NTNU
Kunnskap for en bedre verden

Department of Electric Energy

TET4510 - Specialization Project

Reinforcement Learning for Reserve Markets

Author:
Haakon Nygård Hellebust

Date

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

Electric power systems increasingly rely on fast-acting flexible resources to maintain frequency quality as renewable energy sources displace traditional synchronous generation [1]. In the Nordic power markets, manual Frequency Restoration Reserves (mFRR) are activated to correct sustained imbalances after other faster reserves have been utilized, such as Frequency Containment Reserves (FCR) and automatic Frequency Restoration Reserves (aFRR). This results in a hierarchical reserve structure where mFRR serves as the last line of defense against frequency deviations [2]. mFRR activations are, therefore, relatively infrequent and often substantial when they do occur.

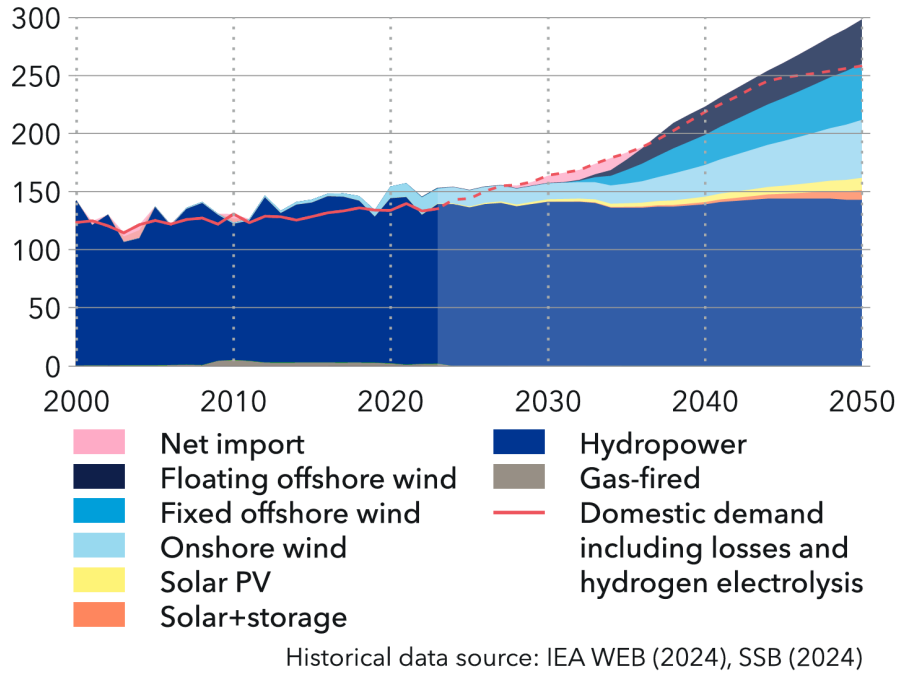


Figure 1: Norway electricity supply by power station and net imports with projections to 2050 [3]

1.2 Motivation

Since mFRR activations are infrequent but significant events in power system operation, accurately predicting their occurrence would provide substantial benefits. For aggregators of distributed flexible resources, such as home batteries and electric vehicles, knowing when mFRR activations are likely to occur can inform bidding strategies in the electricity balancing markets. Capacity markets and activation markets exist side by side, each with distinct revenue opportunities. Capacity markets reward participants for being available to provide reserves, while activation markets compensate for actual reserve activations. Accepted bids in capacity markets guarantees revenue regardless of whether an activation occurs, while activation market bids only yield revenue upon actual activations. Bidding strategies involving activation markets are inherently riskier, as revenue realization is uncertain.

If an aggregator is able to anticipate when mFRR activations are unlikely, the aggregator can confidently bid their resources into the capacity markets without the fear of missing out on high-value activation opportunities. Conversely, if an mFRR activation is predicted to be likely, the aggregator might prioritize bidding into the activation market to capitalize on potential revenues. Thus, a reliable prediction model for mFRR activations can enhance market participation strategies and optimize revenue streams for aggregators of flexible resources.

The recent reform of the mFRR market in March 2025, transitioning to a Nordic-wide mFRR activation market with quarter-hourly resolution [4], further motivates the development of a data-driven prediction model. The increased granularity provides more abundant data points for model training and evaluation, potentially improving prediction quality. Additionally, the reform introduces more nuanced price signals, which in theory could be captured by a well-designed prediction model. However, as the transition is recent at the time of writing, there is limited historical data available at the new resolution. This scarcity of data presents a challenge for model development, as machine learning models typically require substantial amounts of data to learn effectively. *bid acceptance not equal to activation* **not only looking into prices and volumes.**

The aim of this specialization project is, thereby, to provide further work with a mFRR activation prediction model that can be applied to comprehensive and realistic market strategies for aggregators of distributed flexible resources. One possability is to apply reinforcement learning techniques to develop near-optimal bidding strategies based on the predictions from the model. This is an interesting avenue for future research as it, if successful, could provide significant economic benefits for aggregators and enhance the overall efficiency of the electricity balancing markets.

handling uncertainty importanf for further work, near real tiem, continuous

1.3 Research Question

The central research question of this study is: How well can mFRR up- and down-regulation activations be predicted at 15-minute resolution using only real-time-available system state features and historical activation/price data? What is the relative contribution of these systemic features compared to lag-based features?

1.4 Outline

The remainder of this report is structured as follows: Section 2 provides an overview of the Nordic balancing markets and related work. Section 3 formulates the prediction problem and presents the dataset. Section 4 details the modelling methodology, including feature engineering and model selection. Section 5 presents the results and performance evaluation, while Section 6 discusses implications, limitations, and directions for future work

2 Electricity Balancing Market Theory

Electricity balancing markets are mechanisms designed to ensure the stability and reliability of the power grid by managing supply and demand in real-time. These markets facilitate the procurement of balancing services, which are necessary to maintain the equilibrium between electricity supply and demand. Balancing markets operate on the principle of economic efficiency, where market participants can offer their flexibility to the grid operator in exchange for compensation.

2.1 Nordic Balancing Market Structure

In Nordic countries, the balancing hierarchy consists of several layers, each serving a specific purpose in maintaining grid stability. Figure 2 illustrates the different reserve types, their activation times, and their place in the hierarchy.

Not manual anymore

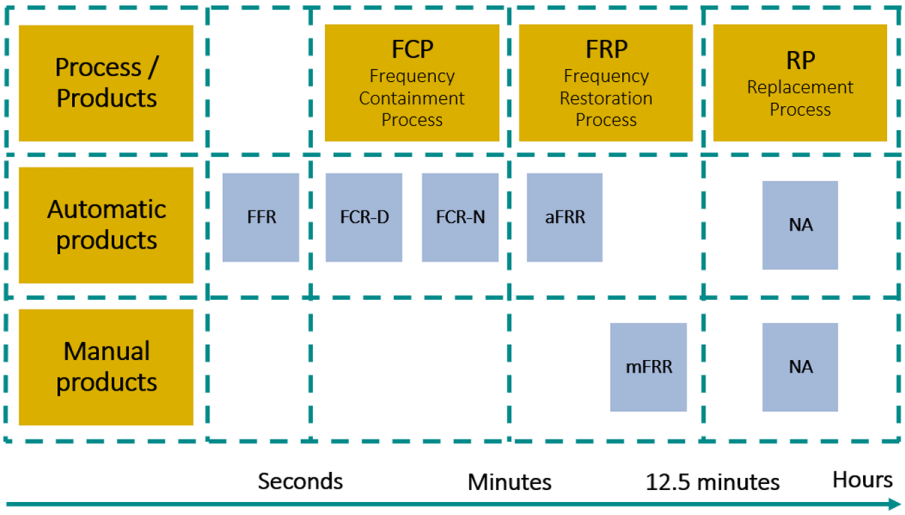


Figure 2: The Nordic balancing market hierarchy, illustrating the different reserve types and their activation times [2].

Frequency Containment Reserves (FCR) were designed to be the first line of defense against frequency deviations in the power grid. These reserves are activated automatically and respond quickly to counteract sudden imbalances between supply and demand. FCR is further divided into two categories: FCR-N and FCR-D. FCR-D is specifically intended to address frequency deviations caused by disturbances in the distribution network, while FCR-N focuses on normal operating conditions in the transmission network. FCR-D should, therefore, be able to respond faster than FCR-N to effectively manage these disturbances.

The Fast Frequency Reserves (FFR) reserve market was implemented in the Nordics in May 2020. These reserves are designed to respond even more rapidly than FCR, ideally in the span of a single second. The need for FFR arises from the increasing penetration of renewable energy sources. Wind power is, for instance, not connected synchronously to the grid, leading to a reduction in system inertia. Lower inertia means that frequency deviations occur more rapidly, necessitating faster-acting reserves like FFR to maintain grid stability. The fast power response provided by FFR is usually sustained for a short duration, stabilizing the frequency slightly before FCR-D takes over [5]. Figure 3 illustrates roughly the activation times and the interplay between the different reserve types in the Nordic balancing market.

