate increases convergence time and increases the risk of the model getting stuck in a local optimum. Altering the learning rate while training through *learning rate schedule* can smooth descent when approaching a local minima. Typically this involves reducing the learning rate after a training condition has been met, for instance having trained on all data n times, i.e n epochs.

6.5.2 Hidden layers and neurons

Altering the depth and width of a neural network can have dramatic and immediate effect on model characteristics. While a general preference for deep networks can be seen in existing literature, the learned representations of shallow and deep networks can often be similar. Nguyen et al. (2021) looks at the difference in predictions made by shallow and deep networks finding that on the whole the predictions had similar accuracy and characteristics.

6.5.3 Optimizers

The learning strategy applied in training neural networks is dictated by the model optimizer. We have so far briefly introduced gradient descent methods, and the use of adaptive learning rates. The preferred optimizer of the machine learning has changed in the course of years, and differs based on the research problem in question. For large neural networks adaptive optimizers are generally preferred as they have a higher chance of convergence, i.e moving towards some minima. In this thesis we have applied mainly two types of optimizer: Ranger (Wright and Demeure, 2021) and Adam (Kingma and Ba, 2014). For a brief walk-through of stochastic weight averaging and its advantages see section C.1.4 in the appendix.

6.6 Benchmark models

6.6.1 ETS

Exponential smoothing method, in contrast to the proposed neural network models, is simple but robust forecasting approach (Billah et al., 2006). The three most common and basic types of exponential smoothing models are: simple exponential smoothing provided by Brown (1959), trend-corrected exponential smoothing by Holt (2004), and Holt-Winters method by Winters (1960). The component used to create ETS forecast are level, trend and seasonal effect, combining the components by either adding or multiplying them together, and adapt the model over time when there is a structural change in the time series (Billah et al., 2006). In total there are 30 combinations of these components, and we opt to perform an exhaustive grid search in order to find the most applicable ETS model on the power market time series using nested cross-validation for hyperparameter selection in section 6.

6.6.2 SARIMA

Seasonal Autoregressive Integrated Moving Average (*SARIMA(p,d,q)(P,D,Q)s)* is a univariate model composed by four terms where autoregressive order term determines how many lagged orders of its target value should be included, integrated order determines number order of differences in order to get stationary time series, moving average order in order to mitigate short-term fluctuations, and length of the seasonality. In this case, where the power prices for both day-ahead and intraday market has hourly frequency, it is usually three types of seasonality: a daily pattern, a weekly pattern and an annual pattern (R. Hyndman and Athanasopoulos, 2018). The methodology for finding a SARIMA model is described by Wang et al. (2013):

1. Elimination of non-stationary time series is an important step in SARIMA models. One way to identify how many differencing to perform is through studying the autocorrelation function (ACF). It is generally common to use either one or two differences in order to overcome the non-stationary problem.

2. In the process of constructing a model that smooths the stationary sequence, the preliminary step is to consider autocorrelation (ACF) and partial autocorrelation (PACF) function, but is not enough in order to identify the optimal SARIMA model.

Another approach is to perform an exhaustive grid search of all possible combination of SARIMA terms in specific range for each terms, using the hyperparameter selection cross-validation approach suggested in the upcoming section 6.8. As an exhaustive grid search is computational expensive, we opt to conduct a random grid search in order to reduce the estimation time.

6.6.3 Energy Quantified forecasts and simpler forecasting methods

In addition to ETS and SARIMA simpler forecasting techniques will be considered as benchmark models for comparison-reason with the complex neural network models. We have chosen three methods: Naïve, Mean and Energy Quantifieds short-term and mid-term day-ahead price forecasts. The Naïve approach is simply taking the last observed value in a train period and extrapolating the forecast period using this value. On the other hand, Mean forecast is taking the average of train period and extrapolate it on the forecast period, where one chose how many recent period to average on. We have simply taken the mean of the whole train period. The day-ahead price forecast provided by Energy Quantified are already finished forecasts. It is interesting to see if our neural network approach is able to outperform forecasting companies such as Energy Quantified. Since only day-ahead price forecast we opt to benchmark it against intraday market in addition, since intraday market has it similarities with the day-ahead market.

6.7 Performance and evaluation metrics

6.7.1 Regression metrics

Evaluating prediction performance across machine learning models, under the same circumstances, is essential to find the best performing model. Common evaluation metrics are *Root Mean Squared Error (RMSE)* and *Mean Absolute Error (MAE)* (Twomei and Smith, 1995). Willmott and Matsuura (2005) indicate that RMSE can be a misleading metric due to the function's three characteristic components of error, in contrast to MAE's simple average of error. The benefit of using MAE as performance metric is the easy interpretation of the mean error, since it is relatable to the measure unit. In other words, for the purpose of price prediction, e.g. 10 in mean absolute error is equal to 10 EUR/MWh in error on average. Hence, we opt to go further with MAE which is calculated by the average of the sum of absolute error, in other words, the deviation between predicted and observed prices presented in 9. Additionally, this metric will be used as loss function in neural networks (see section as well 6.3.1).

$$MAE = \frac{1}{N}\sum_{t=1}^{N}|\hat{y}_t - y_t| \tag{9}$$

The performance metric RMSE assigns more weight to greater prediction errors, because it squares the error before taking the average, which can be seen in 10. It is proven to be very effective in improving model performance, because of the sensitivity to large errors, which it penalizes (Chai and Draxler, 2014). But, Chai and Draxler (2014) does also point out that neither RMSE nor MAE is the best statistics metrics, because they remove a lot of information when aggregating the error to a single measure value. As the underlying assumption of RMSE is that it follows a normal distribution and that error are unbiased, Chai and Draxler (2014) state that it is a better performance metric rather than MAE, due to the fact that model errors are likely to have a normal distribution for n samples ≥ 100 rather than uniformly distributed errors which MAE is more suitable for. It is noteworthy that RMSE can never be smaller than MAE, due to its mathematical expression.

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{y}_t - y_t)^2} \tag{10}$$

Symmetric Mean Absolute Percentage Error (SMAPE) was proposed by Makridakis (1993) and is a modification of MAPE with the divisor divided by two. SMAPE, and also MAE, are scale-independent measures and Makridakis (1993) writes that scale independent measure has been a key characteristic for good measures (Kim and Kim, 2016). Because it is scale independent the metric can be used to compare results across datasets, since the unit is in percentages. However, it is possible that the divisor may approaches zero which makes the metric unstable, due to the fact that actual values can be close to zero and that the predicted value likely will approach zero as well (R. Hyndman and Athanasopoulos, 2018). R. J. Hyndman and Koehler (2006) recommends to not use this performance metric. We chose to include this performance metric, well aware of the warnings, but not using as the main metric to determine the model performance outcome. The formula is as follows:

$$SMAPE = \frac{1}{N}\sum_{t=1}^{N}\frac{|\hat{y}_t - y_t|}{(|y_t| + |\hat{y}_t|)/2} \tag{11}$$

A combination of metrics are comprehensive to assess model performance. The performance metrics will be used to evaluate both validation and test dataset. First for the critical model construction and second to justify how well a given model performs on unseen data.

6.7.2 Classification metrics

The simplest performance metric for classification problem is using the *accuracy performance metric* with the following equation:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

where T and F denotes True and False, respectively. P and N denotes Positive and Negative, respectively.

ROC stands for *receiver operating characteristic*[12] and *AUC* for *Area under the ROC curve.* Bradley (1997) concludes that ROC AUC has number of desirable properties compared to overall accuracy: that it is not dependent on a predefined decision threshold, it indicates how well the

[12] Commonly called “error curve”

negative and positive classes are separated by accounting for the distribution between of the positive to negative classes, and indicates how to the amount "work done" by a classification scheme, penalizing the score to random or "one class only" classifiers. The latter is observable for the Naïve classifier only classifying intraday participation (see figure 7.4). A model without skill is represented as 0.5, and shown as a horizontal line in the ROC AUC plot (Brownlee, 2021). The equation for ROC AUC binary classification metric is shown here:

$$\text{Sensitivity (TPR)} = \frac{TP}{TP+FN}, \quad \text{Fall out (FPR)} = \frac{FP}{FP+TN} \tag{13}$$

$$\text{ROC AUC} = \frac{1 + Sensitivity - Fall\ out}{2} \tag{14}$$

where *TPR* is short for *true positive rate* and *FPR* short for *false positive rate.*

6.8 Model selection and performance evaluation

In selecting model architecture we want to maximize the likelihood of our model performing well across in a wide range of conditions, whether it be change of seasons or exogenous shocks affecting electricity prices. As we shall see, this has multiple implications for how we evaluate our models. Our evaluation process can roughly be described as a three-step process involving:

1. Model selection: Using a nested cross-validation scheme we test a large number of models on a small but unseen validation dataset.

2. Performance: Approximate generalization error by testing a smaller subset of optimal models on a large unseen test dataset

3. Production: Two optimal models, one for each market, are used to forecast day-ahead and intraday prices. The forecasts provide the basis for calculating the probability of intraday price discount before day-ahead bid submission deadline.

Model selection and validation is performed on data ranging from March 28, 2019, to December 31, 2021, where the decision tool is put in production using a smaller subset of 2022 data. The division can be seen in figure 18.

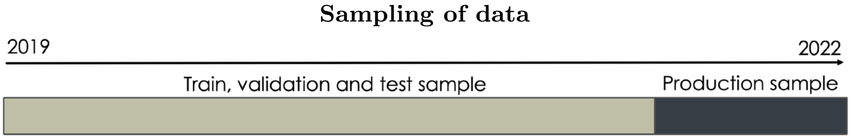


Figure 18: Figure shows how the data has been divided into a train, val, test sample and production sample

6.8.1 Model selection in time series

Validating model performance is crucial for ensuring that all models perform well on unseen data. Performing a single static separation for time series, *hold-out method*, do not capture a model performance in different period of time, introducing the problem of selecting models that only perform well on a subset of data. Evaluating model performance across multiple periods of time reduces the risk of choosing models that may overfit to certain period of time compared to using a single training, validation and test period. There are primarily two concerns which prohibit us from using a holdout method: lack of data, and non-stationarity in electricity price data. Model performance can be approximated by holdout-validation using a sufficiently large test dataset, but this would limit the available training data. Restricted to three years of data, expanding the test-set would inevitably reduce the data used to train our models. Reducing the data available for model training while increasing test-data will lower bias, but at the expense of increased variance (Raschka, 2018).

K-fold cross-validation is a more comprehensive validation technique where validation is conducted on sections of data, i.e. section of time periods (Hastie et al., 2017). Models are fitted for each training fold and the models performance are computed by averaging the performance metric of each validation fold. The purpose of performing k-fold cross validation is to increase generalizability of a model such that it performs well in different time periods. *Rolling cross validation*, as described in R. Hyndman and Athanasopoulos (2022), involves moving or *rolling* a test window across all available data as can be seen in figure 19.

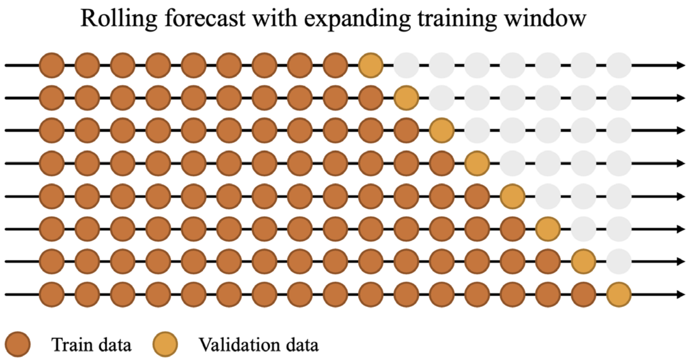


Figure 19: Illustration of rolling forecast using expanding training window.

To save complexity a electricity forecasts are often evaluated using the entirety of the test-set in one sweep (Lago et al., 2021). This is particularly problematic in our case as we seek to mimic as closely as possible the real life usage of deep learning forecasts, meaning that all available data should be utilized. As we are particularly interested in making best possible 38 hours forecasts (see section 6.1) the validation and test size for each fold should contain only 38 observations. In practice this means that the aggregate of our individual testing folds have little time dispersion. This can be mitigated by allowing for gaps between each test fold, at the expense of efficient use of available data.

Addressing the issue of non-stationarity we ideally want several well-dispersed testing folds. To balance the efficient use of data, and representative testing folds, we propose to use a nested sliding-expanding cross validation scheme. The term *nested* means that the cross-validation is done consecutively, first on the outer fold then on the inner fold of each outer fold. In detail, the cross-validation will be performed with sliding window split on the outer fold, while the inner fold is using expanding training window with rolling forecast origin. Varma and Simon (2006) find that nested cross-validation reduced the estimated error estimate. We will apply the outlined cross-validation scheme for hyperparameter-selection and model performance evaluation which can be seen in figure 20 and 21.

6.8.2 Hyperparameter selection for each architecture

In order to find the best possible hyperparameter configuration for each model architecture, i.e. LSTM, GRU, TFT and DeepAR, we conduct the *cross-validation with evaluation on a rolling forecasting origin* using 5 folds with sliding window on the outer fold, as shown in figure 20. Since the neural network models are computational expensive, the time required to perform a single cross-validated run is considerable. Hence, we opt to chose 5-fold cross validation such that more model configurations can be validated which compromises for an higher k-fold that could have lower the expected variance of our validation error estimates. In each outer fold we use time series split with expanding window and arbitrarily repeated six times with a rolling origin.

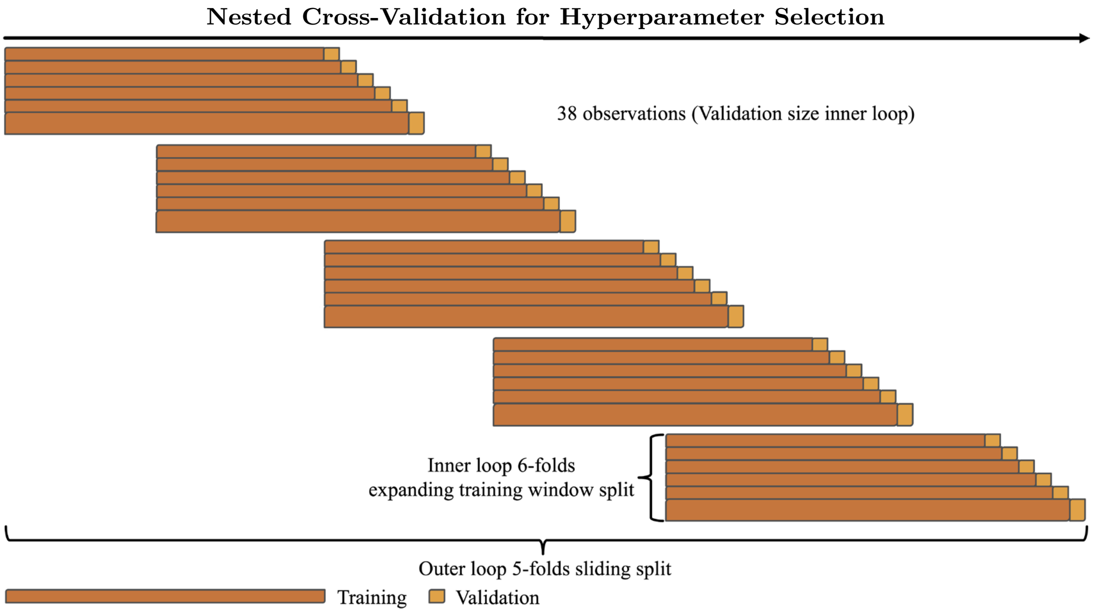

Nested Cross-Validation for Hyperparameter Selection
38 observations (Validation size inner loop)
Inner loop 6-folds
expanding training window split
Outer loop 5-folds sliding split
Training
Validation

Figure 20: Self-produced figure that shows 5-folded cross-validation of hypereparameter selection for each neural network

When the model has finished its cross-validation procedure the total performance for all folds are averaged in order to justify the hyper-parameter selection noted as, e.g. *mean of MAEs* or *mean of RMSEs.*

6.8.3 Model performance evaluation

In order to evaluate the best model configuration for each architecture, i.e. best possible chosen hyperparameters for LSTM, GRU, TFT and DeepAR, it is necessary to evaluate the models performance on a large unseen dataset. In finding optimal model configuration we used a smaller validation size as a compromise between lowered variance of the validation error and computational expense. This compromise is ill-advised when calculating the expected generalizable error of of our trained models. To ensure that the our models can be accurately compared to the benchmarks, test fold size has to be increased considerably.