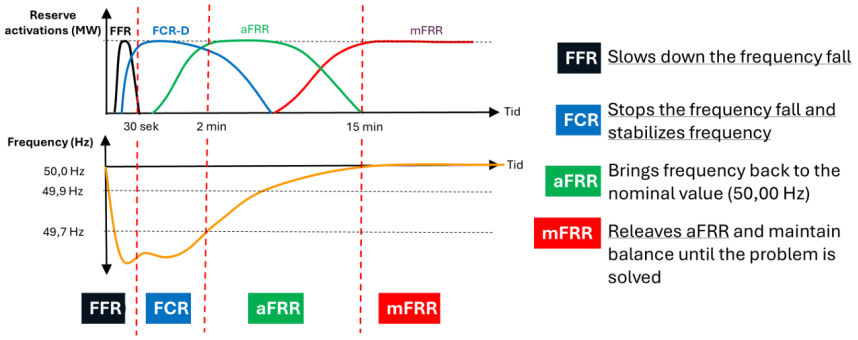


Figure 3: Illustration of the different reserve types and their activation times [2].

After FFR and FCR has stabilized the frequency, something must bring it back to its nominal level. Automatic Frequency Restoration Reserves (aFRR) holds this responsibility. These reserves are often kept activated for a couple of minutes to ensure that the frequency is restored to its normal operating level. Manual Frequency Restoration Reserves (mFRR) then relieves aFRR and maintains the balance until normal operations are restored.

2.2 Reserve Market Concepts

Reserve markets are platforms where market participants can offer their balancing services to the grid operator. These markets operate on the principle of supply and demand, where participants can bid to provide reserves at specific prices. The markets are then cleared based on the bids received, ensuring that the most cost-effective resources are utilized to maintain grid stability. Reserve markets can be broadly categorized into two types: activation markets and capacity markets. Only aFRR and mFRR markets will be discussed further, as they are the most relevant for this study.

2.2.1 Capacity Markets

Capacity is a market mechanism that ensures the availability of sufficient resources to maintain grid stability and reliability. Capacities are procured prior to real-time operation to guarantee the availability of balancing resources at the time of operation [2]. BSPs can offer their capacities to the grid operator through the Nordic aFRR capacity market or in the national (Statnett in Norway) mFRR capacity markets. Capacity market participants are compensated for making their resources available to the grid operator, regardless of whether their resources are activated or not. They are, however, obligated to deliver the offered capacity when called upon by the grid operator.

2.2.2 Energy Activation Markets (EAMs)

Activation markets operate closer to real-time and are designed to procure balancing energy to address immediate imbalances in the power grid. In the Nordic region, the aFRR and mFRR activation markets serve this purpose. In the mFRR EAM in Norway, BSPs submit bids to Statnett at least 45 minutes before the activation period. These bids specify the amount of up- or down-regulation capacity the BSP is willing to provide and the corresponding price. Statnett then forwards the bids to the clearing algorithm Nordic Libra AOF, which provides the activation volumes for the operational quarter-hour. Statnett then activates the selected bids based on the activation volumes provided by Nordic Libra AOF [6].

The mFRR EAM underwent a significant reform March 4th 2025, transitioning from national mFRR activation systems to a Nordic-wide mFRR activation market. The previously hourly resolution was also changed to a quarter-hourly resolution. This resolution change allows for more precise balancing, introducing more nuanced price signals.

2.2.3 Imbalance Prices

2.2.4 mFRR Market in Norway (will refactor this or remove this)

In Norway, the mFRR market plays a crucial role in maintaining the stability of the power grid. Market participants can offer their flexibility to the grid operator through both activation and capacity markets.

3 Literature Review

This chapter reviews literature relevant to the prediction of mFRR activations in a top-down approach. The review begins by outlining the broader context of balancing activation markets and the challenges they present. It then delves into which methodologies exist for handling these challenges, before finally identifying specific research gaps that this study aims to address.

3.1 mFRR Energy Activation Market Characteristics

Balancing markets are by nature unpredictable as their primary function is to maintain system stability in the face of unforeseen imbalances between supply and demand. This inherent unpredictability is handled in most balancing markets by the capacity market mechanism. The mFRR energy activation market distinguishes itself by only compensating participants for actual energy delivered during activation events. Participants must also bid in the correct direction of activation (upward or downward) to be eligible for activation. When an up-regulating activation is required, the up-regulation price is by design higher than the day-ahead market price, and vice versa for down-regulating activations [7]. These market characteristics make it lucrative to predict activation events accurately, as successful predictions can lead to significant financial gains.

In 2022, Klæboe et al. [7] analyzed day-ahead market bidding strategies for flexible generators taking the balancing power market into account. They found near-zero gains from incorporating balancing market predictions into day-ahead bidding strategies. They discuss, however, that the need for balancing services will increase in the future, and that such strategies will therefore become more profitable. In Svenska Kraftnät's balancing market outlook 2030 [8] they present that the mFRR capacity demand has and will steadily increase. *Source*

The mFRR energy activation market transitioned from an hourly to a 15-minute resolution as of 4th March 2025, an endeavor aimed at enhancing market efficiency and integrating renewable energy sources more effectively [9]. An hourly resolution was deemed overly discrete and can miss intra-hour dynamics; a 15-minute resolution reduces discretization and better approximates the continuous variation of supply, demand, and balancing needs. A study by Kallset and Farahmand found that increased resolution significantly reduces such structural imbalances and achieves about 60% of the possible reduction in total balancing, compared to a 5-minute resolution ideal [10]. Thus higher resolution make balancing activations align more closely with actual real- time system needs, and is less restricted by the coarse time blocks of hourly markets.

3.2 Existing Approaches to Balancing Activation Market Forecasting

This chapter reviews relevant literature on balancing market forecasting.

3.2.1 Imbalance Volume and Price Forecasting

Imbalance volume forecasting plays a central role in balancing markets [11] [12]. At the system level, transmission system operators (TSOs) rely on accurate imbalance volume forecasts to minimize balancing costs and ensure system stability. As market participants, accurate imbalance volume predictions are valuable, as this can inform short-term bidding strategies in balancing markets, particularly in markets with high renewable penetration. Consequently, a substantial body of recent research has focused on developing models that forecast system imbalance over short-horizons.

At the system-operator end of the spectrum, several studies develop point forecasting aimed primarily at optimizing TSO operations. One example is the work of Singh et al. [11], who propose a regression-based approach for short-term forecasting of imbalance volumes. They argue that increased renewable variability, combined with the 15-minute activation window, necessitates accurate short-horizon forecasts to enable transmission system operators to anticipate and manage imbalances more effectively. One of their key findings is that their best-performing model reduces

the balancing costs by 44.51% compared to TSO-based forecasting. The lower forecast errors lead to reduced costs of energy not supplied, excess energy, and corrections.

While Singh et al. obtain important results for short-term imbalance forecasting in the Belgian power system, related work examines imbalance forecasting in the Nordic region, where renewable mixes and activation patterns may differ. Edling & Azarang (2025) analyze short-term mFRR activation volumes across the four Swedish bidding zones using LSTM-based (Long Short-Term Memory) regression models [13]. Their study highlights strong geographical hetereogeneity in activation patterns, identifying structural differences between bidding zones. Price region SE2 exhibited the best forecasting accuracy, while zones SE3 and SE4 showed limited predictability due to high zero-activation frequencies. This finding spotlights an important consideration: regions with frequent imbalance, for instance due to a high share of wind power, do not necessarily have high activation frequencies. SE2 appears to have high activation frequencies due to its significant share of Sweden's flexible hydropower resources, which are often called upon to balance the system. Thus, if an imbalance occurs in SE4, the TSO may choose to activate reserves in SE2 if it is more economical and transmission constraints allow it. Overmaat's study on balancing contributions in the Nordic electricity system, though relatively dated (2019), underscores this point further by concluding that SE1 and SE2 dominate the Swedish balancing contributions on the short and medium time scales [14].

Both Singh et al. and Edling & Azarang aim to provide system operators with actionable insights and tools to better forecast and manage imbalances. While better TSO forecasting does not alter physical imbalance maggnitudes, it improves the timing and precision of activation decisions, making activation volumes more closely reflect the real-time system state. For market participants, this suggests that accurate system imbalance forecasts may offer predictive power regarding coming activations. This perspective is developed further in participant-oriented studies such as Plakas et al. [12], who propose a two-stage probabilistic framework for forecasting imbalance volumes and prices sequentially in the Greek balancing market. The first stage employs quantile regression to generate probabilistic forecasts of system imbalances. The second stage leverages the quantiles to predict imbalance prices. Plakas et al. find that system imbalance volumes are critical predictors of imbalance prices, underscoring the correlation between these two variables. By extending from imbalance volume point forecasts to price forecasting, Plakas et al. provide more actionable insights for market participants seeking to capitalize on opportunities in the balancing market. Their results show that imbalance volumes strongly influence imbalance prices, indicating that system-state indicators provide valuable information for bidders seeking to anticipate balancing-market outcomes.