Performing model training with an separated validation set, in this case 38 hours validation size, leads to a 38 hours gap between training and forecasting on the test set. Another approach is to include the validation set as a part of training data, and be well aware that the trained model already is familiar with the validation data, but this avoids the problem of 38 hours gap between training and forecast on test data. The cross-validation architecture in model performance similar to hyperparameter selection using *nested cross-validation with evaluation on a rolling forecast origin* shown in figure 21.

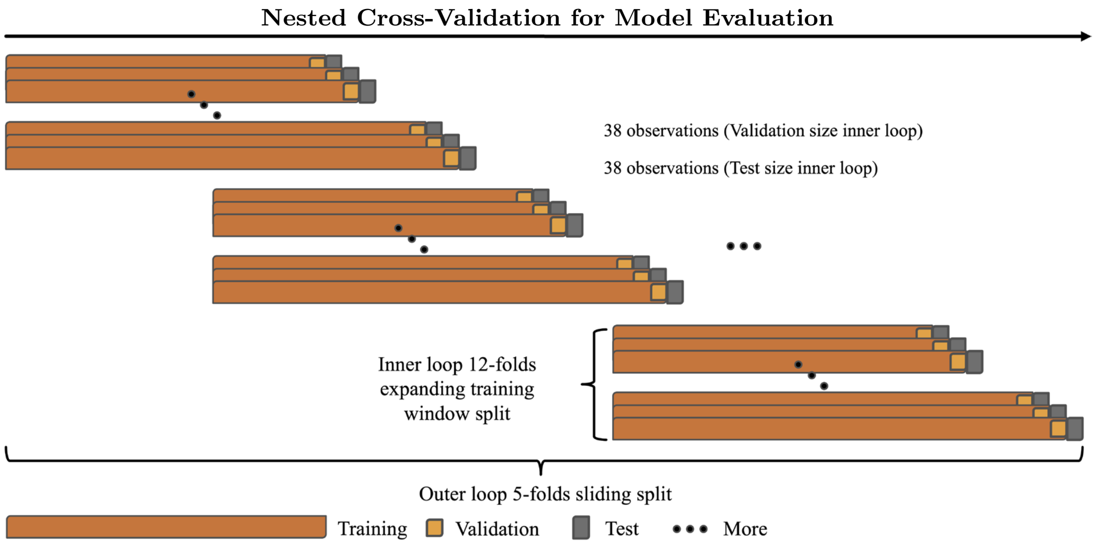

Nested Cross-Validation for Model Evaluation
38 observations (Validation size inner loop)
38 observations (Test size inner loop)
Inner loop 12-folds expanding training window split
Outer loop 5-folds sliding split
Training
Validation
Test
More

Figure 21: Self-produced figure that shows 5-fold cross-validation of model performance

The total number of aggregated test size will then be 2,280 hours (95 days) [13]. The chosen number of outer folds and inner folds may seem arbitrarily chosen, as it is to a certain extent, but it is a trade-off between estimation reliability and computational expenses.

After the refitting the model forecast 38 hours and the total performance for the whole folds test set is calculated by averaging the performance metric on the test data.

6.9 Technical implementation

In this thesis we have relied on Python as our sole programming language. The decision to stick with a single programming language simplified and stream-lined the pipeline from data collection to forecast generation. We have relied heavily on the data processing tools available in the *Pandas*, and *Numpy* Python packages. The forecasting and classification models were developed using the framework Pytorch. From the outset Tensorflow, a popular machine learning framework, was in serious consideration, but later additions to Pytorch such as Pytorch lightning and Pytorch forecasting gave Pytorch a slight edge[14].

A relatively recent development within neural network training is the use of Graphical Processing Units (GPU). GPUs excel at repeated operations, such as the matrix operations performed

[13] 5 folds outer loop x 12 rolling forecast inner loop x 38 hour forecast horizon (test size)

[14] The Python source code used to train the deep neural networks in this thesis is available at a Github repository: https://github.com/sondreid/Buy-on-Intraday-Market-or-not-A-Deep-Learning-Approach

in training neural networks (Oh and Jung, 2004). The models presented in this thesis was trained primarily using a RTX 2060 GPU with 6GB VRAM, 32GB system RAM and a AMD 3600X processor. Additional computational power was rented through Paperspace Gradient notebooks and Paperspace CORE virtual servers. The virtual servers rented generally had less computational power (Nvidia P4000), but with more VRAM (8GB) parallel runs of more than one model configuration were possible. Training models across multiple computer systems[15] is time-demanding endeavour. Running all machines using a common Linux-based operating system helped in this regard[16].

Validating and testing numerous model architectures would be useless unless the results are properly stored and logged. Neptune AI offers an excellent model logging tool which we have used extensively. Storing model configurations, validation and test results on an cloud-based tool such as Neptune has the added benefit of allowing results to be sent seamlessly from multiple devices.

[15] Configurations were run using a desktop computer, Paperspace notebooks and Paperspace virtual server in tandem

[16] Specifically, all models were run using Ubuntu-20.04 as the operating system

7 Results and Analysis

Following the model selection and evaluation procedure as outlined in section 6.8, we present benchmark and neural network architecture results. As emphasized, model selection is done on a limited validation set due to the sheer computational complexity of validating hundreds of candidate models. Selection and evaluation of neural network and non-neural network models are done using the same cross validation approach. In this way we can objectively analyse model performance, and suitability for electricity price forecasting.

Model performance can reasonably be expected to degrade according to how many hours have passed since the forecast was made. Similarly, we have seen how electricity prices in intraday and day-ahead markets are non-stationary and accordingly the optimal forecasting model may vary according to period in question. We therefore place a particular emphasis on model performance across the hours of the forecast horizon, and different folds. The best performing model for intraday and day-ahead markets will be used to make forecasts in the production period. The forecasts are passed to an multilayer feedforward network trained to calculate the likelihood of day-ahead price exceeding intraday prices, that is, an intraday discount.

7.1 Validation performance

After multiple trials of grid search and fine tuning of hyperparameter configurations and validated the architecture performance using nested cross-validation (see section 6.8.2) the final results are revealed in table 4. As aforementioned, we have chosen four different model architectures, i.e. GRU, LSTM, TFT and DeepAR, where each of these model is presented with the best possible hyperparameter combinations. In this setting, each of the architecture is targeted to forecast the both intraday and day-ahead power prices using actual data as validation. Table 6 shows some key hyperparameters, i.e. hidden size, learning rate γ, layers, dropout rate, gradient clipping scale factor and stochastic weighted average starting epoch (SWA). Additionally, the architecture with its final hyperparameter combinations are ranked with respect to the performance metric MAE shown in table 4. Besides, RMSE and SMAPE are included in order to show other performance metric to compare to MAE. The table reveals that GRU with the exact same hyperparameter combinations, both on day-ahead and intraday, is the best performing model only considered on hyperparameter selection validation. Its performance results are 7.412 and 9.441 MAE on day-ahead and intraday, respectively In other words, the neural network deviates from the actual power prices by 7.412 and 9.441 EUR/MWh on average on day-ahead and intraday, respectively.

The models and their optimal hyperparameters are tested using a small but unseen validation set. As important modelling techniques such as learning rate schedules are dependent on validation loss, the validation metrics are likely biased. An more exhaustive evaluation is needed to select the best forecasting model for intraday and day-ahead markets, respectively (see section 7.3).

Hyperparameter Selection of Neural Network Models using Nested Cross-Validation

Model	Target	Mean validation SMAPE	Mean validation MAE	Mean validation RMSE
GRU	Day-ahead	0.131	7.412	10.218
TFT	Day-ahead	0.136	8.023	10.744
LSTM	Day-ahead	0.160	9.245	11.865
DeepAR	Day-ahead	0.254	14.761	17.296
GRU	Intraday	0.167	9.441	11.866
TFT	Intraday	0.189	9.875	12.486
LSTM	Intraday	0.201	10.121	14.231
DeepAR	Intraday	0.223	11.362	14.244

Table 4: Mean validation results of the best hyperparameter selection validated through nested cross-validation for each architecture on day-ahead and intraday price as target.

Using the exact same cross-validation approach as with neural network models, one can observe the following results of the best possible combination for each benchmark forecast models in table 5. Considering figure 11 (see section 5.5.2) it is reasonable that day-ahead and intraday price time series are non-stationary yielding first order differencing in the ARIMA models. The main difference between the ARIMA model with target on day-ahead power prices and intraday power prices are 2nd and 4th order moving average-term and use of fifth and second order autoregressiv-term, meaning that it regresses on its own previous values. It is noteworthy, that non of the optimal ARIMA models using seasonality. The best possible hyperparameter configuration for ARIMA models have the lowest MAE among the other selected benchmark models, with 8.028 and 11.167 MAE for day-ahead and intraday power price forecast on average, respectively. As aforementioned, comparing each architecture this way may be biased, and a more correct comparison comes in the next section.

Just behind the ARIMA models, one can observe that Energy Quantified's short-term day-ahead forecasts performs relative poorly compared to ARIMA with approximately 1 MAE in deviance both on intraday and day-ahead. It is also noteworthy that the Naïve model with day-ahead price as target outperforms the ETS model. The best possible parameter configuration of ETS on day-ahead price target is additive, additive damped and additive, error, trend and seasonal component, respectively. The ETS on intraday price target does only contain an additive error term.

Hyperparameter Selection of Benchmark Models using Nested Cross-Validation

Model	Target	Mean validation SMAPE	Mean validation MAE	Mean validation RMSE
SARIMA (5,1,5)(0,0,0)0	Day-ahead	0.133	8.028	10.836
EQ short term day-ahead benchmark	Day-ahead	0.147	9.311	12.076
Naïve	Day-ahead	0.180	10.015	12.760
ETS (A,Ad,A) season=168 (1 week)	Day-ahead	0.196	10.435	12.984
Mean	Day-ahead	0.777	37.022	38.321
SARIMA (2,1,4)(0,0,0)0	Intraday	0.197	11.167	14.269
EQ short term day-ahead benchmark	Intraday	0.200	12.291	14.860
ETS (A,N,N)	Intraday	22.449	12.353	19.858
Naïve	Intraday	0.266	14.069	16.811
Mean	Intraday	0.797	36.316	38.036

Table 5: Mean validation results of the best hyperparameter selection validated through nested cross-validation for each architecture on day-ahead and intraday price as target.

Comparing the GRU's MAE results with the benchmark models in table 5, it outperforms the best ARIMA benchmark model with 0.904 and 1.726 EUR/MWh on average on day-ahead and intraday, respectively.

7.2 Final Model Architectures

The end result of the model selection is narrowing down to a single model architecture for each family of models. For the sake of brevity, we will only present the hyperparameters determined to be the most influential for model performance. For a full overview of the model hyperparameters in this section, see E.3.1 in the appendix. Some configuration is worth noticing from table 6 is that the neural networks use hidden size in a range between 64 and 256 and relative low learning rate γ below 0.001, except for TFT which uses a 0.05 learning rate. The recurrent neural layers used in this architectures are between 2 and 4. Furthermore, the neural networks seem to perform well using 5 and 10 % dropout values, a modest degree of regularization. Using gradient clipping value of 30 % to 60 % stabilizes training, and improves model performance.

Optimal Hyperparameter Configuration of each Neural Network

Model	Target	Hidden size	Learning rate	Layers	Dropout	Gradient clipping	SWA starting epoch
GRU	Day-ahead	256	0.00100	2	0.100	0.600	12
TFT	Day-ahead	128	0.05000	3	0.050	0.300	10
LSTM	Day-ahead	128	0.00010	2	0.100	0.500	15
DeepAR	Day-ahead	64	0.00100	2	0.100	0.600	15
GRU	Intraday	256	0.00100	2	0.100	0.600	12
TFT	Intraday	64	0.05000	3	0.050	0.300	10
LSTM	Intraday	128	0.00056	2	0.100	0.005	12
DeepAR	Intraday	128	0.00010	2	0.100	0.050	13

Table 6: Optimal Hyperparameter Configuration of each Neural Network.

For all model architectures we found that averaging the predictions of the n best performing model states improved performance[17].

[17] This in-expensive ensembling technique is called checkpoint ensembling (H. Chen et al., 2017)

Switching optimizer to stochastic weight averaging (SWA) mid-training, increased out-of-sample performance considerably. SWA was generally switched to after 10 to 15 epochs worth of training. For a brief introduction of SWA and checkpoint-ensembling, we refer the reader to C.1.4 and C.1.1 in the appendix.

7.3 Performance evaluation

As described in the methodology section (see section 6.8.3) we want to fairly compare each model against each other by performing a new nested cross-validation using larger unseen test data folds. Notice that validation fold shares the same data records the last 38 records in the last part of each training fold to avoid a temporal gap between train and test period.

Table 7 shows the nested cross-validation performance results for LSTM, GRU and DeepAR for intraday and day-ahead forecasts. The results reveals that LSTM outperforms other models in terms of MAE by a significant margin, averaging a 9.484 EUR/MWh discrepancy between forecasts and observed prices. Of particular note is the differing relative performance of the deep neural architectures. In intraday markets the gated recurrent unit model performs best in terms of MAE with an average error of 12.118 EUR/MWh. The order of relative performance is the same for RMSE and MAE, while SMAPE suggests that DeepAR performs better than GRU. This may be a sign of lower relative forecast variance.

Neural network performance using Nested Cross-Validation[a]

Model	Target	Mean test SMAPE	Mean test MAE	Mean test RMSE	Mean train SMAPE	Mean train MAE	Mean train RMSE
LSTM	Day-ahead	0.193	9.484	12.383	0.156	2.118	2.960
GRU	Day-ahead	0.251	11.639	15.275	0.510	5.926	7.168
DeepAR	Day-ahead	0.264	14.832	18.434	0.300	4.128	5.187
GRU	Intraday	0.303	12.118	15.853	0.581	5.927	7.172
DeepAR	Intraday	0.313	12.227	15.653	0.292	3.119	4.049
LSTM	Intraday	0.351	12.553	16.046	0.317	2.776	3.674

Table 7: Model performance using Nested Cross-Validation

[a] The performance results of TFT is not included in the further results and analysis (see section 8.3 for rationale)

For the benchmark models in table one observes that ETS(A,Ad,A) with 1 week seasons (168) is the best performing model both for day-ahead and ARIMA on intraday price target. The average test MAE for day-ahead and intraday on GRU and ARIMA price forecast are 13.667 and 17.067, respectively. It is noteworthy, that the Naïve model on day-ahead price target is almost as good as GRU on average only deviating results by 0.238 on MAE. Considering the test SMAPE performance metric, the Naïve model actually outperforms all the benchmark models on day-ahead on average and benchmark Energy Quantified's short-term day-ahead forecast on intraday.