3.2.2 Activation Forecasting

Whereas imbalance forecasting aims to estimate the magnitude of system imbalances, activation forecasting focuses on predicting the TSO's decision to activate upward, downward, or no mFRR energy. Activation direction is directly relevant for market participants, as balancing market bids must be placed in the correct direction to be eligible for activation. As the Nordic system transitions to a 15-minute market resolution, short-term activation forecasting has become increasingly important, yet the academic literature on this topic remains sparse.

Svedlindh and Yngveson [15] explore the general price formation in intraday and mFRR markets. Among other explorations, they develop logistic regression and ANN (Artificial Neural Network) models to predict activation direction in the mFRR activation market. The ANN model outperforms the logistic regression, achieving solid *accuracy* and *F1-scores*. They identify, however, that *class imbalance* poses a significant challenge, as no-activation events dominate the dataset. This imbalance skews model performance, making it difficult to accurately predict the less frequent upward and downward activations. Despite these acknowledged challenges, Svedlindh and Yngveson achieve above expected performance, seemingly sustained by the inclusion of mFRR capacity market cleared volumes and prices. Cleared mFRR capacity market data is not necessarily available to market participants in real-time, and thus limits the practical applicability of their models. **Is this true Jay? As far as I know, we do not have access to capacity market data (volumes, prices). They do not explain how the data is gathered.**

A recent study in 2025, Porras [16] investigated the use of XGBoost for short-term forecasting of mFRR activation direction and imbalance prices with a two-stage multi-horizon model. This paper is closely related to the present study, as it also focuses on predicting mFRR activations. However, Porras leans heavily into XGBoost, while this study explores a broader range of models. Additionally, Porras produces predictions at an hourly resolution, whereas this study aims for a finer 15-minute resolution - aligning with the recent Nordic mFRR market reform. Porras targets the SE2 price area in Sweden, thereby incorporating data from different sections of the Nordic power system. The dynamics may be similar, but this study focuses exclusively on the NO1 price area in Norway, which may exhibit different characteristics. The primal gap in Porras' study, which this study aims to address, is the opportunity opened by the recent 15-minute market reform, which provides more granular data and potentially improved prediction capabilities.

3.2.3 Methodological Distinctions Across Studies

In its most basic form, regression can be used to make *point forecasts* of future values - providing a single expected value for each time step. Regression can also be extended to produce *probabilistic distributions*, thus capturing the uncertainty inherent in balancing markets.

3.3 Research Gap

While existing literature has made significant strides in predicting balancing market activations, several gaps remain. Notably, there is a lack of studies that leverage and discuss potential impacts of the recent transition to 15-minute time intervals in the Nordic mFRR markets. This reform presents an opportunity to enhance prediction accuracy by utilizing more granular data. Additionally, while various machine learning techniques have been explored, there is a need for a comprehensive comparison of different models specifically tailored to the unique characteristics of the mFRR energy activation market. Many of the recent studies on this topic have focused on regions with high activation volumes and frequencies, such as SE2 in Sweden. However, there is a lack of research focusing on regions like NO1 in Norway, where activations are less frequent. While SE2 and other high-activation regions most likely would yield better prediction results due to more balanced classes, it is still valuable to explore the challenges and opportunities in regions with lower activation frequencies. This study aims to fill these gaps by focusing on the NO1 price area in Norway, utilizing the new 15-minute market data, and comparing a range of machine learning models to identify the most effective approaches for predicting mFRR activations.

litterature review, compare litterature, gap,

4 Problem Formulation

Note: This section is a working draft and may be removed later; it currently preserves some ideas for potential reuse. The primary objective of this study is to develop a machine learning model capable of predicting manual Frequency Restoration Reserves (mFRR) activations in the Nordic electricity market, specifically for the bidding zone NO1 in Norway. The problem is formulated as a classification task, where the model aims to classify time intervals into distinct categories based on the occurrence of mFRR activations. The classification task is approached in two stages: binary classification and ternary classification.

4.1 Binary Classification Problem

In the binary classification problem, the model predicts whether an mFRR upregulating activation occurs within a given time interval. This is a straightforward classification task with two possible outcomes: activation or no activation. This formulation is a good starting point, as it allows for exploration of the model's ability to identify patterns associated with upregulating activations specifically. Binary models are also generally easier to train and tune, as they focus on a single class of interest, thus providing a base understanding of expected model performance before moving on to the more complex ternary classification problem.

4.2 Ternary Classification Problem

In the ternary classification problem, the model extends its predictive capabilities to include both upregulating and downregulating mFRR activations. This introduces a third class to the classification task, resulting in three possible outcomes: upregulating activation, downregulating activation, or no activation. This formulation captures the full spectrum of mFRR activation dynamics, making it ideal for practical applications in the electricity market. This approach, however, is more complex than the binary classification problem, as the binary models can be more focused on the specific characteristics of upregulating activations. The ternary models, on the other hand, must learn to distinguish between three classes, which can be more challenging, possibly causing the model to be less certain in its predictions.

4.3 Restrictions

In an ideal world, a market participant would be able to choose to bid or not in real time. The mFRR activation market, however, is not ideal in this sense. Bids are accepted for time slots $t+4$ intervals into the future, i.e. one hour into the future. This greatly restricts market actors in terms of how fast they can act on recent information. In fact, at the time of bid closing for a time slot $t+4$, only mFRR activation data from $t-3$ and earlier is available. Thus there are seven 15-minute intervals of unavailable data. This data would be the most useful for the model, since activation data at time t correlates with the system state and activations close to t.

It would, for instance, be considerably easier to identify streaks of activations, as activations often occur in long contiguous sequences. In the real world scenario, data from $t-3$ and before is still useful, but can often be misleading and cannot be trusted in isolation. The dataset developed in this project investigates ways to provide the model with crucial context about the system's state of stress. The intuition is that data such as physical cross-zonal flows in and out of NO1 may correlate with NO1 mFRR activations. When many such features are combined, the model may develop complex relationships between them, enhancing its predictive performance.

Most of the available and useful data are not available in real time. For instance, consumption

and production data are often published with a delay of several hours. The model must therefore rely on features that are available in real time, or with minimal delay. Such features often take the form of forecasts, which are available ahead of time. Forecasts are inherently uncertain, but they still provide valuable information about the expected system state.

4.4 Class Imbalance

5 Data and Preprocessing

The gathered data spans a period from January 1, 2024, to December 1, 2025, providing a comprehensive view of the mFRR activation patterns over nearly two years. This extensive dataset allows for robust model training and evaluation, capturing seasonal variations and other temporal dynamics in the electricity market.

5.1 Data Sources

5.2 Data Coverage and Time Span

5.3 Nord Pool Data

Nord Pool is the leading power market in Europe, facilitating the trading of electricity across multiple countries. The data from Nord Pool includes information on market prices, trading volumes, and other relevant metrics that can influence mFRR activations. Data was obtained manually through the Nord Pool data portal as CSV files and then combined into a single dataset for analysis [17]. This tedious manual approach was necessary due to the Nord Pool API not being available for this project. Data API access requires a commercial agreement with Nord Pool, which was not feasible within the project's scope. This is not an issue for this project and research as real-time data is not required for training and evaluating the models. If the models were to be deployed in a real-time setting, however, access to real-time data through the API would be essential. The following subsections describe the specific datasets obtained from Nord Pool.

5.3.1 mFRR activation data

The primary dataset used in this study consists of manual Frequency Restoration Reserves (mFRR) activation data from the Nordic electricity market, specifically for the bidding zone NO1. This data includes accepted and activated up- and down-regulation bids at a 15-minute resolution. The activated volumes provide the target variable for the prediction models, indicating whether an mFRR activation occurred in a given 15-minute interval. Figure 4 displays the distribution of mFRR activations over the dataset period. The figure especially highlights the infrequent nature of up-activations, which occur far less often than down-activations. This phenomenon induces a *class imbalance* in the prediction task, which in general makes it more challenging for models to accurately predict the minority class [18].