Benchmark performance using Nested Cross-Validation

Model	Target	Mean test SMAPE	Mean test MAE	Mean test RMSE	Mean train SMAPE	Mean train MAE	Mean train RMSE
ETS(A,Ad,A)168	Day-ahead	0.230	13.667	17.434	0.138	1.151	2.916
ARIMA(5,1,5)	Day-ahead	0.218	13.709	17.749	0.045	0.943	2.861
Naïve	Day-ahead	0.200	13.898	17.823			
EQSTDF[1]	Day-ahead	0.279	14.755	18.795			
Mean	Day-ahead	0.980	46.549	48.514			
ARIMA(2,1,4)	Intraday	0.404	17.067	21.187	0.193	1.717	3.608
EQSTDF[1]	Intraday	0.401	17.103	21.320			
ETS(A,N,N)	Intraday	0.429	19.616	23.550	0.180	1.695	3.713
Naïve	Intraday	0.429	19.622	23.556			
Mean	Intraday	1.002	42.538	45.138			

Table 8: Model performance using Nested Cross-Validation. (1) EQSTDF = Short-term Day-ahead forecast

Comparing the nested cross-validated neural network LSTM (day-ahead) and GRU (intraday) in table 7 against the best performing benchmark models ETS and ARIMA in table 8 it significantly shows that the neural network models outperforms the benchmark models by 30.6 % measured on test MAE and by 29 % on day-ahead and intraday on average, respectively. The neural network models outperforms Energy Quantified's forecasts, an example of which can be seen in 22. This particular fold is chosen because it highlights the stability of deep neural network forecasts. [18]

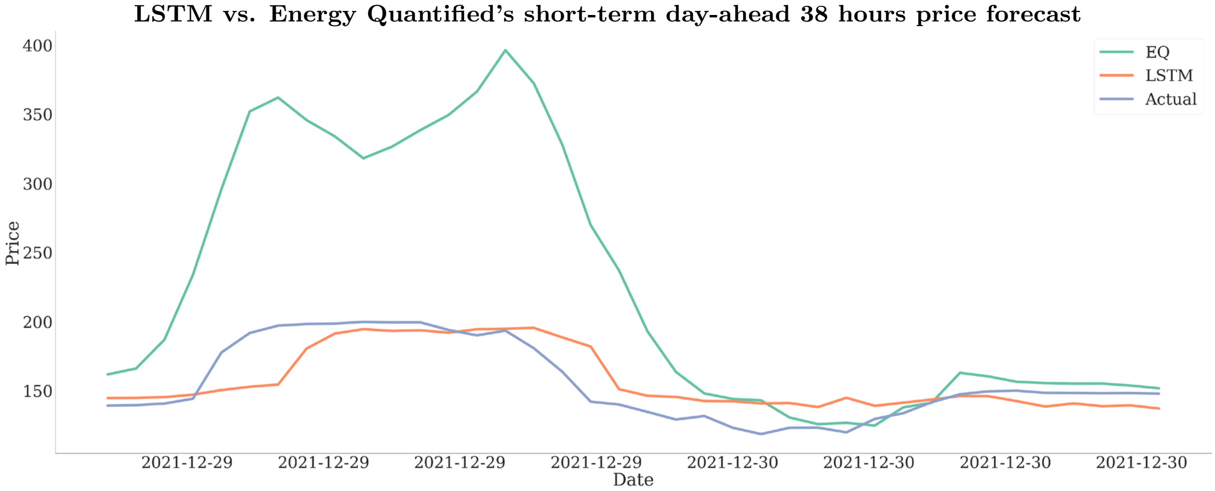


Figure 22: LSTM vs. EQ Day-ahead

As observed in figure 11 (see section 5.5.2) that the price evolution on intraday and day-ahead market NO2 are relative stationary from 2019 up until 2021, when the series become non-stationary and more volatile. This can also be observed in figure 23 and 24 which present the MAE for each nested cross-validate fold on day-ahead and intraday. Fold 1 to 4 have relative

[18] A compilation of all evaluation forecasts is available in the "forecasts" directory of the github repository: https://github.com/sondreid/Buy-on-Intraday-Market-or-not-A-Deep-Learning-Approach

low MAE in contrast to the last 5th fold where the MAE for the best neural network and four of the top benchmarks MAE increase significantly. 23 and 24 highlights the relative performance-premium of deep neural networks in periods high price variance.

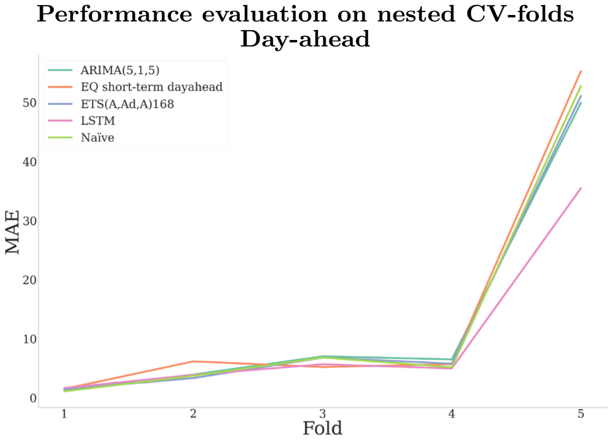


Figure 23: MAE per cross validation fold in day-ahead markets

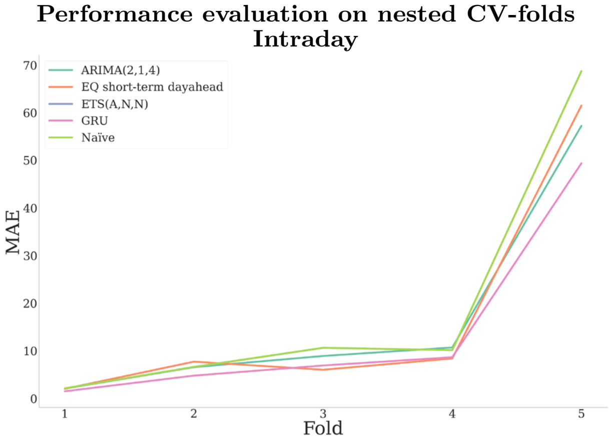


Figure 24: MAE per cross validation fold in intraday markets

It is also interesting to analyse how well the neural network models perform across the forecast horizon. Figure 25 and 26 shows the average MAE achieved by benchmarks and deep neural networks per outer fold. It reveals that LSTM and GRU outperforms the best benchmark models almost exclusively.

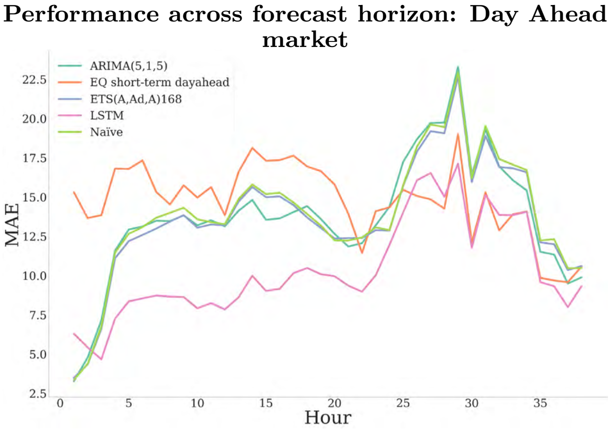


Figure 25: MAE per hour from forecast origin in day ahead market for all cross-validated folds

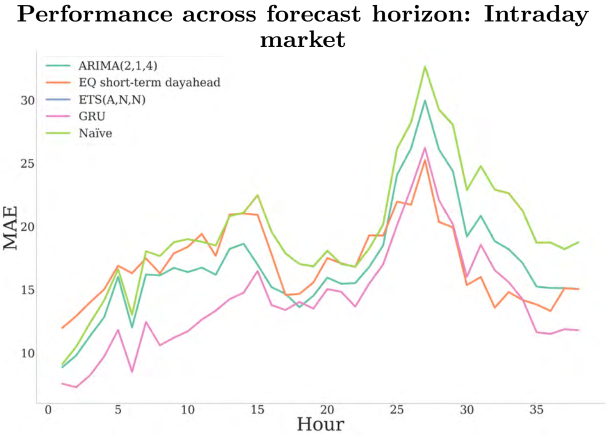


Figure 26: MAE per hour from forecast origin in intraday market for all cross-validated folds

The total performance across delivery hours for the nested performance validation is presented in figure 27 and 28 for day-head and intraday, respectively. Calculating the test MAE for each hours the plots showcase that LSTM and GRU almost always beating the best performing benchmark models ETS and ARIMA, except from hour 16 and 17 on intraday.

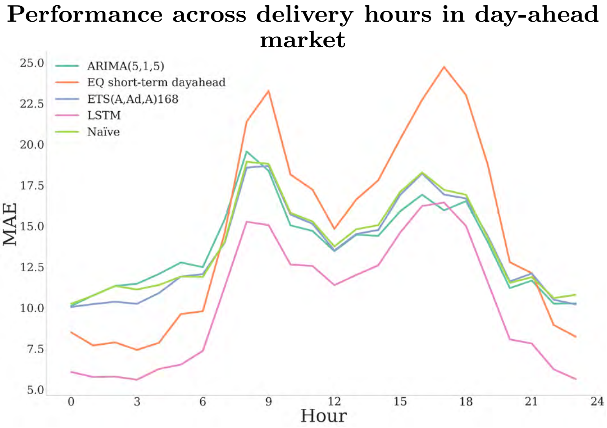


Figure 27: Performance evaluation across hour on Day-ahead for all cross-validated folds

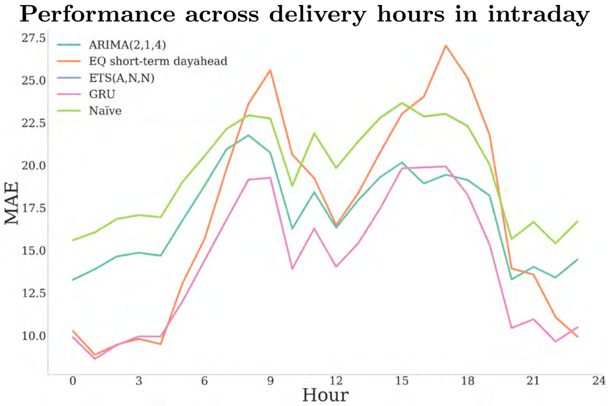


Figure 28: Performance evaluation across hour on Intraday for all cross-validated folds

Summarized, the neural networks, i.e. LSTM on day-ahead and GRU on intraday, are the best performing models as determined by nested cross-validation. The forecasts produced by respective best-performing networks will be used in a simulated production environment to showcase the potential use-case of the decision tool.

7.4 Neural network classifiers

We will compare the performance of our classifier networks to two benchmarks: A majority class prediction, and the sign-difference of intraday and day-ahead forecasts in the production period. In table 9 the Naïve model that always predict that day-ahead price always exceeds intraday and LSTM neural network yield the highest accuracy of 67.9 %.

Performance of probabilistic classification

Model	Test ROC AUC	Test Accuracy
MLP model	0.598	0.671
Logit model	0.560	0.613
LSTM	0.504	0.679
Naïve ($Prob(\Delta\hat{P}_{h,d}) = 1$)	0.500	0.679

Table 9: Performance of probabilistic classification models

On the other hand, measuring model performance using ROC the MLP model outperforms both the simple benchmark models and LSTM network with a measure of 59.8 %. It is noteworthy that the simple sigmoid function (which is a logit model) squishing the exceeding price difference between day-ahead and intraday directly, is quite close to MLP only deviating 3.8 %-points.

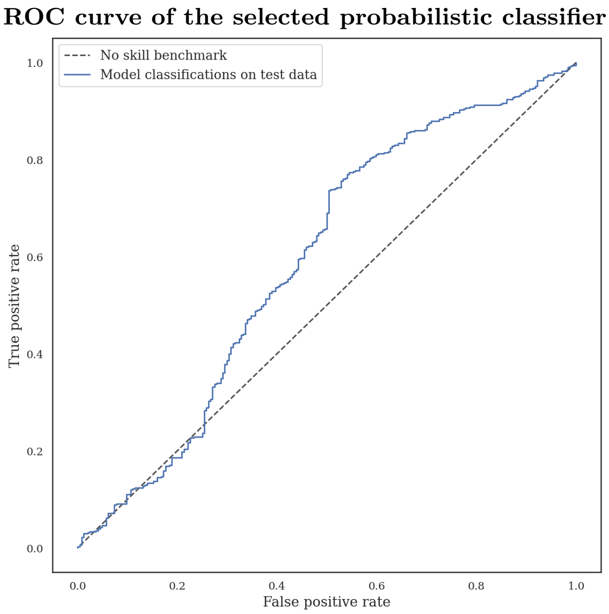


Figure 29: ROC curve of probalistic classifer

Figure 29 shows the ROC curve of the classifier. In low threshold outcomes we see that our model performs worse than a no-skill benchmark, seen in the bottom-left section of the plot. This suggests that our model generally underestimates the likelihood of a day-ahead discount.

7.5 Production

We apply the best performing LSTM model and GRU model to forecast day-ahead and intraday prices, in a simplified production environment. The forecasts are made before the day-ahead market submission deadline, at 12 p.m. The forecast horizon spans to midnight the following day, a total of 38 hours (see section 6.1). The probabilistic classifier outlined in previous section is used to generate probabilities of day-ahead price exceeding intraday prices for a given delivery hour. Depending on the risk profile of a potential buyer μ, a decision can be made on whether to buy the entirety of the bid volume in intraday or day-ahead markets. We opt to simulate a real life production scenario from January 2, 2022, to January 31, 2022, as presented in table 10.

Table 10 shows the actual profit (see section 6.2) based on a given probability threshold μ between 0 % and 100 %. Studying the total performance for the production month one can observe that a power buyer choosing to have a risk probability threshold of 50 % is gaining 6,414.3 Euro for its participation on intraday compared to a day-ahead strategy. For a very risk-averse electricity buyer at a μ of 95 %, the total profit is reduced to 1,435.1 Euro.

Daily profits of the decision tool in production January 2022

Day/μ	0/10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	100%
2	513.3	513.3	513.3	513.3	486.5	293.5	273.0	273.0	209.9	−3.9	0
3	756.8	756.8	756.8	756.8	756.8	756.8	724.3	711.6	658.0	606.0	0
4	136.3	136.3	136.3	129.7	139.1	196.0	−22.1	−22.1	−22.1	−22.1	0
5	24.7	24.7	24.7	24.7	24.7	24.7	24.7	26.8	83.6	62.0	0
6	347.2	347.2	347.2	347.2	347.2	348.7	263.6	148.7	106.4	88.6	0
7	557.9	557.9	557.9	557.9	557.9	306.1	192.1	133.8	79.4	51.0	0
8	209.9	209.9	209.9	209.9	209.9	207.4	201.7	181.6	163.5	93.0	0
9	0.3	0.3	0.3	0.3	0.3	0.3	0.3	−0.8	21.0	40.7	0
10	−85.0	−85.0	−85.0	−85.0	−85.0	−85.0	−85.0	−84.9	−105.3	−111.1	0
11	60.7	60.7	60.7	60.7	60.7	64.8	23.8	−13.2	−18.9	−18.9	0
12	598.8	598.8	598.8	598.8	598.8	598.8	598.8	598.8	320.6	30.0	0
13	632.1	632.1	632.1	632.1	632.1	15.0	16.7	0	0	0	0
14	216.5	216.5	216.5	216.5	216.5	165.5	2.0	0	0	0	0
15	79.4	79.4	79.4	79.4	79.4	−32.2	0	0	0	0	0
16	294.9	294.9	294.9	294.9	294.9	294.9	298.5	233.7	246.6	151.9	0
17	−59.3	−59.3	−59.3	−59.3	−59.3	−59.3	−59.3	−101.8	0	0	0
18	69.3	66.8	66.8	31.7	31.7	20.4	43.0	55.7	23.5	3.3	0
19	276.1	276.1	276.1	276.1	276.1	276.1	276.1	269.5	160.6	−9.7	0
20	387.7	387.7	387.7	387.7	387.7	184.0	0	0	0	0	0
21	88.0	88.0	88.0	88.0	88.0	88.0	88.0	85.7	7.5	0	0
22	103.9	103.9	103.9	103.9	103.9	50.2	11.3	2.4	0	0	0
23	240.1	240.1	240.1	240.1	240.1	216.9	57.7	21.6	0	0	0
24	137.2	137.2	137.2	137.2	137.2	127.5	94.0	75.4	27.9	0	0
26	144.0	144.0	144.0	144.0	144.0	140.5	112.4	13.5	0	0	0
27	254.0	254.0	252.0	238.2	227.0	158.7	154.0	117.8	130.7	88.7	−0.8
28	314.6	314.5	314.8	315.6	303.5	311.4	336.2	335.1	384.9	384.9	201.1
29	794.6	794.6	794.6	794.6	794.6	794.6	794.6	715.6	110.1	0	0
30	248.0	248.0	248.0	248.0	248.0	248.0	248.0	249.2	60.2	0.2	0
31	−1047.6	−1047.6	−1047.6	−1047.6	−828.0	−331.9	−104.7	0.4	0.4	0.4	0
Total	**6294.3**	**6291.7**	**6290.1**	**6235.4**	**6414.3**	**5380.4**	**4563.7**	**4026.8**	**2648.7**	**1435.1**	**200.4**

Table 10: Daily profits given a probability threshold μ

A buyer willing to take on more risk has a higher expected net profit, while total profits decrease with lower risk tolerance. The decrease in risk is evident from the variation in daily profits, where a lower threshold μ gives higher fluctuations in daily profits.

It should again be emphasized that the validity of the results presented depend on the assumptions outlined in section 6.1. The assumption of a transaction cost in particular are likely to skew the results in favor of intraday trade. This can be seen in comparing the net marginal profits of an exclusive intraday trade strategy when μ is 0 to 10 %, and a near-exclusive day-ahead strategy when μ is 100 %. Furthermore, a 30-day production period is too small a sample for generalizable results, and should be read as a proof-of-concept.

8 Discussion

8.1 Main findings

A central working hypothesis of this work has been the potential under-utilisation of intraday electricity markets. When left with a choice to participate in intraday-markets or balancing markets, wind-traders often ignore intraday markets (Mauritzen, 2015). Similar findings by Scharff and Amelin (2016), lends some credence that intraday markets remain underutilised. A central point of discussion in this thesis is the potential gains of an electricity buyer by considering intraday trading. To explore this possibility, we have used deep learning to make multi-day forecasts before the day-ahead market submission deadline. Our goal has been to determine whether forecasts provided by deep neural architectures can provide sufficient information for a market participant to choose its market of choice. The decision tool, a probabilistic neural classifier, provides probabilities of an intraday price discount for a given delivery hour. The classifier, does to some degree, rectify the error in day-ahead and intraday forecasts, leading to somewhat better performance compared to using a simple logit model. The tested classifiers generate point-point classifications, and the inclusion of sequence-sequence classifiers might have improved classifications, as done in Sutskever et al. (2014) for translation. The use of deep learning forecasts and classifications complements previous work done by Maciejowska et al. (2019), using an ARX forecasting model and a probit classification model. Although, a direct-comparison of results is difficult as the underlying electricity markets differ.

We have placed a special emphasis on thorough model validation and performance review. Through a nested cross-validation scheme, we ensure that the unseen test data is drawn from a data sample that is as independent and temporally spread as possible. The added complexity this entails reduces the number of model configurations we are able to test. The benefit of more trustworthy results, outweighs the complexity and tediousness of a more thorough cross-validation scheme.

In a simplified production environment, we have demonstrated the potential use-case of our decision tool. By participating in intraday trade, an electricity buyer in our production period can expect marginal net profits to increase by 6,414.3 Euro when compared to exclusively trading in the day-ahead market. We find that given a high risk tolerance expected profits as well as uncertainty increases. This is in line with with previous work from Maciejowska et al. (2019). Our simulated production use-case should be read with an ample degree of scepticism. Due to our limited dataset, and computational concerns, the production period is limited to only 29 days, comprising 696 hours. Drawing any far-reaching conclusions from such a limited sample is problematic.

In addition, the assumptions made in our production environment are likely not valid given large buy volumes. Specifically, in lack of any clear method of estimating transaction cost we have assumed that the ex-post observed intraday price is the price ultimately paid by a producer. However, a large buy volume added to a relatively illiquid intraday-market will likely have a net-positive effect on intraday prices. As seen in Weber (2010) any influence a bidder has one the intraday price can be viewed as a major transaction cost. This may cause an overestimation of the potential gains of intraday trading when compare to a purely day-ahead-based trading strategy. Similarly, regulation prices will likely be affected by an increase in consumption imbalance. In the event of an existing negative system imbalance, added buy volume will likely increase upward regulation prices, reducing the potential gains compared to day-ahead trade. As a result, we view it as unlikely that the production results can be viewed as anything more than a proof-of-concept.

8.2 Suitability of Neural Networks in Electricity price forecast

As we have seen, deep learning architectures outperform all tested benchmark models for a wide selection of test data. The average MAE in the evaluation period was 62.11 EUR/MWh in intraday markets and 61.04 EUR/MWh in day-ahead markets. The average discrepancy between forecasts and observed prices for LSTM (day-ahead) and GRU (intraday) were 9.48 and 12.12 EUR/MWh. In comparison the best performing benchmarks achieved mean absolutes errors of 13.67 and 17.07 EUR/MWh in day ahead and intraday markets. In percentage terms this constitutes a performance premium of 30.6 % and 29 % on day-ahead and intraday markets, respectively.

Significantly, deep learning techniques outperform Energy Quantified's preferred short-term model ("Spot price model improvements", 2022)[19]. The benchmarks presented in this thesis does not represent an exhaustive suite of forecasting models, and the inclusion of other benchmark models may have altered the relative performance of the neural networks. Regardless of the potential shortcomings of the benchmark models presented in this thesis, the efficacy of deep neural networks for forecasting has been shown in existing literature. Comparing the performance of our deep neural networks directly is problematic as no previous work exists for intraday and day-ahead forecasts in the NO2 price region to our knowledge. Nonetheless, exist-

[19] The Energy Quantified forecasts represent the latest iteration of forecasting methods per January 25, 2022

ing literature for other European electricity markets such as Beigaitė et al. (2018) and Lago et al. (2021) identify deep learning architectures as their respective best performing class of models.

We find that the performance premium of deep neural networks are particularly noteworthy when in periods of high intraday and day-ahead price volatility. Our findings are in line with previous work such as Polson and Sokolov (2019) who found that deep learning forecasting techniques were particularly effective for periods with high price volatility. With some confidence we can say that the performance of our deep neural networks is credible, and supports existing literature.

There are however some major caveats to using deep neural networks for forecasting. The discrepancy in performance between the the first tested model, and the final models presented in this thesis is such that model performance is often paid for with time. In comparison, a well-performing ETS model can be validated in a fraction of time while achieving similar performance on some sections of the test data. A refit on available training data takes *neural network training time* for best performing LSTM network while the day-ahead ETS model uses *ETS training time.* The increase in model performance achieved by applying more complex models might not be a good trade-off in all cases. However, the economic benefits gained by minimizing forecast errors in day-ahead and intraday markets might justify time spent.

8.3 The anatomy of tested neural networks

For the purposes of maximizing forecast performance, a large number of architectures have been tested. Generally, our initial design choices have been based on previous work, such as Lago et al. (2018). Aside from techniques meant to speed-up convergence such as the use of stochastic weight averaging and gradient clipping, the width and depth of the neural networks have an immediate impact on model performance. Generally the best performing model were relatively shallow, only spanning two-layers in depth. While shallow and narrow networks (as determined by the number of weights in each layer) frequently under-fitted, shallow and broad networks were generally found to outperform their deep counterpart. This contrasts recent findings in literature and what seems to have been a developing consensus in machine learning circles. Specifically, Goodfellow et al. (2016) finds that increasing model depth beyond 4 layers increases model performance.

We can only speculate as to the reason for this discrepancy between our findings and that of existing literature. One reason might be our relatively modest data-sampling period. This may be plausible as model convergence was often found to be unstable, a sign of high-variance parameter updates. Increasing the number of training observations might have altered the optimal architectures.

We were particularly interested in model architectures that has been shown to perform well as electricity price forecast models in existing literature. Temporal fusion transformers (Lim et al., 2021) seemed promising, but proved to be a challenge family of models to train. The numerous sub-networks of the TFT model such as encoder-decoder and variable-selection networks increases complexity, and reduces the potential configurations tested given limited computational resources. A pragmatic approach would be to reduce the number of neurons and layers, but this in turn lead to under-fitting and more or less static forecasts in line with the mean of the preceding electricity prices. The end result were a set of under-performing model, found to be unworthy of further time-investment. As a consequence, TFT models were withheld from performance evaluation. It is possible, but not likely, that testing TFT models on a larger test during performance evaluation would have altered the relative order of the best performing models.

8.4 Relevance of decision tool: Barriers and Possibilities

The decision tool described in this thesis may enable electricity buyers to take an informed decision on whether or not to participate in intraday trading. Despite its severe limitations, narrowing down the market-participation choice down to a binary choice offers several benefits. For one, it allows for automation of intraday trade, a possibility that according to Scharff and Amelin (2016) is rarely used today[20]. This is particularly relevant as intraday trade is expected to increase in tandem with the share of variable renewable energy sources (vRES) and particularly wind power Mauritzen (2013), plausibly making intraday trade more difficult to monitor without the aid of automated systems.

The application of deep neural networks comes with a significant added complexity, for practitioners and the machines required to train and validate countless iterations. Neural network frameworks typically have a higher barrier-of-entry in terms of programming skills than other machine learning methods, examples including Pytorch and Tensorflow. We would argue that the skills needed to implement deep neural networks are non-negligible, but does not present a barrier for wide-spread adoption. The application of deep neural networks comes with a considerable computational premium. However, it is reasonable to assume that the minimum-required computational resources are available to a potential user of the decision tool outlined in this thesis. We find this to be particularly plausible as the training of forecast and classification models presented in this thesis was done mostly on a modestly powerful desktop computer and

[20] Specifically, Scharff and Amelin (2016) looks at observed trading patterns and suggests the patterns are in line with manual trading

virtual servers.

8.5 Weaknesses

In this thesis we rely on linear interpolation of non-traded hours and potentially missing data. We have discussed that there exist a multitude of methods to interpolate the missing records, and that linear interpolation is a simple method to overcome this problem (see section 5.4.1). It is a stark assumption to assume that two observed neighboring values are linear. It is however difficult to ascertain the potential faultiness of linear interpolation as no obvious test to judge it suitability exists. Creating values for non-existing or missing values has potentially severe impact on the prediction, considering that 33 % of the intraday data are non-traded on or missing (see figure 14). Additionally, the ticker bids on intraday market not only occur for a single period, may last for several hours, referred to as block order (see section 5.1). We have chosen to extrapolate the price and volume for given block orders represented as several hours bids, for the purpose of calculating the hourly volume weighted price. We highlight this as a potential weakness, because large order may affect the hourly price calculation, and the fact that block orders can be partially accepted during its lifetime.

By including meteorological, production, and capacity-related forecasts we can potentially enrich our models. There are however, some plausible caveats to using forecasts as covariates. Each forecast is a result of an underlying forecast model, inevitably including simplifying assumptions. The sum of numerous forecasts errors ϵ_i may have degraded, rather than improved our forecasts.

Additionally, we opt to only include 2 days ahead old forecasts to ensure that only data that would be available to a prospective electricity buyer is used. This is crucial in validation and evaluation where forecasts are made on a rolling origin. However, in a production environment all forecasts made before 10 .a.m should be considered. As such we have potentially left out valuable information present in newer forecasts, potentially degrading model performance. The Energy Quantified short-term day-ahead forecasts is also penalized by forecast restriction, and may have been an unrealistic benchmark in the validation of models.

The validation of our models has been conducted using non-overlapping 38 hour forecasts, using all available data. In order to strictly use data that would be available in a real-life use-case, Energy Quantified forecasts published within the forecast horizon have been intentionally left out. Another viable approach could have been to validate the models by performing 38 hours forecast each day at 10 a.m. such that we could make use of Energy Quantified's newest forecast before issued at 10 a.m. each day. In practice this would mean leaving out large portions of available data, a trade-off we deemed unattractive.

The use of two different neural networks in order to predict the probability of day-ahead exceeding intraday price classification may have been an unnecessary complication. Using the actual price exceedance $\Delta P = P^{\text{Day-ahead}} - P^{\text{Intraday}}$ directly as dependent variable may reduce the modeling complexity. Early attempts to forecast using the day-ahead exceedance directly suggested that predictions improved by modelling the intraday and day-ahead prices in separate neural networks. The discrepancy in relative performance of the same model architecture applied to intraday and day-ahead markets may lend some support to this hypothesis, as seen in table 4. This suggests that the optimal intraday and day-ahead forecasting models differ depending on whether they are applied to day-ahead or intraday markets. Additionally, predicting prices are beneficial as it is allows for easier comparisons with existing literature and day-ahead forecasts made by Energy Quantified.

8.6 Avenues of further research

In this thesis we have made several major simplifications which may have affected our results. Extending trading behaviour by allowing for mixed day-ahead and intraday positions may alter the conclusions drawn in this thesis.

The decision to trade in intraday markets is largely based on the expectation that volume may be adequate for a given delivery hour. Complementary forecasts of expected volume could potentially be of great value in determining the optimal market for a given delivery hour.

We have been primarily concerned with the perspective of an electricity buyer, but dispatchable energy producers are faced with a similar decision problem as that of electricity buyers. Adapting a similar decision tool for electricity producers could be an interesting future research topic.

Preliminary findings suggests that forecasting the price exceedance of day-ahead and intraday prices directly ΔP degrades forecast performance. An exhaustive analysis is needed to conclude either way.