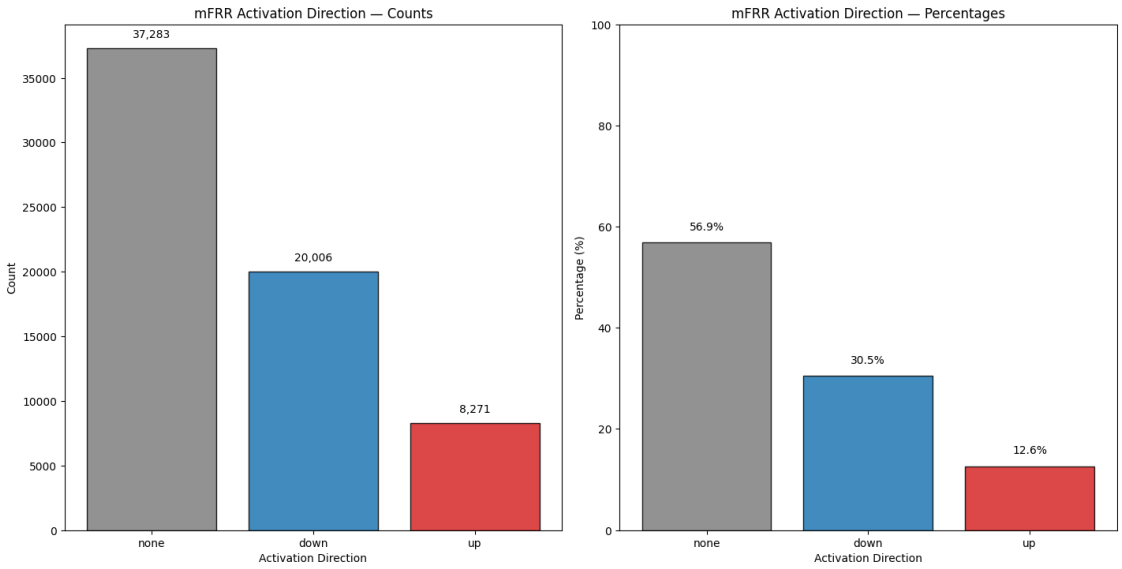


Figure 4: mFRR activation distribution.

The distribution also reveals the majority class with 56,9% of all intervals having no activation. Activations occur in 43,1% of the intervals, with down-activations being the most common at 32,5% and up-activations being the least common at 10,6%. Thus, if one were to consider the binary case of activation vs. no activation, the classes would be approximately balanced. Distinguishing between up- and down-activations, however, makes the problem more nuanced and challenging.

5.3.2 Cross-zonal flows

Cross-zonal flows refer to the electricity flows between different bidding zones in the Nordic market. In this project, only cross-zonal flows involving the NO1 bidding zone are considered: flows between NO1 and SE3, NO1 and NO3, NO1 and NO5, and NO1 and SE3. These flows are crucial for maintaining grid stability and optimizing the use of available resources. The dataset includes information on cross-zonal flows to provide additional context for mFRR activations. Figure 5 shows the distribution of cross-zonal flow directions for the NO1 bidding zone. The figure indicates that flows between NO1 and NO3, NO5, and SE3 are drastically skewed towards imports into NO1, while flows between NO1 and NO2 are reversely skewed. NO1-NO3 and NO1-NO5 show the most pronounced skewness, with very few occurrences of exports from NO1 to these zones. **This might be interesting to talk about later, as imbalance here may make the directionality more predictive.**

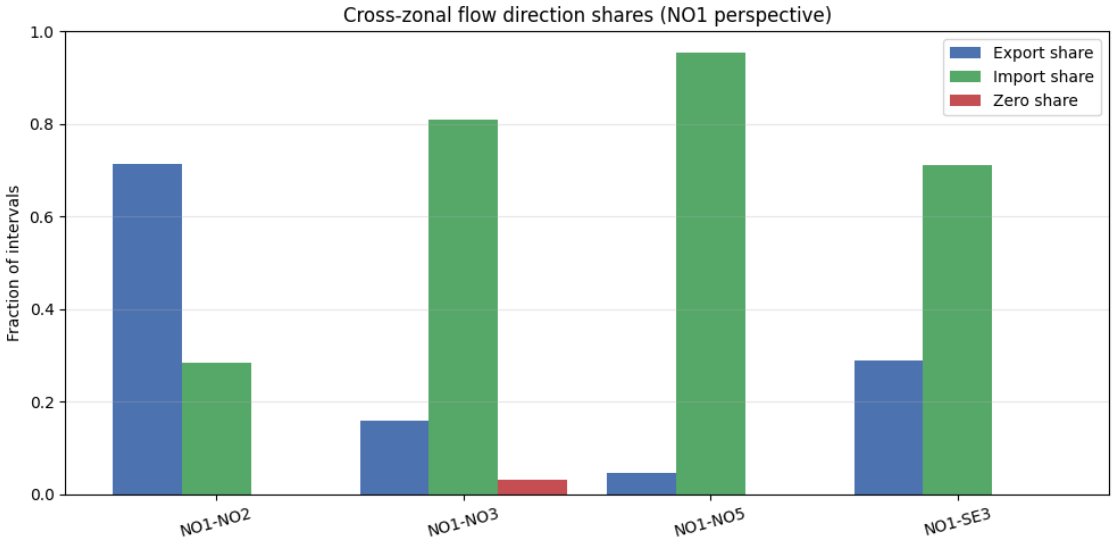
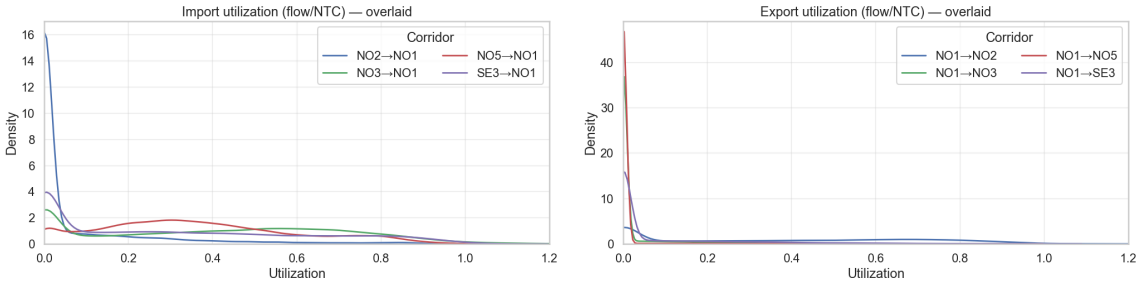


Figure 5: Cross-zonal flow distributions for the NO1 bidding zone.

Figure 6a and 6b show the relative utilization of the cross-zonal connections for NO1 as density plots. The figures illustrate how heavily the connections are utilized for imports and exports, respectively. The utilization is calculated as the ratio between actual flow and the NTC (Net Transfer Capacity) capacity of the connection. The export utilization figure highlights the lack of exports from NO1, except for the NO2 connection, indicated by the tail thickness between 0.4 and 1.0 utilization. The import utilization figure, on the other hand, shows that all connections, except NO2-NO1, are heavily utilized for imports, with many thick tails approaching full utilization. The NO1-NO2 connection is thus almost exclusively used for exports from NO1 to NO2, whilst the other connections are primarily used for imports into NO1.



(a) Import utilization.

(b) Export utilization.

Figure 6: Cross-zonal flow utilizations for the NO1 bidding zone calculated as the ratio between actual flow and NTC capacity.

5.3.3 Load and production data

Load and production forecasts provide insights into the expected system state. Forecast may on their own provide valuable information about potential mFRR activations, but when combined with actual load and production data, the model can learn to identify discrepancies between expected and actual system states. Such discrepancies often lead to imbalances that require mFRR activations to restore balance. The different production sources (e.g., hydro, wind, thermal) have varying characteristics and impacts on grid stability. Among them, wind power is particularly relevant due to its intermittent nature, which can lead to sudden changes in generation levels. Wind power production data is therefore predicted to have the biggest impact on mFRR activations among the different production types.

Table 1: Summary statistics for wind-related features (2024–2025, NO1)

Metric	Mean	Std	Min	P10	P50	P90	Max	Count
Wind DA Forecast	121.64	96.34	0.0	16.0	95.0	274.0	370.0	62,680
Wind Intraday Forecast	135.80	104.55	0.0	15.0	111.0	294.0	376.0	38,972
Wind Actual Production	120.38	102.72	0.0	8.0	91.0	283.0	380.0	65,568
Wind Revision (ID–DA)	12.59	15.27	0.0	1.0	7.0	30.0	151.0	38,972
DA–Actual Error	0.03	38.85	-250.0	-44.0	-2.0	47.0	221.0	62,680
ID–Actual Error	2.97	35.30	-227.0	-38.0	1.0	46.0	189.7	38,972
Abs DA Error (%)	0.58	0.99	0.0	0.038	0.24	1.36	5.0	61,486
Abs ID Error (%)	0.42	0.75	0.0	0.032	0.19	0.92	5.0	38,370
Wind Share	0.054	0.047	0.0	0.0036	0.040	0.128	0.228	65,568

Load/consumption data is much simpler in nature, as *who* or *what*, essentially the source of consumption, is not as relevant as the source of production. Consumption forecasts and actual consumption data can still be useful, however, as sudden changes in consumption patterns can lead to imbalances that require mFRR activations.