9 Conclusion

Deep neural networks represents a powerful forecasting tool. This thesis support the general findings in existing literature, and find that deep neural networks outperform all tested benchmarks by a significant margin. The LSTM and GRU models chosen for day-ahead and intraday markets surpasses the benchmark by 30.6 % and 29 %, respectively.

We find that the performance premium achieved by the implemented neural networks is particularly promising in time periods of high price variance. Furthermore, the forecasts produced by the neural networks are less susceptible to performance degradation across delivery hours and length of forecast horizon. As a result this thesis provides some degree of evidence that neural networks applied to the electricity price domain, increases performance and stability when compared to traditional forecasting techniques. The use of a multilayer feedforward network to classify the probabilities of an intraday discount modestly increased performance when compared to using the forecasts directly in a logit model. Though ripe with potential sources of error, our decision tool is shown to increase expected marginal profits when compared to a day-ahead-only trading strategy. Given the allevement of the error sources, the decision tool presents a promising avenue for automation of intraday trade. This is particularly relevant as wind-power and other variable intermittent energy sources make up an ever-larger share of total electricity production.

References

Articles

Ahsan, M. M., Mahmud, M. A. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, *9*(3). https://doi.org/10.3390/technologies9030052

Bahdanau, D., Cho, K., & Bengio, Y. (2016). Neural machine translation by jointly learning to align and translate. https://arxiv.org/pdf/1409.0473.pdf

Beigaitė, R., Krilavičius, T., & Man, K. (2018). Electricity price forecasting for nord pool data, 1–6. https://doi.org/10.1109/PlatCon.2018.8472762

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192–213. https://doi.org/10.1016/j.ins.2011.12.028

Billah, B., King, M., Snyder, R., & Koehler, A. (2006). Exponential smoothing model selection for forecasting. *International Journal of Forecasting*, *22*, 239–247. https://doi.org/10.1016/j.ijforecast.2005.08.002

Boucher, M.-A., Anctil, F., Perreault, L., & Tremblay, D. (2011). A comparison between ensemble and deterministic hydrological forecasts in an operational context. *Advances in Geosciences*, *29*, 85–94. https://doi.org/10.5194/adgeo-29-85-2011

Boulila, W., Driss, M., Al-Sarem, M., Saeed, F., & Krichen, M. (2021). Weight initialization techniques for deep learning algorithms in remote sensing: Recent trends and future perspectives. *abs/2102.07004*. https://doi.org/10.48550/arXiv.2102.07004

Bourry, F., & Kariniotakis, G. (2009). Strategies for wind power trading in sequential short-term electricity markets. *Proceedings European Wind Energy Conference & Exhibition (EWEC) 2009*.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Bye, T., & Hope, E. (2005). Deregulation of electricity markets—the norwegian experience. *Economic and Political Weekly*, *40*, 5269–5278. https://doi.org/10.2307/4417519

Cao, X. H., Stojkovic, I., & Obradovic, Z. (2016). A robust data scaling algorithm to improve classification accuracies in biomedical data. *BMC bioinformatics*, *17*, 359. https://doi.org/10.1186/s12859-016-1236-x

Chai, T., & Draxler, R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?– arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*, 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chen, H., Lundberg, S., & Lee, S.-I. (2017). Checkpoint ensembles: Ensemble methods from a single training process.

Chen, X., Dong, Z., Meng, K., Xu, Y., Wong, K., & Ngan, H. (2012). Electricity price forecasting with extreme learning machine and bootstrapping. *Power Systems, IEEE Transactions on*, *27*, 2055–2062. https://doi.org/10.1109/TPWRS.2012.2190627

Cho, K., Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. https://doi.org/10.3115/v1/W14-4012

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling.

Clo, S., Cataldi, A., & Zoppoli, P. (2015). The merit-order effect in the italian power market: The impact of solar and wind generation on national wholesale electricity prices. *Energy Policy*, *77*, 79–88. https://doi.org/10.1016/j.enpol.2014.11.038

Cludius, J., Hermann, H., Matthes, F. C., & Graichen, V. (2014). The merit order effect of wind and photovoltaic electricity generation in germany 2008–2016: Estimation and distributional implications. *Energy Economics*, *44*, 302–313. https://doi.org/10.1016/j.eneco.2014.04.020

Eldan, R., & Shamir, O. (2016). The power of depth for feedforward neural networks. *Conference on Learning Theory*. https://doi.org/10.48550/arXiv.1512.03965

Faria, E., & Fleten, S.-E. (2011). Day-ahead market bidding for a nordic hydropower producer: Taking the elbas market into account. *Computational Management Science*, *8*, 75–101. https://doi.org/10.1007/s10287-009-0108-5

Geissmann, T., & Obrist, A. (2018). Fundamental price drivers on continental european day-ahead power markets, 75. https://doi.org/10.2139/ssrn.3211339

Gelabert, L., Labandeira, X., & Linares, P. (2011). An ex-post analysis of the effect of renewables and cogeneration on spanish electricity prices. *Energy Economics*, *33*, S59–S65. https://doi.org/10.1016/j.eneco.2011.07.027

Hochreiter, J. (1991). Untersuchungen zu dynamischen neuronalen netzen. https://people.idsia.ch/~juergen/SeppHochreiter1991ThesisAdvisorSchmidhuber.pdf

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, *20*(1), 5–10. https://doi.org/10.1016/j.ijforecast.2003.09.015

Holttinen, H. (2005). Optimal electricity market for wind power. *Energy Policy*, *33*(16), 2052–2063. https://doi.org/10.1016/j.enpol.2004.04.001

Holttinen, H., Saarikivi, P., Repo, S., Ikäheimo, J., & Koreneff, G. (2006). Prediction errors and balancing costs for wind power production in finland.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., & Wilson, A. (2018). Averaging weights leads to wider optima and better generalization. https://doi.org/10.48550/arXiv.1803.05407

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, *38*(18), 2895–2907. https://doi.org/10.1016/j.atmosenv.2004.02.026

Ketterer, J. (2014). The impact of wind power generation on the electricity price in germany. *Energy Economics*, *44*, 270–280. https://doi.org/10.1016/j.eneco.2014.04.003

Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, *32*(3), 669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1412.6980

Knapik, O. (2017). Modeling and forecasting electricity price jumps in the nord pool power market. *Aarhus University, Department of Economics and Business Economics, CREATES - Center for Research in Econometric Analysis of Time Series,* https://pure.au.dk/ws/files/109020806/rp17_07.pdf

Kyritsis, E., Andersson, J., & Serletis, A. (2017). Electricity prices, large-scale renewable integration, and policy implications. *Energy Policy*, *101*, 550–560. https://doi.org/10.1016/j.enpol.2016.11.014

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

Lago, J., Ridder, F. D., & Schutter, B. D. (2018). Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. *Applied Energy*, *221*, 386–405. https://doi.org/10.1016/j.apenergy.2018.02.069

Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, *37*(4), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

Lu, Z., Pu, H., Wang, F., Hu, Z., & Wang, L. (2017). The expressive power of neural networks: A view from the width.

Maciejowska, K., Nitka, W., & Weron, T. (2019). Day-ahead vs. intraday—forecasting the price spread to maximize economic benefits. *Energies*, *12*(4). https://doi.org/10.3390/en12040631

Makridakis, S. (1993). Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting*, *9*(4), 527–529. https://doi.org/10.1016/0169-2070(93)90079-3

Mauritzen, J. (2013). Dead battery? wind power, the spot market, and hydropower interaction in the nordic electricity market. *The Energy Journal*, *Volume 34*. https://doi.org/10.5547/01956574.34.1.5

Mauritzen, J. (2015). Now or later? trading wind power closer to real-time and how poorly designed subsidies lead to higher balancing costs. https://doi.org/10.5547/01956574.36.4.jmau

Mohamed Noor, N., Abdullah, M. M. A. B., Yahaya, A. S., & Ramli, N. (2014). Comparison of linear interpolation method and mean method to replace the missing values in environmental data set. *Materials Science Forum*, *803*, 278–281. https://doi.org/10.4028/www.scientific.net/MSF.803.278

Mosquera-López, S., Uribe, J. M., & Manotas-Duque, D. F. (2018). Effect of stopping hydroelectric power generation on the dynamics of electricity prices: An event study approach. *Renewable and Sustainable Energy Reviews*, *94*, 456–467. https://doi.org/10.1016/j.rser.2018.06.021

Narajewski, M., & Ziel, F. (2020). Econometric modelling and forecasting of intraday electricity prices. *Journal of Commodity Markets*, *19*, 100107. https://doi.org/10.1016/j.jcomm.2019.100107

Nguyen, T., Raghu, M., & Kornblith, S. (2021). Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. https://doi.org/10.48550/ARXIV.2010.15327

Oh, K.-S., & Jung, K. (2004). GPU implementation of neural networks. *Pattern Recognition*, *37*(6), 1311–1314. https://doi.org/10.1016/j.patcog.2004.01.013

Perez-Arriaga, I. J., & Batlle, C. (2012). Impacts of intermittent renewables on electricity generation system operation. *Economics of Energy &amp; Environmental Policy*, *Volume 1*. https://doi.org/10.5547/2160-5890.1.2.1

Polson, M., & Sokolov, V. (2019). Deep learning for energy markets. https://doi.org/10.48550/arxiv.1808.05527

Psiloglou, B. E., Giannakopoulos, C., Majithia, S., & Petrakis, M. (2009). Factors affecting electricity demand in athens, greece and london, UK: A comparative assessment. *Energy*, *34*(11), 1855–1863. https://doi.org/10.1016/j.energy.2009.07.033

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *ArXiv*, *abs/1811.12808*. https://doi.org/10.48550/arXiv.1811.12808

Rintamäki, T., Siddiqui, A., & Salo, A. (2017). Does renewable energy generation decrease the volatility of electricity prices? an analysis of denmark and germany. *Energy Economics*, *62*. https://doi.org/10.1016/j.eneco.2016.12.019

Ruder, S. (2017). An overview of gradient descent optimization algorithms. https://arxiv.org/pdf/1609.04747.pdf

Russo, M. A., Carvalho, D., Martins, N., & Monteiro, A. (2022). Forecasting the inevitable: A review on the impacts of climate change on renewable energy resources. *Sustainable Energy Technologies and Assessments*, *52*, 102283. https://doi.org/10.1016/j.seta.2022.102283

Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, *36*(3), 1181–1191. https://doi.org/10.1016/j.ijforecast.2019.07.001

Scharff, R., & Amelin, M. (2016). Trading behaviour on the continuous intraday market elbas. *Energy Policy*, *88*, 544–557. https://doi.org/10.1016/j.enpol.2015.10.045

Shah, D., Campbell, W., & Zulkernine, F. (2018). A comparative study of LSTM and DNN for stock market forecasting. https://doi.org/10.1109/BigData.2018.8622462

Sprangers, O., Schelter, S., & Rijke, M. d. (2022). Parameter-efficient deep probabilistic forecasting. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2021.11.011

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. https://doi.org/10.48550/ARXIV.1409.3215

Tangerås, T., & Mauritzen, J. (2014). Real-time versus day-ahead market power in a hydro-based electricity market. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2398857

Twomei, J. C., & Smith, A. E. (1995). Performance measures, consistency, and power for artificial neural network models*. *Elsevier Science*, *21*, 243–258. https://doi.org/10.1016/0895-7177(94)00207-5

Unger, E. A., Ulfarsson, G. F., Gardarsson, S. M., & Matthiasson, T. (2018). The effect of wind energy production on cross-border electricity pricing: The case of western denmark in the nord pool market. *Economic Analysis and Policy*, *58*, 121–130. https://doi.org/10.1016/j.eap.2018.01.006

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection." BMC bioinformatics, 7(1), 91. *BMC bioinformatics*, *7*, 91. https://doi.org/10.1186/1471-2105-7-91

Wang, S., Feng, J., & Liu, G. (2013). Application of seasonal time series model in the precipitation forecast. *Mathematical and Computer Modelling*, *58*(3), 677–683. https://doi.org/10.1016/j.mcm.2011.10.034

Weber, C. (2010). Adequate intraday market design to enable the integration of wind energy into the european power systems. *Energy Policy*, *38*(7), 3155–3163. https://doi.org/10.1016/j.enpol.2009.07.040

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*, 79–82. https://doi.org/10.3354/cr030079

Winters, P. (1960). Forecasting sales by exponentially weighted moving averges. *Management Science*, *6*, 324–342.

Wright, L., & Demeure, N. (2021). Ranger21: A synergistic deep learning optimizer. *ArXiv*, *abs/2106.13731*. https://doi.org/10.48550/arXiv.2106.13731

Wu, W., Liao, W., Miao, J., & Du, G. (2019). Using gated recurrent unit network to forecast short-term load considering impact of electricity price. *Energy Procedia*, *158*, 3369–3374. https://doi.org/10.1016/j.egypro.2019.01.950

Würzburg, K., Labandeira, X., & Linares, P. (2013). Renewable generation and electricity prices: Taking stock and new evidence for germany and austria. *Energy Economics*, *40*, S159–S171. https://doi.org/10.1016/j.eneco.2013.09.011

Zhao, P., Wang, Q. J., Wu, W., & Yang, Q. (2021). Which precipitation forecasts to use? deterministic versus coarser-resolution ensemble NWP models. *Quarterly Journal of the Royal Meteorological Society*, *147*(735), 900–913. https://doi.org/10.1002/qj.3952

Zhu, Y. (2005). Ensemble forecast: A new approach to uncertainty and predictability. *Advances in Atmospheric Sciences*, *22*(6), 781–788. https://doi.org/10.1007/BF02918678

Books

Brown, R. (1959). *Statistical forecasting for inventory control*. McGraw-Hill.

Canale, R., & Chapra, S. (1998). *Numerical methods for engineers with programming and software applications* (3rd ed.). McGraw-Hill.

Chollet, F. (2018). *Deep learning with python*. Manning. https://tanthiamhuat.files.wordpress.com/2018/03/deeplearningwithpython.pdf

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning* (2nd ed.). Springer. https://hastie.su.domains/Papers/ESLII.pdf

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning* (12th ed.). Springer. https://hastie.su.domains/ElemStatLearn/printings/ESLII_print12_toc.pdf

Haykin, S. (2009). *Neural networks and learning machines*. New Jersey: Pearson Education Upper Saddle River.

Hyndman, R., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts.com/fpp2

Hyndman, R., & Athanasopoulos, G. (2022, May 23). *Forecasting: Principles and practice* (3rd ed.). OTexts. OTexts.com/fpp3

In books

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier & G. Saporta (Eds.), *Proceedings of COMPSTAT'2010* (pp. 177–186). Physica-Verlag HD.

Glorot, X., Bordes, A., & Bengio, Y. (2011, April 11). Deep sparse rectifier neural networks. In G. Gordon, D. Dunson, & M. Dudík (Eds.), *Proceedings of the fourteenth interna-*

tional conference on artificial intelligence and statistics (pp. 315–323). PMLR. https://proceedings.mlr.press/v15/glorot11a.html

Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.-X., & Yan, X. (2019). Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Proceedings of the 33rd international conference on neural information processing systems*. Curran Associates Inc.

Online articles

About us. (n.d.). https://www.nordpoolgroup.com/About-us/

Arik, S., & Pfister, T. (n.d.). *Interpretable deep learning for time series forecasting*. https://ai.googleblog.com/2021/12/interpretable-deep-learning-for-time.html

Baijayanta, R. (2020, April 6). *All about feature scaling* [Towards data science]. https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35

Berk, M. (2021, August 4). *How to forecast time series data using deep learning* [Towards data science]. https://towardsdatascience.com/deep-learning-for-time-series-data-ed410da30798

Bidding areas. (n.d.). https://www.nordpoolgroup.com/the-power-market/Bidding-areas/

Brownlee, J. (2020, September 12). *Understand the impact of learning rate on neural network performance* [Machine learning mastery]. https://machinelearningmastery.com/understand-the-dynamics-of-learning-rate-on-deep-learning-neural-networks/

Brownlee, J. (2021, January 13). *How to use ROC curves and precision-recall curves for classification in python* [Machine learning mastery]. How%20to%20Use%20ROC%20Curves%20and%20Precision-Recall%20Curves%20for%20Classification%20in%20Python

Data types for curves [Energy quantified]. (n.d.). https://app.energyquantified.com/knowledge-base/articles/1

DeepAR forecasting algorithm [Amazon]. (n.d.). https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html

Energy and marine resources [Norway]. (n.d.). https://www.norway.no/en/missions/eu/values-priorities/energy-marine-res/

Instance tags [Energy quantified]. (n.d.). https://app.energyquantified.com/knowledge-base/articles/43

Kraftmarkedet [Energi fakta norge]. (n.d.). https://energifaktanorge.no/norsk-energiforsyning/kraftmarkedet/

Mest kraftutveksling med norden. (2021, December 3). https://www.statnett.no/om-statnett/nyheter-og-pressemeldinger/nyhetsarkiv-2021/mest-kraftutveksling-med-norden/

Nord pool announces 2021 trading figures. (2022, January 18). https://www.nordpoolgroup.com/message-center-container/newsroom/exchange-message-list/2022/q1/nord-pool-announces-2021-trading-figures/

NordLink [Statnett]. (n.d.). https://www.statnett.no/en/our-projects/interconnectors/nordlink/

Olah. (2015, August 27). *Understanding LSTM networks.* https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Or, B. (2020, October 10). *The exploding and vanishing gradients problem in time series.* https://towardsdatascience.com/the-exploding-and-vanishing-gradients-problem-in-time-series-6b87d558d22

An overview of the nordic electricity market. (2019, January 10). https://www.nordic-energyregulators.org/about-nordreg/an-overview-of-the-nordic-electricity-market/

The power market. (n.d.). https://www.nordpoolgroup.com/the-power-market/

Regulation information per area [Nord pool]. (n.d.). https://www.nordpoolgroup.com/en/Market-data1/Regulating-Power1/Regulating-Power--Area1/NO11/Norway/?view=table

Reserves and balancing power [Fingrid]. (n.d.). https://www.fingrid.fi/en/electricity-market/reserves_and_balancing/#reserve-products

Security of electricity supply. (n.d.). https://energifaktanorge.no/en/norsk-energiforsyning/forsyningssikkerhet/

The skagerrak 4-interconnector - cable contracts signed [Statnett]. (2013, August 4). https://www.statnett.no/en/about-statnett/news-and-press-releases/News-archive-2011/the-skagerrak-4-interconnector---cable-contracts-signed/

Spot price model improvements. (2022, January 25). https://app.energyquantified.com/login?next=/knowledge-base/articles/54

Trial operation at NSL starts on 1 october [Statnett]. (2021, September 28). https://www.statnett.no/en/for-stakeholders-in-the-power-industry/news-for-the-power-industry/trial-operation-at-nsl-starts-on-1-october/

UK and norway. (n.d.). https://www.northsealink.com/en/news/national-grid-powers-up-world-s-longest-subsea-interconnector-between-the-uk-and-norway

Understanding the term 'dispatchable' regarding electricity generation [NMPP energy]. (2021, March 17). https://www.nmppenergy.org/energy-education/understanding-term-dispatchable-regarding-electricity-generation

Weather forecast models and schedule. (n.d.). https://app.energyquantified.com/knowledge-base/articles/4

Weather indexes [Energy quantified]. (n.d.). https://app.energyquantified.com/knowledge-base/articles/55

What is the kyoto protocol? [United nations climate change]. (n.d.). https://unfccc.int/kyoto_protocol

Thesis

Kolberg, J., & Waage, K. (2018). *Artificial intelligence and nord pool's intraday electricity market elbas: A demonstration and pragmatic evaluation of employing deep learning for price prediction* (Doctoral dissertation). Norwegian School of Economics. https://openaccess.nhh.no/nhh-xmlui/handle/11250/2560898

Presentations

High energy prices [EU Agency for the Cooperation of Energy Regulators (ACER)]. (2021, September). https://documents.acer.europa.eu/en/The_agency/Organisation/Documents/Energy%20Prices_Final.pdf

Hinton, G., Bengio, Y., & LeCun, Y. (2015). *Deep learning* [Deep Learning Tutorial]. http://www.iro.umontreal.ca/~bengioy/talks/DL-Tutorial-NIPS2015.pdf

Appendix

A Background

A.1 Power Market

A.1.1 Balancing markets

The regulating volume available to a TSO can broadly be divided into three types: primary reserves (FCR), secondary reserves (aFFR) or tertiary reserves (mFRR) ("Reserves and balancing power", n.d.). The primary and secondary reserves act automatically as a response to deviation in frequency, while the tertiary reserves are manually initiated by the TSO. The procedure is as follows: using primary reserve when the imbalance occurs, the second reserves take effect when the imbalance lasts for several minutes in order to free up primary reserve for a new imbalance, and tertiary reserve initiated when the first two reserves come in short with up to 15 minute response time. Primary and secondary reserves are traded in separate hourly and weekly markets while tertiary reserves are paid upfront to bidders guaranteeing available regulation regardless of utilization of the available resources.

A.1.2 Balancing market: price-determination for consumers and producers

The Norwegian regulating market follows one-price system for consumption-imbalances, and a two-price system for production imbalances. The one-price system entails that consumers pay the regulating price of the dominating regulated power volumes. "Regulation information per area" (n.d.). Producers however, are only penalized for increasing the existing regulation imbalance, that is, moving in the opposite direction of the regulation measure taken by the TSO (Bourry and Kariniotakis, 2009). On the other hand, should a power producer alleviate an imbalance by for example buy available electricity when a surplus exists, the volume is paid for by the day-ahead/spot price for that delivery hour (Holttinen et al., 2006).

B Data

B.1 Pre-Processing

B.1.1 Feature scaling

Max-Min Scaler, *Standard Scaler*, *Max Absolute Scaler* and *Robust Scaler* are some of the scaling methods that can be used. Each of the existing scaling methods also requires that certain assumptions are fulfilled and are not always appropriate for certain data. *Standard Scaler* assumes that the data are Gaussian distributed, and observing all the histograms of all features in figure 30 and 31 (see Appendix) indicates that not all are normally distributed (Baijayanta, 2020). *Max-Min Scaler* and *Max Absolute Scaler* are sensitive to outliers, and studying the boxplots of all features in figure 32 and 33, in the appendix, one can see that most of the features on hourly frequency suffers from outliers.

B.1.2 Time zone

Coordinated Universal Time (UTC) and *Central European Time (CET)* are two standardized time zone where CET is one hour ahead of UTC during winter time and two hours ahead during summer time (daylight saving time), commonly called *CEST*. Since most of our features from Energy Quantified uses CET, especially hourly data, we opt to use this as hour standard time zone. Some of the daily data in Energy Quantified and Nord Pool uses UTC, which needs transformation to CET.

C Methodology

C.1 Neural Networks

C.1.1 Checkpoint ensembling

Checkpoint ensembling is an inexpensive technique that has been shown to increase generalizability by averaging predictions of the n best performing renditions of a given model (H. Chen et al., 2017). Checkpoint ensembling is used extensively, for all models presented in this thesis.

C.1.2 Regularization techniques

Early stopping

Early stopping refers to ending model training once out-of-sample performance has degraded or is no longer improving after new epochs of training. Early stopping has been described as “free lunch”, implying that its performance increase suffers no penalty (Hinton et al., 2015).

Dropout

Dropout is a dramatic yet effective regularization technique. The technique approximates the performance of multiple renditions of the same model by leaving out random neurons (Srivastava et al., 2014).

C.1.3 Model convergence

Gradient clipping Gradient clipping seeks to increase the stability and thus the expected speed of model convergence. Simply put, the step performed in parameter updates may become too large and "skip" over local minima undoing previous valuable training (Goodfellow et al., 2016). Clipping restricts the value of the parameter gradient just before a parameter update occurs.

C.1.4 Stochastic Weight Averaging

As we have seen Stochastic Gradient Descent (SGD) involves calculating training loss and updating loss for all samples drawn from a random batches z_t. A single weight update is done on the basis of subtracting the calculated loss from the random sample: $w_{t+1} = w_t - \delta_w Q(z_t, w_t)$, where λ_{yt} is a learning rate (Bottou, 2010). We remind the reader that all optimization techniques involving gradient descent bears the risk of learning stalling in local minima far from the global minima. We emphasize for the reader that the purpose of learning and corresponding parameter updates is to approximate as closely as possible the global minimum. In neural network training involves carefully balancing the number of steps, batch size, type of optimizer and learning rate to carefully balance the risk of converging too quick or overfitting. The authors of this thesis can vouch for the frustrations that choosing optimizers and finetuning learning rates may involve. A recent technique seeks to mitigate the risk of getting stuck i local optima by using a higher learning rate to explore nearby optima (Izmailov et al., 2018). Stochastic Weight Averaging is used in our model selection as a callback implemented in Pytorch Lightning (Pytorch, 2022). After an attempt of finding a local optimum using a low learning rate with an non-averaging optimizer such as ADAM or Ranger, we switch to SWA after a preset number of epochs (frequently, 12 or 15). This results in some rather unusual-looking learning curves, as the constant learning rate is applied. In the burn-in period, training and validation loss typically increases before approaching a new optima. This can be seen in several of the learning curves presented in the appendix, most notably in the intraday GRU model seen in figure 40.