5.4 NUCS

NUCS, or the Nordic Unavailability Collection System, is a service for collection of data on unavailable data in the Nordic power system. NUCS is an important part of this project, as it provides otherwise unavailable data that served as features in the models. NUCS is unique from the other data sources used in this project, as it provides data through an API (Application Programming Interface) [19]. This allows for automated data retrieval, which is especially useful for real-time applications. However, as this project does not have access to comprehensive real-time data, the NUCS API was only used to gather historical data for the training and evaluation of the models.

5.4.1 aFRR procurement prices

aFRR procurement data is generally unavailable. However, through the NUCS API, historical aFRR procurement prices and volumes for the NO1 bidding zone were obtained. This data provides insights into the costs associated with aFRR activations and can be used as additional features in the models.

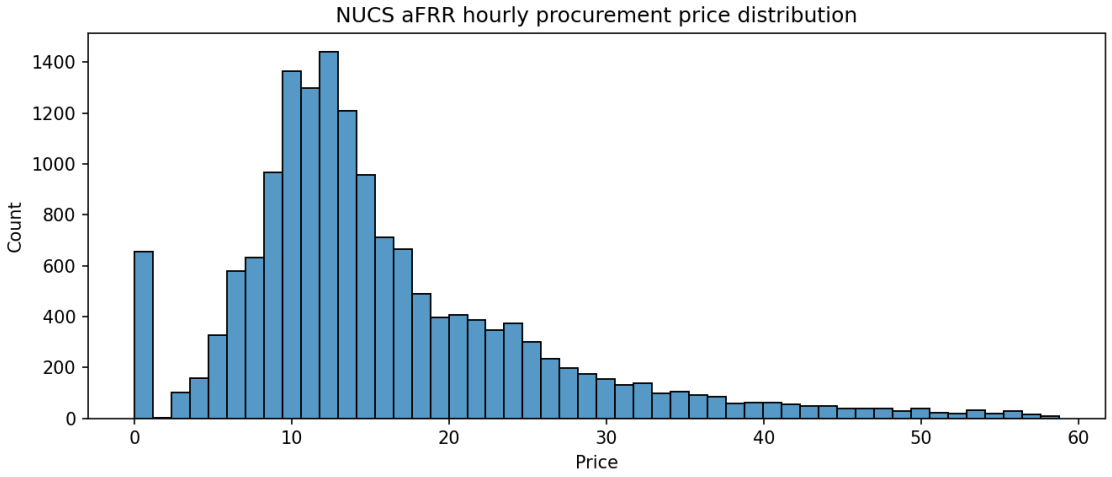


Figure 7: A histogram of hourly aFRR procurement prices for the NO1 bidding zone from NUCS data.

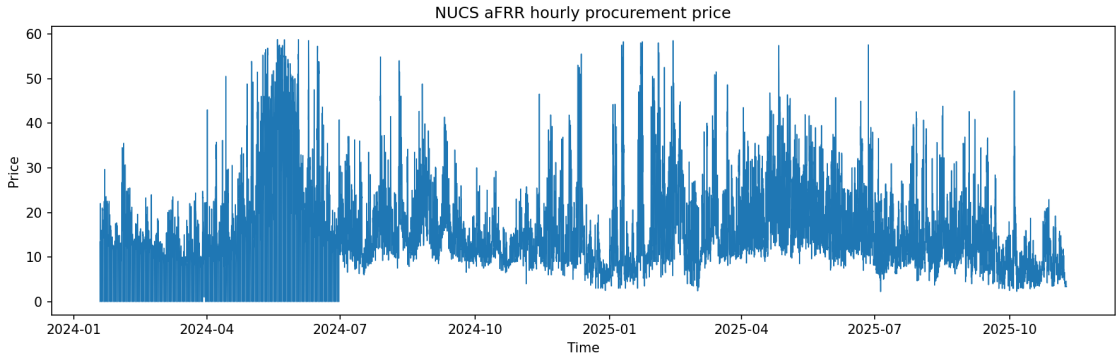


Figure 8: Hourly aFRR procurement prices for the NO1 bidding zone from NUCS data.

Figure 7 and 8 show a histogram and time series plot of the hourly NO1 aFRR procurement prices between January 1, 2024, and 1. December 2025. Both figures highlight a problem in the data before July 2024, where there are many missing values represented as zeros. The data appears complete after this date. Intermittent missing values can be handled during preprocessing in multiple ways. In time series data, simply removing the rows with missing values is not an option, as it would break the time continuity. One approach is to use interpolation to estimate the missing values based on the surrounding data points. Another approach is to use forward or backward filling, where missing values are filled with the last known value or the next known value, respectively. The choice of method depends on the nature of the data and the specific requirements of the analysis or model being used.

5.5 ENTSO-E

ENTSO-E, the European Network of Transmission System Operators for Electricity, is a key organization in the European electricity market. ENTSO-E provides a wide range of data related to electricity generation, consumption, and grid operations across Europe [20].

5.5.1 aFRR Activation Data

aFRR prices and capacities are available through NUCS as discussed earlier, but aFRR activation data is not. However, ENTSO-E provides detailed data on aFRR regulations via their "Accepted Offers and Activated Balancing Reserves" dataset. This dataset is supposed to provide information about up and down regulations from all balancing markets. Only aFRR data seems to be available, however, which is sufficient for this project. The data is available at an hourly resolution, which is coarser than the 15-minute resolution of the mFRR data. This data is, in the same manner as other 1-hour resolution data, resampled with forward filling to match the 15-minute resolution of the main dataset.

Crucial: Time restriction. (Probably move parts of this to somewhere else as these concepts are important.) aFRR activation data is naturally not available at the target time $t + 4$ and prediction time t as these activations has not yet occurred or finished. The 1-hour resolution is problematic as it essentially causes the data to become available in clusters of four 15-minute intervals at a time. This causes inconsistencies in the what specific lag features are available at different prediction times. For instance, assume data becomes available immediately after the hour. Then, at time $t = 10{:}00$, the last available aFRR activation data would be from 09:00-10:00. This means that lag features for $t - 1$, $t - 2$, $t - 3$, and $t - 4$ would be available. However, in the worst case, at a prediction time $t = 10{:}45$, the last available aFRR activation data would still be from 09:00-10:00. This means that only lag features starting at $t - 4$ would be available, while $t - 1$, $t - 2$, and $t - 3$ would not. This inconsistency in available lag features

complicates the model training and evaluation process. There are multiple ways to handle this issue:

- **Option 1:** Only use lag features that are always available, i.e., only use lag features starting from $t-4$ and further back. This is the easiest option to implement but it sacrifices potentially valuable information from more recent time steps.
- **Option 2:** Create separate models for different prediction times within the hour (e.g., one model for predictions at 00, 15, 30, and 45 minutes past the hour). Each model would be trained with the each with their own set of available lag features. This approach maximizes the use of available data but requires maintaining multiple models. The more gathered datasets are of a 1-hour resolution, the more valuable this approach becomes.
- **Option 3:** Use imputation techniques to estimate the missing lag features based on available data. This approach allows for a single model to be used but introduces uncertainty due to the imputation process.

5.6 Dataset structure

The data is represented as a time series, where each record in the dataset consists of a set of attributes connected to one point in time. More specifically, the data contains a sequence of 15-minute interval time stamps. Each time stamp may or may not have an associated activation, which is the target variable the model is trying to predict. The features describe the system state at that time stamp, providing context for the model to learn from.

6 Feature Engineering

Features are attributes in a dataset that describe each data point. A dataset for predicting a person's income could, for instance, have features like gender, job type, and age. Then, each data point represents a person and relevant information about the person in terms of the target variable – income. In this project, each data point represents a specific time stamp in the mFRR activation dataset, and the features describe the system state at that time. This includes information such as electricity demand, generation capacity, and market prices, all of which can influence mFRR activations.

Already available features can be transformed to create new higher-level features that may better capture the underlying patterns in the data. For example, if one has features for year of death and year of birth, a new feature for age at death can be created by subtracting the year of birth from the year of death. This is known as *feature construction* [21]. Features can be constructed in various ways, such as through mathematical operations or aggregations. This project leverages this concept to create new features that may enhance the model's predictive capabilities.

Feature selection is crucial. There are many features that may seem useful and relevant in isolation, but sometimes they mislead the models, or they work poorly in combination with other seemingly good features. Theoretical analysis of the usefulness of certain features can be helpful, but only trial-and-error together with feature-importance analysis will uncover the features' actual impact on performance. Feature selection will be subject to restrictions outlined in Section 4.3 to ensure that only real-time available features are used.