D Figures

D.1 Data

D.1.1 Data Exploration

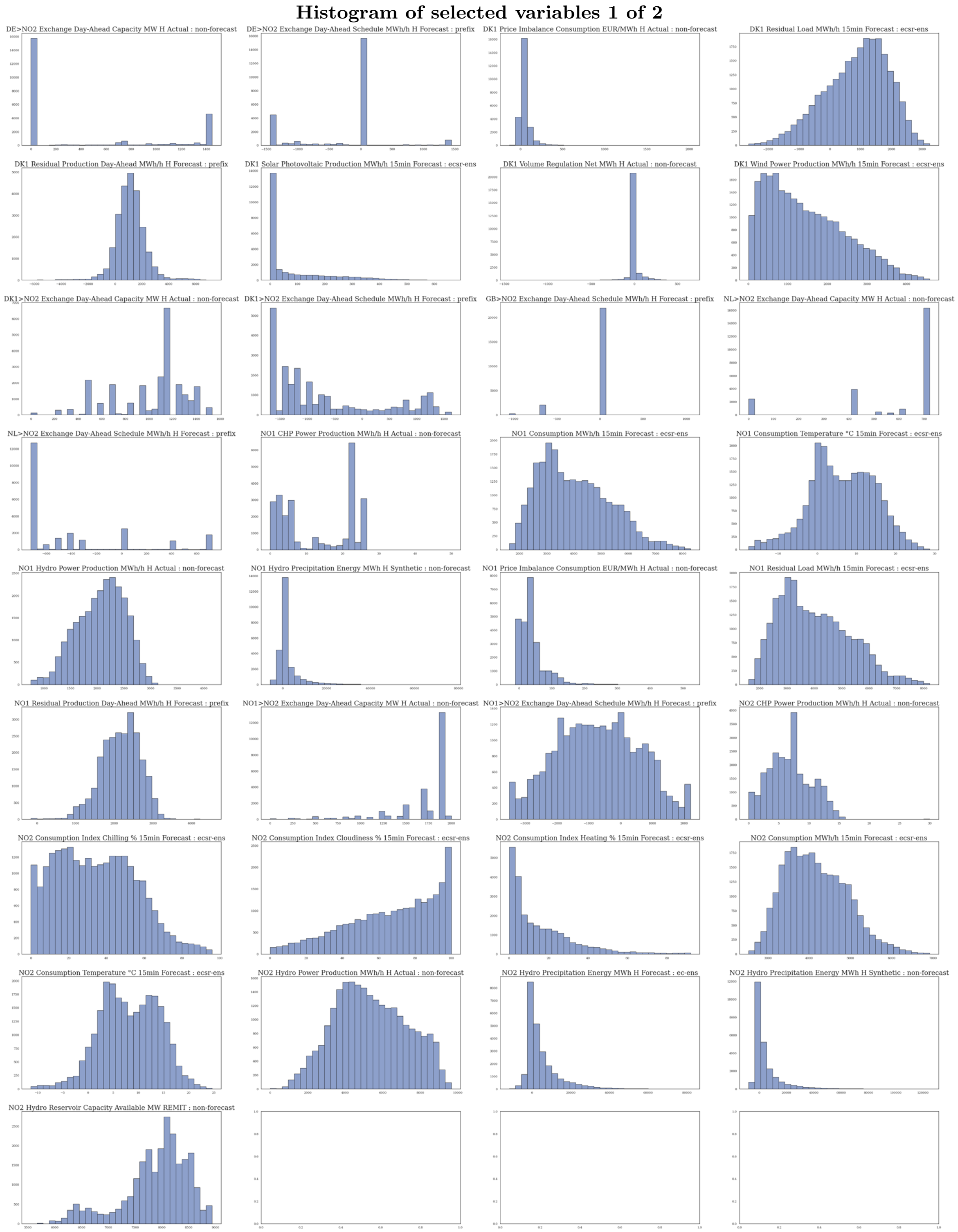


Figure 30: Histograms of all variable used

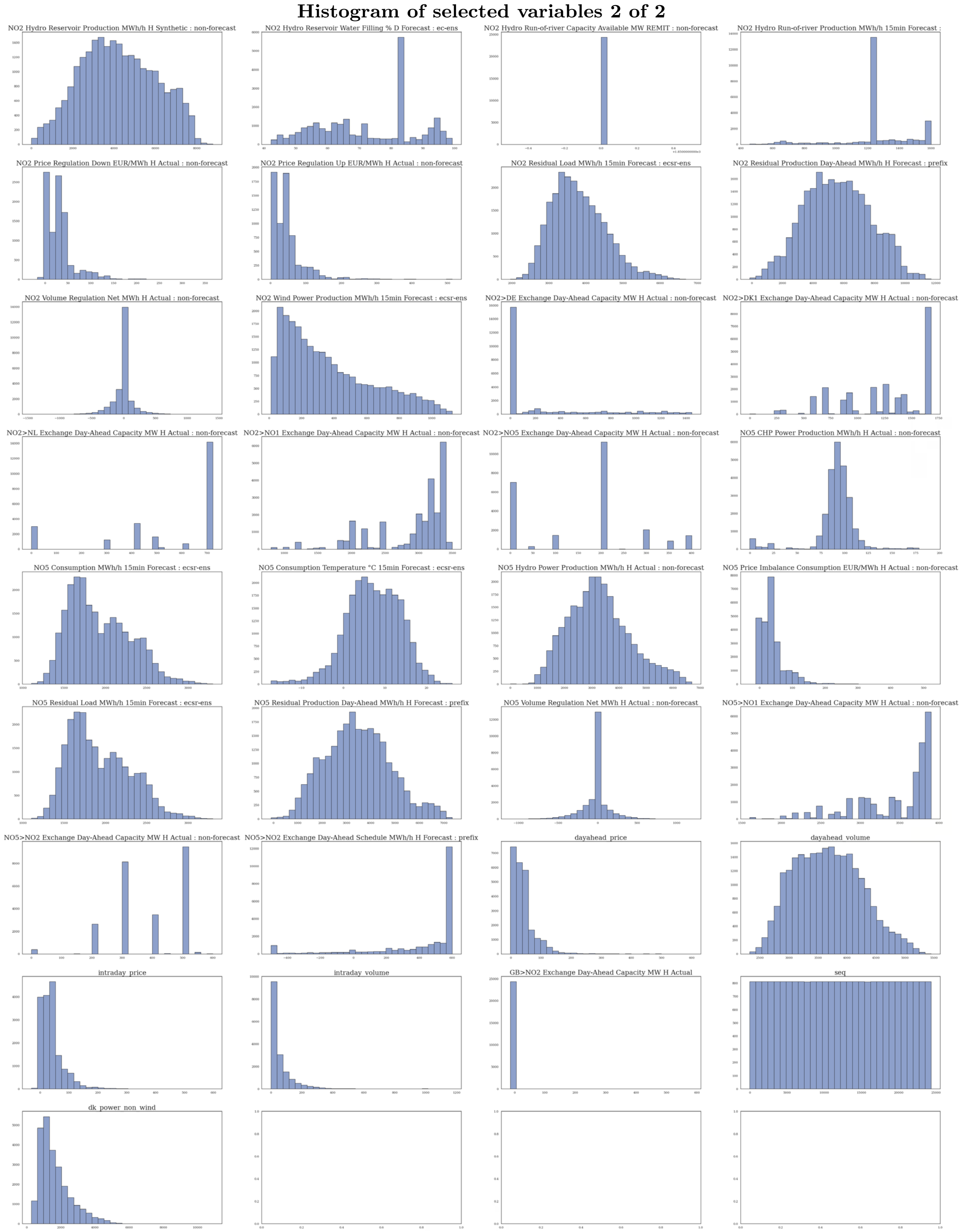


Figure 31: Histograms of all variable used

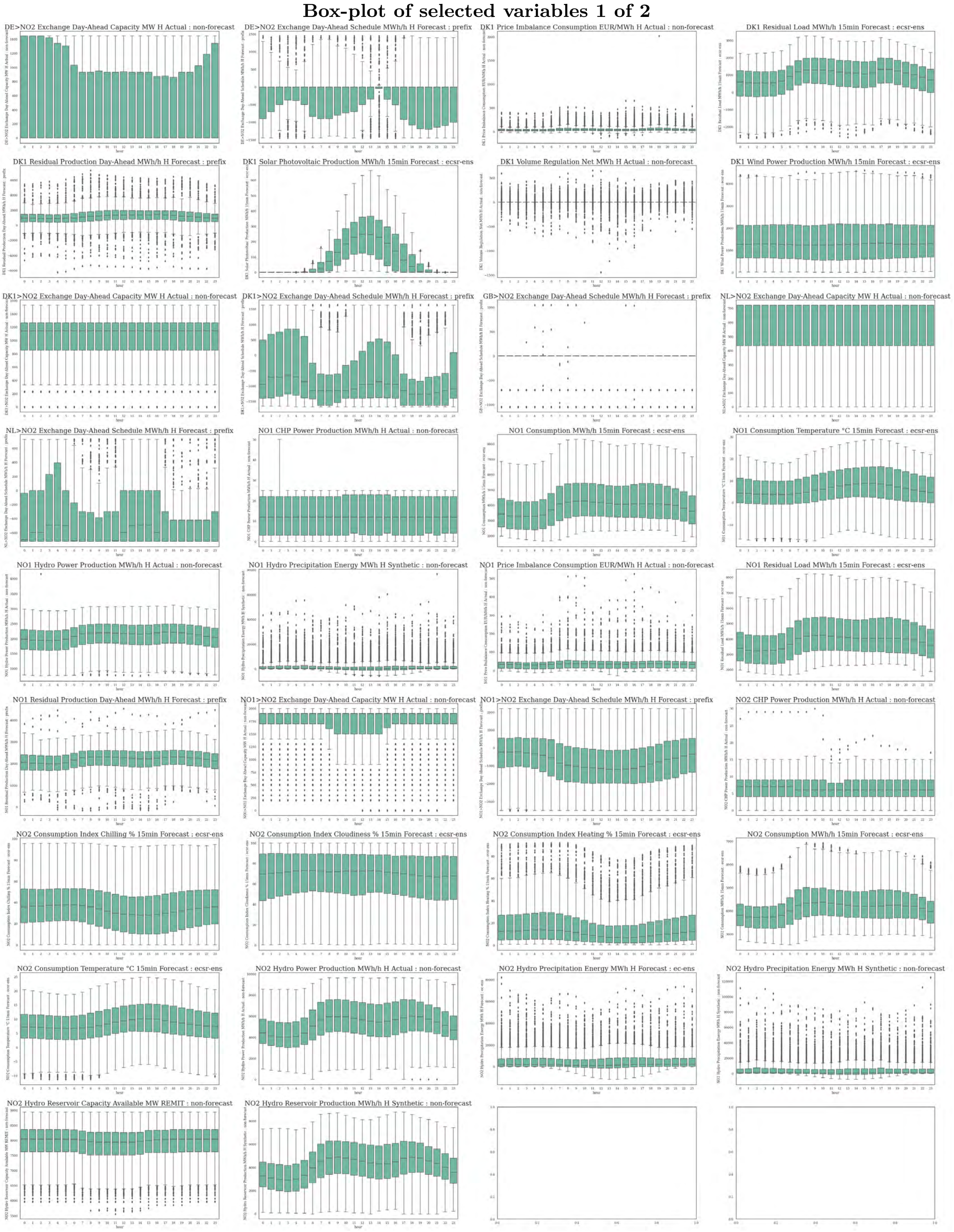
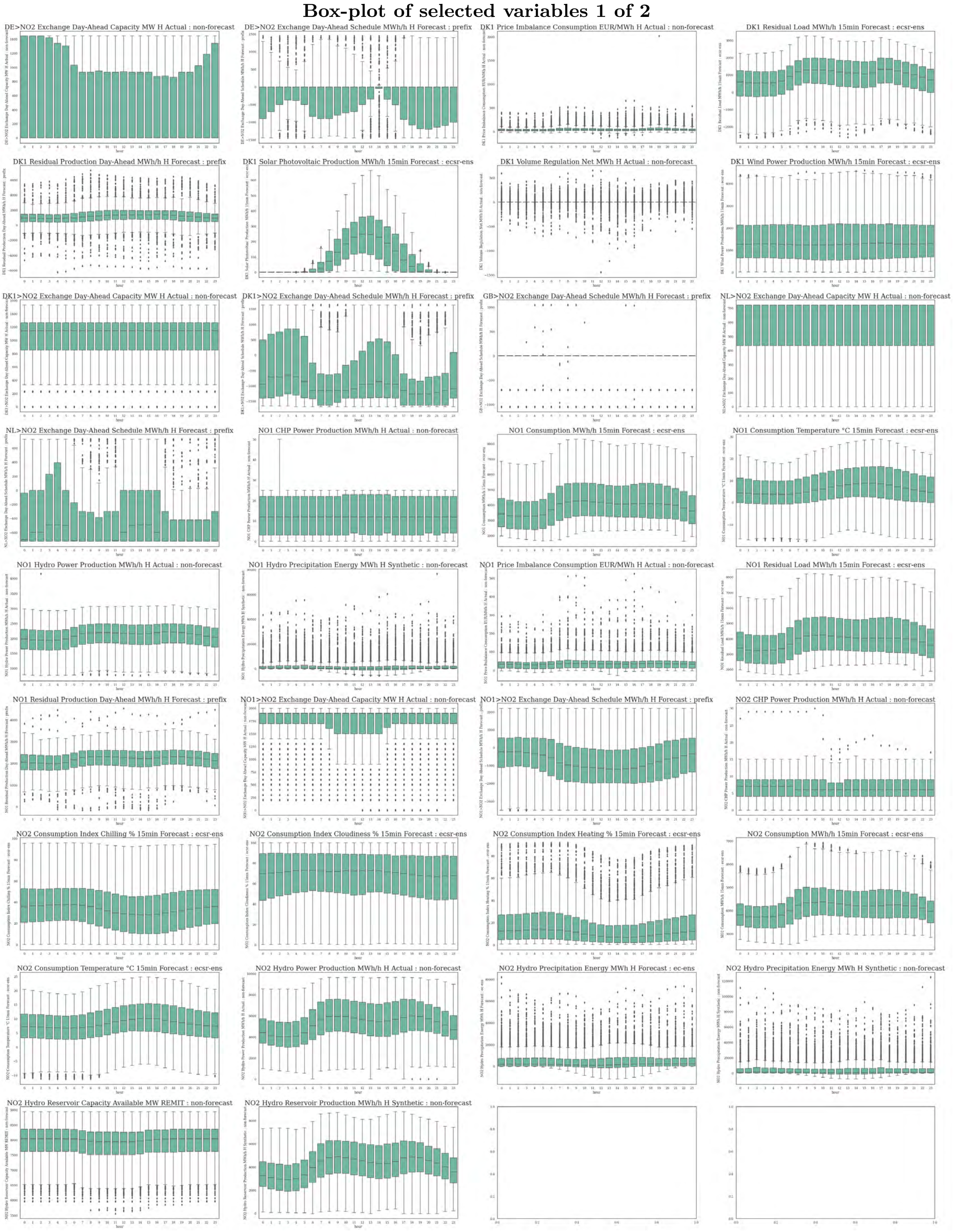