6.1 Lag features

A lag feature is a feature that represents the value of a variable at a previous time step. Lag features are commonly used in time series analysis to capture *temporal* dependencies and trends in the data [22]. By including lag features, the model can leverage historical information to make more informed predictions about future mFRR activations. For instance, if there was an upregulating activation in the previous time step, it may indicate a higher likelihood of another upregulating activation in the current time step.

Activation lag features Activation lag features are the most important lag features for this problem, as they convey important information about recent temporal activation trends. They are, however, restricted by the real-time limitations, so the model may only use activation lag features from $t-3$ and earlier for predicting activations at time $t+4$. As a result, lag features for upregulating and downregulating activations are created for time steps $t-3$, $t-4$, $t-5$, ..., $t-9$. These features are useful by themselves, but they also serve as a basis for creating other features that capture activation trends more effectively.

Persistency (streak length) Let t denote the reference time such that the model predicts activation at $t+4$. At bid close, only activation data up to and including $t-3$ are available. Define binary activation indicators $A^{\uparrow}(\cdot)$ and $A^{\downarrow}(\cdot)$ for up- and down-regulating activations, respectively. The up- and down-persistency features are the lengths of the most recent consecutive activation runs ending at $t-3$: **TODO: Consider only formulating persistency with words, as the mathematical expression is more complex than the concept.**

$$S^{\uparrow}(t) = \sum_{j=0}^{\infty} \prod_{i=0}^{j} A^{\uparrow}(t-3-i), \qquad S^{\downarrow}(t) = \sum_{j=0}^{\infty} \prod_{i=0}^{j} A^{\downarrow}(t-3-i).$$

Here, j is the (zero-based) window length and i is the offset index within that window. The product is 1 only while all the last $j+1$ values are 1, so when a 0 is encountered, the product becomes 0 for all larger j. Thus, the sum counts how many consecutive intervals back from $t-3$ remain all ones before the first zero. In words, $S^{\uparrow}(t)$ (resp. $S^{\downarrow}(t)$) counts how many consecutive up- (resp. down-)

activations occurred immediately before $t-3$. If there was no activation at $t-3$, the corresponding streak length is 0.

Design choice: separate vs. signed. When designing the persistency features, a choice had to be made between using two separate non-negative integer features or a single signed feature (e.g., $S^{\uparrow}(t) - S^{\downarrow}(t)$). Separate features avoid conflating direction with magnitude and let the model learn asymmetric effects. This option may be easier for the model to interpret, as it provides a clearer distinction between up- and down-regulating activations. The signed variant is more compact, but it may introduce ambiguity in how the model interprets the values. It is, however, important to consider the possibility of persistency overreliance. If the model relies too heavily on persistency features, it may overlook other important factors influencing activations. This could lead to suboptimal predictions, especially in scenarios where activation patterns change. Only having one persistency feature may reduce this risk, as persistency is encoded in a single column. Both options will be explored and evaluated during model development, and the pros and cons of each approach will be assessed based on empirical performance.

Importance. Persistency features are perhaps the single most important features in the dataset. They provide compact and direct information about recent activation trends, making it easier for the model to identify activation streaks. As activations often come in long contiguous runs, and even though there are real-time limitations, the persistency features still provide the model with valuable context. The longer the streak, the more likely it is that the activation trend will continue. In some cases, the model might be able to predict an activation solely based on a high persistency value, without needing to consider other features.

Much of the motivation behind this project is, as outlined in Section 4, to develop models that can predict activations based on system state features rather than just relying on past activation trends. The reality is, however, that persistency features are extremely dominant, and they may overshadow the contributions of other features.

6.2 Cross-zonal flow features

Cross-zonal flow features capture information about electricity flows between different zones or regions in the power grid - in this case, in and out of the NO1 bidding zone. These flows can indicate the grid's stress level and influence mFRR activations. For instance, high inflows into NO1 may signal increased demand or generation shortages, potentially causing upregulating activations. Conversely, high outflows may indicate surplus generation, potentially causing downregulating activations. It is unlikely, though, that cross-zonal flow features alone can predict activations. Combining them with available transfer capacity provides a picture of how close the grid is to its operational limits. For instance, if the inflow into NO1 is close to the maximum available transfer capacity, only a small margin remains for additional inflows, which could increase the likelihood of upregulating activations. Such situations often occur in zones that are short, i.e. zones where consumption exceeds production. In such cases, the grid operator may need to activate expensive mFRR reserves to maintain grid stability when no more cheap imports are possible. It is important that the models developed in this project are able to capture these kind of relationships as they are among the most valuable for a potential user of the models.

Capacity-normalized cross-zonal flow. Raw cross-zonal flow magnitudes are not comparable across interconnections or over time because each line has different capacity and the available transfer capacity (ATC) varies. The same absolute flow can be insignificant on a strong interconnection but critical on a constrained one. To obtain a dimensionless, capacity-normalized measure of proximity to operational limits—and to make features comparable across borders and time—flows are expressed as a ratio to the relevant directional ATC.

Let $F_i(t)$ be the *signed* flow for interconnection i at time t (positive into NO1, negative out of NO1). Let $ATC_i(t) \geq 0$ denote the available transfer capacity magnitude used for normalization

(e.g., a symmetric ATC for interconnection i at time t). The capacity-normalized ratio is then

$$F^i_{\text{ratio}}(t) \;=\; \frac{F_i(t)}{ATC_i(t)}.$$

Hence $F^i_{\text{ratio}}(t) \in [-1, 1]$ when flows are within limits, approaching $+1$ as inflow nears the capacity and -1 as outflow nears the capacity.

6.3 Temporal features

Temporal features capture time-related patterns in the data. These features help the model understand how mFRR activations vary with time, such as daily or weekly cycles. Basic temporal features include hour of the day, day of the week, and month of the year. These features allow the model to learn patterns related to specific times. mFRR activations could, for instance, be caused by completely different factors during peak hours on weekdays compared to off-peak hours on weekends. Temporal features like these are most often represented using cyclical encoding to reflect their periodic nature. For example, 1 AM and 11 PM are close in time, even though their numerical representations (1 and 23) are far apart. Cyclical encoding uses sine and cosine transformations to capture this periodicity [23]. Hourly features are, for instance, encoded as:

$$\text{Hour}_{\sin} = \sin\left(2\pi \cdot \frac{\text{Hour}}{24}\right), \qquad \text{Hour}_{\cos} = \cos\left(2\pi \cdot \frac{\text{Hour}}{24}\right).$$

Monthly features can be encoded similarly, using 12 as the divisor instead of 24.

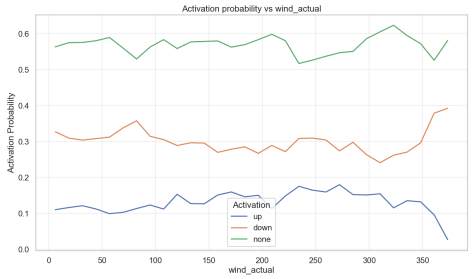
6.4 Price features

Price features capture information about the various electricity market prices. The mFRR activaiton market is closely linked to other electricity makets, such as the day-ahead market, the intraday market, and the aFRR market. Most prices may not have direct impacts on activations, but by crafting features that capture important relationships between prices, the model may be able to infer system stress levels that could lead to mFRR activations. Large discrepencies between day-ahead prices and intraday prices may, for instance, indicate unexpected changes in supply or demand, which should correlate with mFRR activations. Similarly, the difference between aFRR prices and mFRR prices may provide insights into the relative costs of balancing services, which could influence activation decisions.

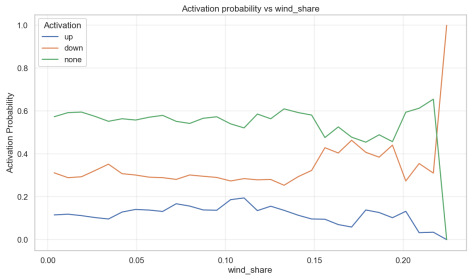
6.5 Load features

6.6 Production features

Production features capture information about electricity generation, particularly from renewable sources like wind power. Wind power production features were considered promising candidates for predicting mFRR activations, as wind power is intermittent and can cause sudden changes in supply. figures 9a and 9b show values of realized wind production and wind share (wind production as a fraction of total production) plotted against the distribution of mFRR activations. These figures indicate that there is no direct correlation between wind production and mFRR activations. The existence of such a correlation would have made it easy for the model to leverage wind production features for predicting activations. The hope is, however, that wind production features will prove useful when combined with other features, as the model captures complex relationships between features.



(a) Realized wind production plotted against mFRR activation distribution.



(b) Forecasted wind production plotted against mFRR activation distribution.

Figure 9: Load and production data distributions.

6.7 Interaction features

Interaction features are created by combining two or more existing features to capture complex relationships that may influence mFRR activations.

7 Methodology

This section delves into the methodologies and techniques employed in this study.

7.1 Problem Setup

Explain how the problem is constructed in code - this will probably be quite in-depth and technical.

7.2 Metrics

Classification problems are often evaluated using accuracy, precision, recall, and F1-score. These metrics are defined as follows [24]:

- **Accuracy**: The ratio of correctly predicted observations to the total observations. It is calculated as:
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$
where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.
- **Precision**: The ratio of correctly predicted positive observations to the total predicted positive observations. It is calculated as:
$$\text{Precision} = \frac{TP}{TP + FP}$$
- **Recall**: The ratio of correctly predicted positive observations to all observations in the actual class. It is calculated as:
$$\text{Recall} = \frac{TP}{TP + FN}$$
- **F1-score**: The weighted average of Precision and Recall. It is calculated as:
$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

While accuracy is a commonly used metric, it can be misleading in cases of imbalanced datasets. For instance, if only 5% of the data points belong to the positive class, a model that always predicts the negative class would achieve 95% accuracy, but would be useless for identifying positive cases. In such scenarios, precision, recall, and F1-score provide a more nuanced evaluation of model performance, especially in applications where the costs of false positives and false negatives differ significantly. In the context of mFRR activation prediction, false negatives (failing to predict an activation) may lead to missed opportunities for market participation, while false positives (predicting an activation when there isn't one) could result in unnecessary costs or penalties. Therefore, a balanced consideration of these metrics is essential for developing an effective classification model.

Precision-recall trade-off . This project primarily focuses on maximizing the F1-score, as it balances precision and recall, providing a comprehensive measure of the model's performance in predicting mFRR activations. Accuracy is essentially neglected due to the imbalanced nature of the dataset. Between precision and recall, recall is slightly prioritized, as missing an activation prediction is considered more detrimental than a false alarm in this context. Precision can also be optimized more easily as the confidence threshold can be adjusted post-training to favor precision over recall or vice versa. To achieve a high recall score, the model must be capable of identifying as many actual activations as possible, even if it means occasionally predicting an activation when there isn't one. The model must be gutsy and attempt to identify patterns that indicate an upcoming activation, not just follow recent activation history. This approach aims to ensure that the model has practical utility in real-world market participation scenarios. If the model had no such pattern recognition capabilities, it would be of little use beyond simple statistical analysis of recent activation trends, which could be performed without machine learning.

There is, however, a limit to how much recall can be prioritized. If the model predicts an activation for most time intervals, it will achieve a high recall but at the cost of precision, rendering it ineffective. Therefore, the model must strike a balance, ensuring that it is both sensitive to actual activations and specific enough to avoid excessive false positives. This balance is crucial for the model's success in practical applications, where both types of errors have significant implications. The exact precision-recall trade-off can be adjusted based on the specific use-case. Three potential cases are outlined below:

- **Case 1 - Recall-focused**: A recall-focused approach can be useful for analysis purposes, where the goal is to identify periods of increased risk for activations. In this case, the model can be optimized to achieve a high recall score, even if it means sacrificing precision. Such a model would be valuable for understanding the conditions that lead to activations, but may not be suitable for direct market participation due to the high number of false positives. This approach might even be the best for market participation as the ability to discover more activations could outweigh the costs of false positives.
- **Case 2 - Balanced approach**: For general applications, maintaining a balance between precision and recall is often desirable. The model can be optimized to achieve a high F1-score, ensuring that both metrics are adequately addressed. This involves fine-tuning the model's parameters and threshold settings to find an optimal trade-off. Ideally, this approach would be used, achieving good performance in both precision and recall, making the model versatile for various applications, including market participation. Achieving such results is, however, quite challenging as either precision or recall often needs to be sacrificed to some extent to improve the other.
- **Case 3 - High precision focus**: In situations where false positives carry significant costs, the model can be adjusted to prioritize precision. This may involve raising the confidence threshold for predicting an activation, reducing false positives but potentially missing some true activations. For multi-market actors, such a model is useful, as they can afford to be selective about which market to participate in, only bidding into activation markets when the model is very certain of an upcoming activation.

In this project, the focus is primarily on Case 1 and Case 2, with an emphasis on achieving a high F1-score while slightly prioritizing recall. This approach aims to ensure that the model has

a chance to give users novel insights into mFRR activations, potentially providing a competitive edge in market participation.

The transition metric is one more evaluation metric which is important to consider for this specific problem: how often the model manages to predict the start of an activation streak. This is measured by looking at all sequential time interval pairs where the first interval did not have an activation, but the second one did. If the model predicted an activation for the second interval, it is counted as a successful prediction *transition*, which is how this metric will be referred to. This metric is important because it concretizes the model's ability to catch the onset of activation periods, which is the most valuable aspect for market participants. If the model can often enough predict these transitions, it can provide significant strategic value, even if its overall precision and recall are not perfect.

Probability correctness. Although the models make hard classifications, they do so based on predicted probabilities for each class. It is, therefore, useful to evaluate how well these predicted probabilities align with actual outcomes. For example, if the model predicts an activation with a probability of 0.8 and an activation does not occur, this should be considered a more severe error than if the model falsely predicted an activation when the predicted probabilities were more evenly distributed.

7.3 Data Splitting

Proper data splitting is crucial for evaluating the performance of machine learning models. In this study, the dataset is divided into training, validation, and test sets based on temporal order to prevent data leakage and ensure that the model is evaluated on unseen data. The training set is used to train the model, the validation set is used for hyperparameter tuning and model selection, and the test set is reserved for the final evaluation of the model's performance. **Explain more.**

7.4 Model Selection

Many different models were considered and tested during the project. The main models that were evaluated include Random Forest, Extra Trees, LightGBM, XGBoost, CatBoost, and neural network models implemented using AutoGluon. Each model has its strengths and weaknesses, and their performance can vary depending on the specific characteristics of the dataset and the problem at hand. **Explain more.**

7.4.1 Random Forest and Extra Trees

Random Forest and Extra Trees are ensemble learning methods that combine multiple decision trees to improve predictive performance and reduce overfitting.

7.4.2 CatBoost

CatBoost is a gradient boosting algorithm that was found to perform well with particular feature sets and hyperparameter configurations [25]. Random Forest and Extra Trees were generally favoured throughout the project, but CatBoost provided competitive results in some scenarios. Since

7.4.3 XGBoost and LightGBM

Need to check this out - haven't really given them a real chance with hyperparameter tuning and such yet, as they are suspected to not be ideal for contiguous time-series data.

7.5 Adjusting Classification Thresholds

A critical aspect of optimizing classification models involves adjusting the decision thresholds to balance precision and recall effectively. It is often useful to find the threshold that maximizes the F1-score, which is the harmonic mean of precision and recall. This approach ensures that both false positives and false negatives are minimized, leading to a more balanced model performance. This was implemented by evaluating the model's performance across a range of thresholds and selecting the one that yielded the highest F1-score. A typical precision-recall curve is shwon in figure 10. During training and model evaluation, this threshold was heavily utilized as a general indicator of the model's performance. This streamlined the evaluation process, making it easier to compare different models and configurations based on a single metric. It is important to note that while maximizing the F1-score provides a balanced approach, this is not necessarily the optimal strategy threshold for all applications. Depending on the specific use case, one might prioritize either precision or recall more heavily.

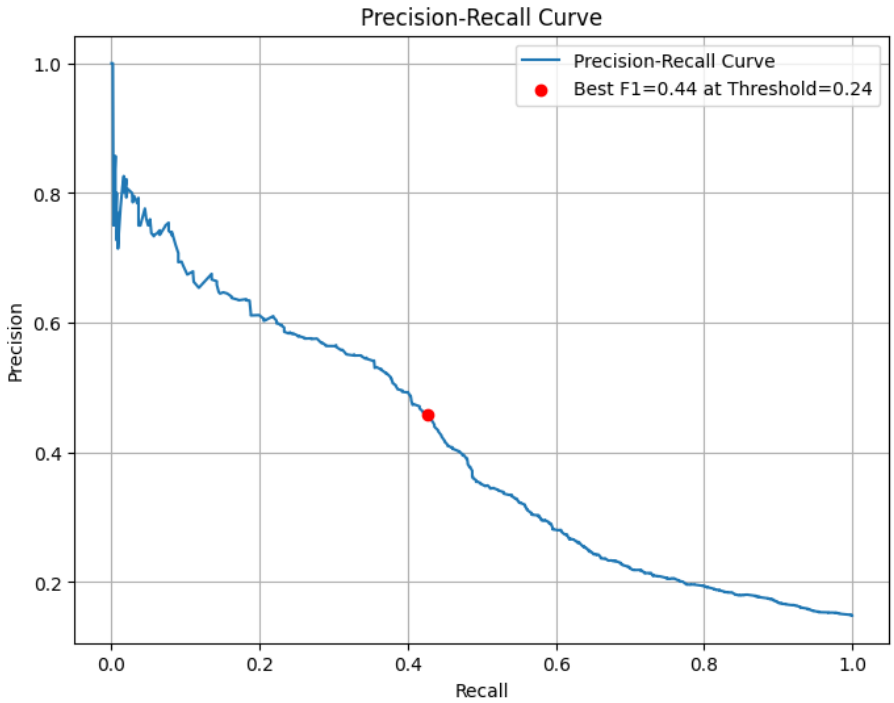


Figure 10: Typical Precision-Recall Curve with F1-score maximization point indicated.