Figure 32: Boxplot of all variable used



Box-plot of selected variables 1 of 2

DE>NO2 Exchange Day-Ahead Capacity MW H Actual : non-forecast

DE>NO2 Exchange Day-Ahead Schedule MWh/h H Forecast : prefix

DK1 Price Imbalance Consumption EUR/MWh H Actual : non-forecast

DK1 Residual Load MWh/h 15min Forecast : ecsr-ens

DK1 Residual Production Day-Ahead MWh/h H Forecast : prefix

DK1 Solar Photovoltaic Production MWh/h 15min Forecast : ecsr-ens

DK1 Volume Regulation Net MWh H Actual : non-forecast

DK1 Wind Power Production MWh/h 15min Forecast : ecsr-ens

DK1>NO2 Exchange Day-Ahead Capacity MW H Actual : non-forecast

DK1>NO2 Exchange Day-Ahead Schedule MWh/h H Forecast : prefix

GB>NO2 Exchange Day-Ahead Schedule MWh/h H Forecast : prefix

NL>NO2 Exchange Day-Ahead Capacity MW H Actual : non-forecast

NL>NO2 Exchange Day-Ahead Schedule MWh/h H Forecast : prefix

NO1 CHP Power Production MWh/h H Actual : non-forecast

NO1 Consumption MWh/h 15min Forecast : ecsr-ens

NO1 Consumption Temperature °C 15min Forecast : ecsr-ens

NO1 Hydro Power Production MWh/h H Actual : non-forecast

NO1 Hydro Precipitation Energy MWh H Synthetic : non-forecast

NO1 Price Imbalance Consumption EUR/MWh H Actual : non-forecast

NO1 Residual Load MWh/h 15min Forecast : ecsr-ens

NO1 Residual Production Day-Ahead MWh/h H Forecast : prefix

NO1>NO2 Exchange Day-Ahead Capacity MW H Actual : non-forecast

NO1>NO2 Exchange Day-Ahead Schedule MWh/h H Forecast : prefix

NO2 CHP Power Production MWh/h H Actual : non-forecast

NO2 Consumption Index Chilling % 15min Forecast : ecsr-ens

NO2 Consumption Index Cloudiness % 15min Forecast : ecsr-ens

NO2 Consumption Index Heating % 15min Forecast : ecsr-ens

NO2 Consumption MWh/h 15min Forecast : ecsr-ens

NO2 Consumption Temperature °C 15min Forecast : ecsr-ens

NO2 Hydro Power Production MWh/h H Actual : non-forecast

NO2 Hydro Precipitation Energy MWh H Forecast : ec-ens

NO2 Hydro Precipitation Energy MWh H Synthetic : non-forecast

NO2 Hydro Reservoir Capacity Available MW REMIT : non-forecast

NO2 Hydro Reservoir Production MWh/h H Synthetic : non-forecast

Figure 33: Boxplot of all variable used

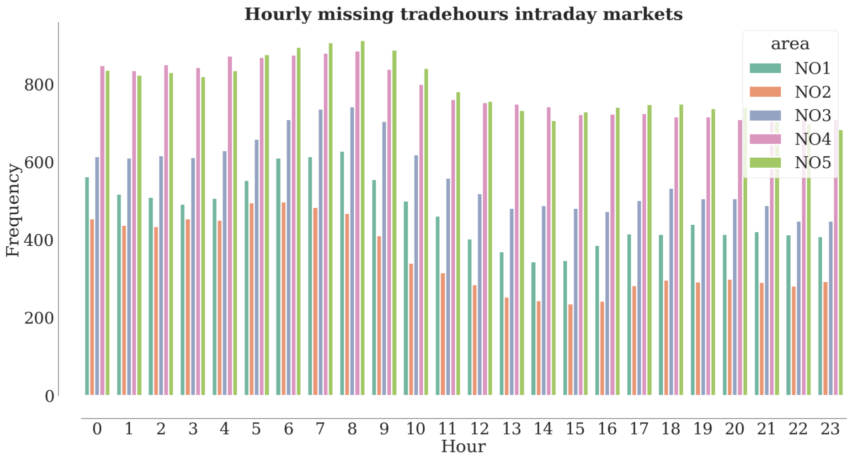


Figure 34: Frequency of missing tradehours on intraday markets per hour

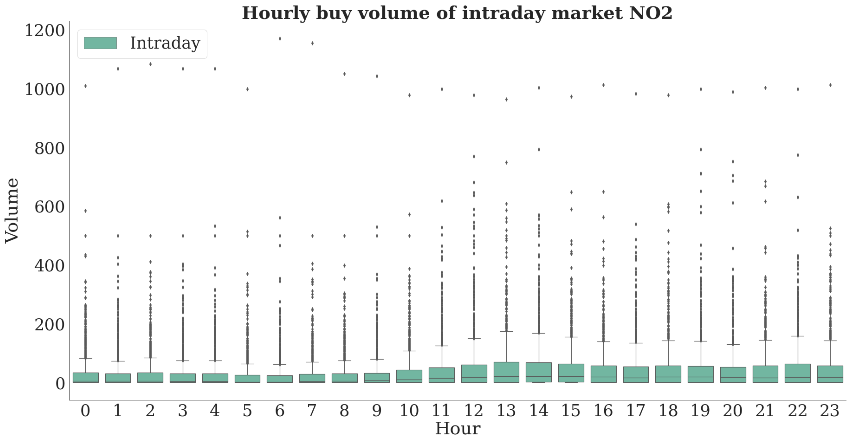


Figure 35: Boxplot of intraday buy volume in NO2 per hour in 2019-2022

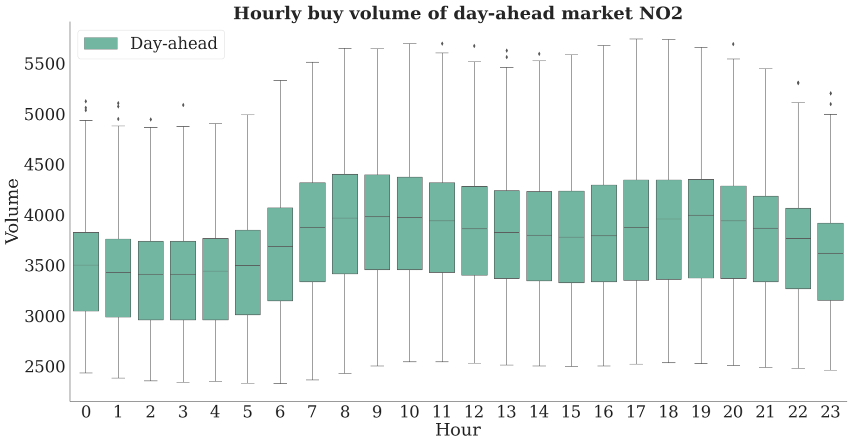


Figure 36: Boxplot of day-ahead buy volume in NO2 per hour in 2019-2022

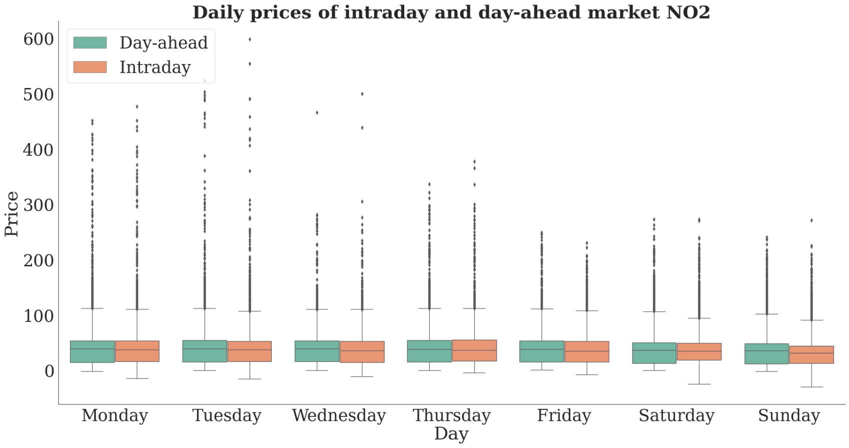


Figure 37: Boxplot of intraday and day-ahead prices per day for area NO2 in 2019-2022

E Tables

E.1 Price-Drivers

Time features

Series	Unit	Example
Hour	Integer	[0, 1, ..., 23]
Day	String	[Monday, Tuesday, ..., Sunday]
Week	Integer	[1, 2, ..., 52]
Month	String	[January, February, ..., December]

Table 11: Time features

Binary structural change features

Series	Unit	State	Period
Covid-19	String	[Yes, -]	12.03.2020 - 25.09.2021
Nord Link	String	[Yes, -]	31.03.2021
North Sea Link	String	[Yes, -]	01.10.2021

Table 12: Binary structural change features

E.2 Data

E.2.1 Consolidated time series

Consolidated features are mainly combined features, where the main purpose is to reduce the amount of features under the assumption that these are consider to explain the independent target variable together. From the literature (see section 1, 2 and 4.1) it is found that wind and solar photovoltaic have a significant impact on power prices, likely because of its intermittent property, and the fact that Norway (NO2) imports this type of energy because of the strongly reduced prices due to production surplus in Denmark, we opt to only include these power sources and leaving all the dispatchable power sources into a single feature as shown in table 13.

Consolidated time series

Series	Unit	Type	Resolution	Area
+ CHP biomass power production	MWh	Actual	Hourly	DK1
+ CHP central power production	MWh	Actual	Hourly	DK1
+ CHP decentral power production	MWh	Actual	Hourly	DK1
+ Hard coal power production	MWh	Actual	Hourly	DK1
+ Natural gas power production	MWh	Actual	Hourly	DK1
+ Oil power production	MWh	Actual	Hourly	DK1
+ Other power production	MWh	Actual	Hourly	DK1
+ Waste power production	MWh	Actual	Hourly	DK1
= Dispatchable power production	**MWh**	**Actual**	**Hourly**	**DK1**

Table 13: Showing how the features are combined into a single time series. The dispatchable power production series is named *dk power non wind* in the code

E.2.2 Pre-processing

Summary of missing hours for all selected time series

Series	Missing values	Share of total
NO5 Residual Load MWh/h 15min Forecast	38	0.16%
NO2 Residual Load MWh/h 15min Forecast	38	0.16%
NO1 Residual Load MWh/h 15min Forecast	38	0.16%
DK1 Residual Load MWh/h 15min Forecast	38	0.16%
NO2 Consumption Index Cloudiness % 15min Forecast	31	0.13%
NO5 Consumption Temperature °C 15min Forecast	24	0.10%
NO5 Consumption MWh/h 15min Forecast	24	0.10%
NO2 Wind Power Production MWh/h 15min Forecast	24	0.10%
NO2 Consumption Temperature °C 15min Forecast	24	0.10%
NO2 Consumption MWh/h 15min Forecast	24	0.10%
NO2 Consumption Index Heating % 15min Forecast	24	0.10%
NO2 Consumption Index Chilling % 15min Forecast	24	0.10%
NO1 Consumption Temperature °C 15min Forecast	24	0.10%
NO1 Consumption MWh/h 15min Forecast	24	0.10%
DK1 Wind Power Production MWh/h 15min Forecast	24	0.10%
DK1 Solar Photovoltaic Production MWh/h 15min Forecast	24	0.10%
NO2 Hydro Reservoir Production MWh/h H Synthetic	23	0.09%
NO5 CHP Power Production MWh/h H Actual	7	0.03%
NO2 CHP Power Production MWh/h H Actual	7	0.03%
NO1 CHP Power Production MWh/h H Actual	7	0.03%
NO1 Hydro Power Production MWh/h H Actual	6	0.02%
Day-ahead Trade Volume	3	0.01%
Day-ahead Price	3	0.01%

Table 14: Summary of missing data for each time series

E.2.3 Data Exploration

Summary statistics of Nord Pool volumes in 2019-2022

Market	Area	Count	Mean	St.dev	Min	25%	50%	75%	Max
Day-ahead	NO1	26709	3980.1	1504.7	1514.8	2758.6	3678.0	5097.1	8717.2
Day-ahead	NO2	26709	3738.5	608.9	2324.3	3256.4	3721.6	4158.1	5740.4
Day-ahead	NO3	26709	2962.5	456.1	1928.6	2585.7	2912.0	3318.7	4182.1
Day-ahead	NO4	26709	1656.9	313.4	948.5	1405.0	1662.7	1903.3	2571.2
Day-ahead	NO5	26709	1691.0	408.9	830.7	1390.0	1653.0	1984.2	2801.1
Intraday	NO1	26688	21.1	37.3	0.0	0.0	5.0	26.3	467.0
Intraday	NO2	26688	39.7	75.1	0.0	0.0	10.8	45.3	1170.0
Intraday	NO3	26688	15.6	35.1	0.0	0.0	0.0	15.0	646.4
Intraday	NO4	26688	6.7	20.1	0.0	0.0	0.0	4.0	440.0
Intraday	NO5	26688	15.5	48.4	0.0	0.0	0.0	5.0	946.0

Table 15: Summary statistics of Nord Pool volumes in 2019-2022

Summary statistics of Nord Pool prices in 2019-2022

Market	Area	Count	Mean	St.dev	Min	25%	50%	75%	Max
Day-ahead	NO1	26733	42.6	40.9	-2.0	13.6	37.6	50.6	600.2
Day-ahead	NO2	26733	42.8	40.7	-2.0	13.6	37.6	52.2	600.2
Day-ahead	NO3	26733	29.8	22.0	-0.0	13.0	31.2	41.4	360.0
Day-ahead	NO4	26733	27.5	21.0	-0.0	11.9	26.3	39.3	360.0
Day-ahead	NO5	26733	42.5	40.5	-0.1	13.7	37.6	50.5	600.2
Intraday	NO1	15398	40.8	39.6	-29.7	16.5	35.0	48.7	563.1
Intraday	NO2	18147	42.1	41.6	-29.8	15.5	35.2	51.5	598.0
Intraday	NO3	12999	28.4	23.9	-45.0	12.2	27.1	39.0	437.0
Intraday	NO4	7837	27.1	25.5	-36.0	11.5	22.9	37.2	357.5
Intraday	NO5	7693	46.9	43.2	-29.0	21.0	36.4	60.9	580.9

Table 16: Summary statistics of Nord Pool prices in 2019-2022

E.3 Results

E.3.1 Model configurations

Hyperparameter configurations of day-ahead neural networks

Hyperparameters	LSTM	GRU	TFT	DeepAR
Accumulate gradient batch size	2	4	3	5
Batch size	128	128	64	64
Dropout	0.1	0.1	0.05	0.1
Encoding length	336	336	96	200
Gradient clipping	0.5	0.6	0.3	0.6
Hidden size	128	256	128	128
Layers	2	2	3	2
Learning rate	0.0001	0.001	0.05	0.001
Minimum delta	0.05	0.05	0.01	0.05
Number of ensembles	3	3	2	3
Optimizer	Ranger	Ranger	Ranger	Ranger
Patience	17	17	7	16
Reduce on plateu patience	2	2	2	2
Reduce on plateu reduction	3	2	2	2
SWA epoch start	15	12	10	15

Table 17: Hyperparameter configurations of day-ahead neural networks

Hyperparameter configurations of intraday neural networks

Hyperparameters	LSTM	GRU	TFT	DeepAR
Accumulate gradient batch size	2	4	3	2
Batch size	64	128	218	64
Dropout	0.1	0.1	0.05	0.1
Encoding length	120	300	168	120
Gradient clipping	0.005	0.6	0.3	0.05
Hidden size	128	256	64	128
Layers	2	2	3	2
Learning rate	0.00056	0.001	0.05	0.0001
Minimum delta	0.05	0.05	0.1	0.03
Number of ensembles	0	3	2	0
Optimizer	Adam	Ranger	Ranger	Adam
Patience	17	17	7	17
Reduce on plateu patience	2	2	2	3
Reduce on plateu reduction	2	2	2	3
SWA epoch start	12	12	10	13

Table 18: Hyperparameter configurations of day-ahead neural networks

E.3.2 Neural network classifiers

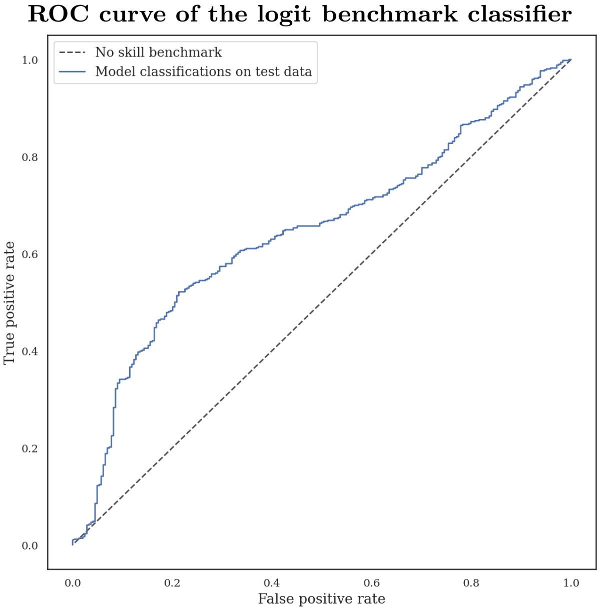


Figure 38: ROC curve of logit classifier

E.3.3 Performance evaluation

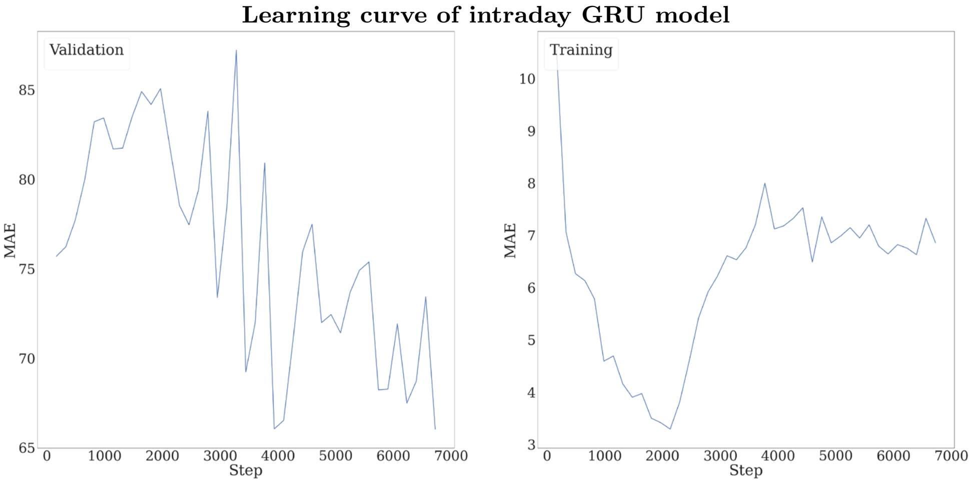


Figure 39: Learning curves displaying validation and training loss per training step

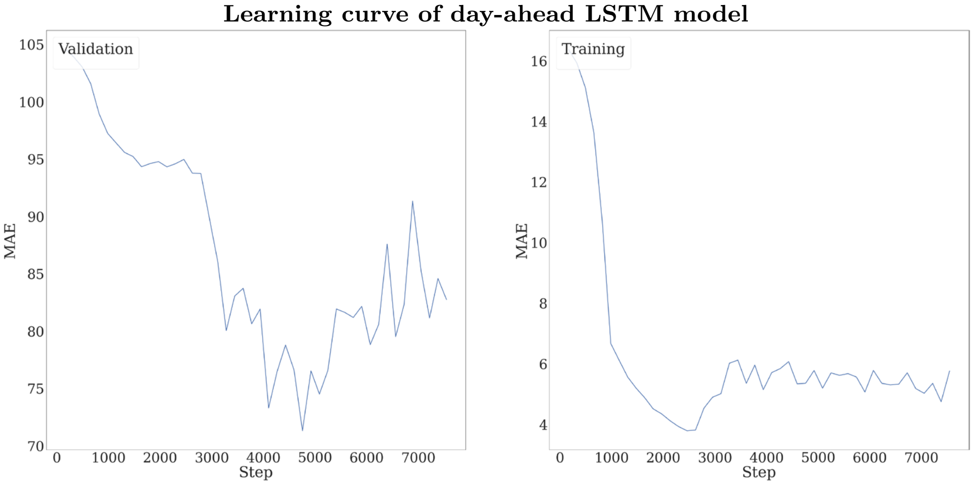


Figure 40: Learning curves displaying validation and training loss per training step