7.6 Decision-bias tuning

Simply adjusting the classification threshold to maximize F1-score is simple in a binary classification setting, but in a ternary classification setting, where there are three classes, such a straightforward approach is not possible. The 'up' class is, as discussed in section 4.4, a heavily imbalanced class. Imbalanced classes are often more difficult for models to learn, as crucial information about the minority class is more sparse. After the switch from binary to ternary classification, it was observed that the model's performance on the 'up' class degraded significantly. To mitigate this, a decision-bias tuning approach was implemented.

Decision-bias tuning is very similar to adjusting classification thresholds, as both methodologies tune the model's decision-making process post-training. Decision-bias tuning specifically focuses on modifying the decision boundary for the 'up' class to improve its F1 score.

Class-specific decision bias for "up". Given predicted class probabilities $p_c(x)$ for $c \in \{\text{up}, \text{down}, \text{none}\}$, we adjust the score for the "up" class by a multiplicative factor $\alpha > 1$:

$$p'_{\text{up}}(x) \;=\; \alpha\, p_{\text{up}}(x), \qquad p'_k(x) \;=\; p_k(x) \;\text{ for } k \neq \text{up},$$

and predict

$$\hat{y}(x) \;=\; \arg\max_c \; p'_c(x).$$

For $\alpha > 1$, the effective decision threshold for predicting "up" is lowered:

$$p_{\text{up}}(x) \;\geq\; \max_{k \neq \text{up}} \frac{p_k(x)}{\alpha}.$$

Validation-based selection of α. Let $\mathcal{A}$ be a finite candidate set (e.g., $\mathcal{A} = \{1.0, 1.25, 1.5, 2.0, 3.0\}$). We select α by grid search on a validation set $\mathcal{D}_{\text{val}}$ to maximize the F1 score for the "up" class:

1. For each $\alpha \in \mathcal{A}$:
   (a) Compute adjusted scores $p'(x; \alpha)$ by multiplying $p_{\text{up}}(x)$ by α and leaving other classes unchanged.
   (b) Predict $\hat{y}(x; \alpha) = \arg\max_c p'_c(x; \alpha)$ for $x \in \mathcal{D}_{\text{val}}$.
   (c) Compute $\text{F1}_{\text{up}}(\alpha)$ by treating "up" as positive and $\{\text{down}, \text{none}\}$ as negative.
2. Choose
$$\alpha^* \;\in\; \arg\max_{\alpha \in \mathcal{A}} \; \text{F1}_{\text{up}}(\alpha).$$
3. Use α^* at inference time to adjust $p_{\text{up}}(x)$ before the final arg max.

Intuition and benefits. This method effectively lowers the decision threshold for predicting the "up" class, making the model more inclined to predict "up" when there is uncertainty. This will improve recall for the "up" class, at the cost of precision. By selecting α based on F1 score, we ensure that the trade-off between precision and recall is optimized for practical performance.

7.7 AutoGluon

8 Model Results

During the project, various models were trained on different dataset iterations as new data was added and features were engineered and refined. Models were initially trained on data from 2025 only. This dataset served as the starting point, excluding data from 2024 for simplicity and to establish a baseline. Using a one-year dataset is inherently problematic due to the limited amount of data and the fact that the model can not learn from seasonal patterns. It is also likely that the model overfits to specific events in the year, which may not generalize well to other years. Later, data from 2024 was included to provide more training data and to allow the model to learn from seasonal variations. A two-year dataset is, however, still problematic, as it only captures two instances of seasonal patterns, which is likely to not be representative of previous or coming years. Ideally, the dataset would span multiple years to capture a wider range of seasonal patterns and outliers, but this was not possible due to data availability constraints.

The problem started as a binary classification problem, where the model's task was to predict whether an up-activation would occur in a given hour or not. However, as the project progressed,

it was expanded to also include the prediction of down-activations, making it a multi-class classification problem with three classes: up-activation, down-activation, and no activation. The results presented here touch on both the binary and multi-class classification problems, but focuses primarily on the multi-class problem, as it is the most comprehensive and relevant to the real-world application.

8.1 The 2025 Dataset

Depending on the specific metrics used for evaluation, different models perform best. It was quite early on discovered that Random Forest and Extra Trees models performed best overall, so the results focus on these two models. The models were trained and evaluated based on F1-score, with raw precision and recall also considered after training. The aforemontentioned transition metric is also used to evaluate the models, as it provides insight into the model's ability to predict activation transitions specifically.

8.1.1 Extra Trees

Figure 11 shows the Precision-Recall curve for an Extra Trees model trained on the 2025 dataset. Three dots are scattered on the curve, representing the precision-recall pairs for three interesting thresholds: the threshold given by the predictor leaderboard, the threshold that maximizes F1-score, and the threshold that gives a recall of 0.5. It is interesting that the leaderboard threshold gives a lower F1-score than the maximum F1-score threshold after training, as one would expect the leaderboard threshold to be optimal. Nevertheless, all three thresholds provide solid performance, with the maximum F1-score threshold achieving the best balance between precision and recall. The threshold that gives a recall of 0.5 sacrifices some precision to achieve higher recall, while simultaneously improving the transition metric.

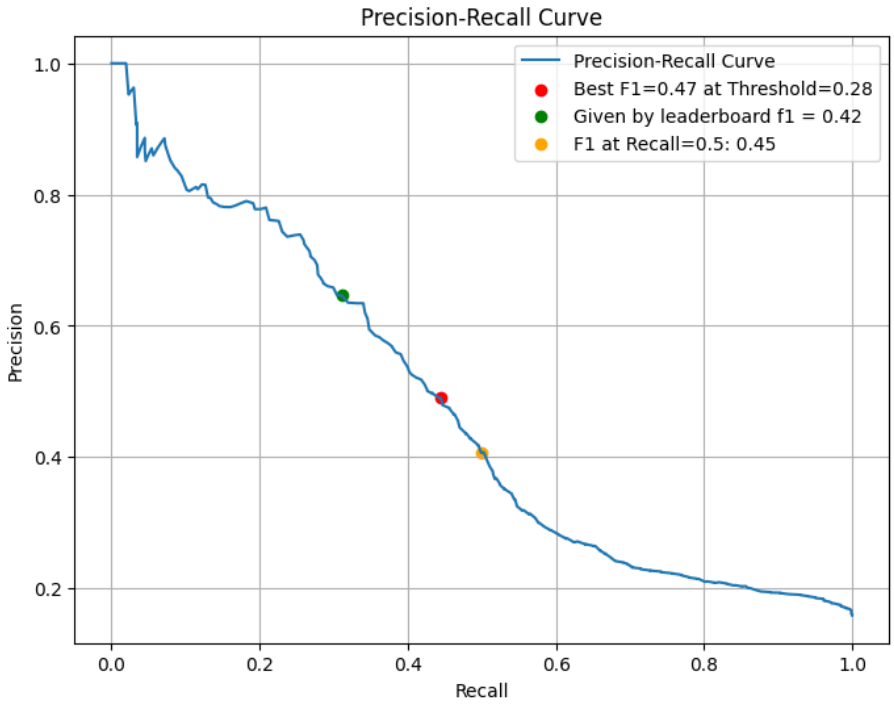


Figure 11: Precision-Recall Curve for Extra Trees model trained on 2025 dataset with highest F1-score.

Threshold	Precision	Recall	F1-score	Transition Metric
Leaderboard	0.65	0.31	0.42	9.49%
Max F1-score	0.49	0.45	0.47	19.76%
Recall = 0.5	0.41	0.50	0.45	28.06%

Table 2: Performance metrics for Extra Trees model on 2025 dataset at different thresholds.

Table 2 summarizes the performance metrics for the Extra Trees model on the 2025 dataset at the three different thresholds. The transition metric is highly correlated with recall, as expected, since higher recall means more activation events are correctly identified, leading to better transition detection. The recall-focused threshold achieves a transition success rate of 28.06%, significantly higher than the threshold proposed by the leaderboard, and 10% higher than the maximum F1-score threshold. With a precision of 0.41 at this threshold, the model is relatively precise, not overwhelmingly predicting activations, which is crucial.

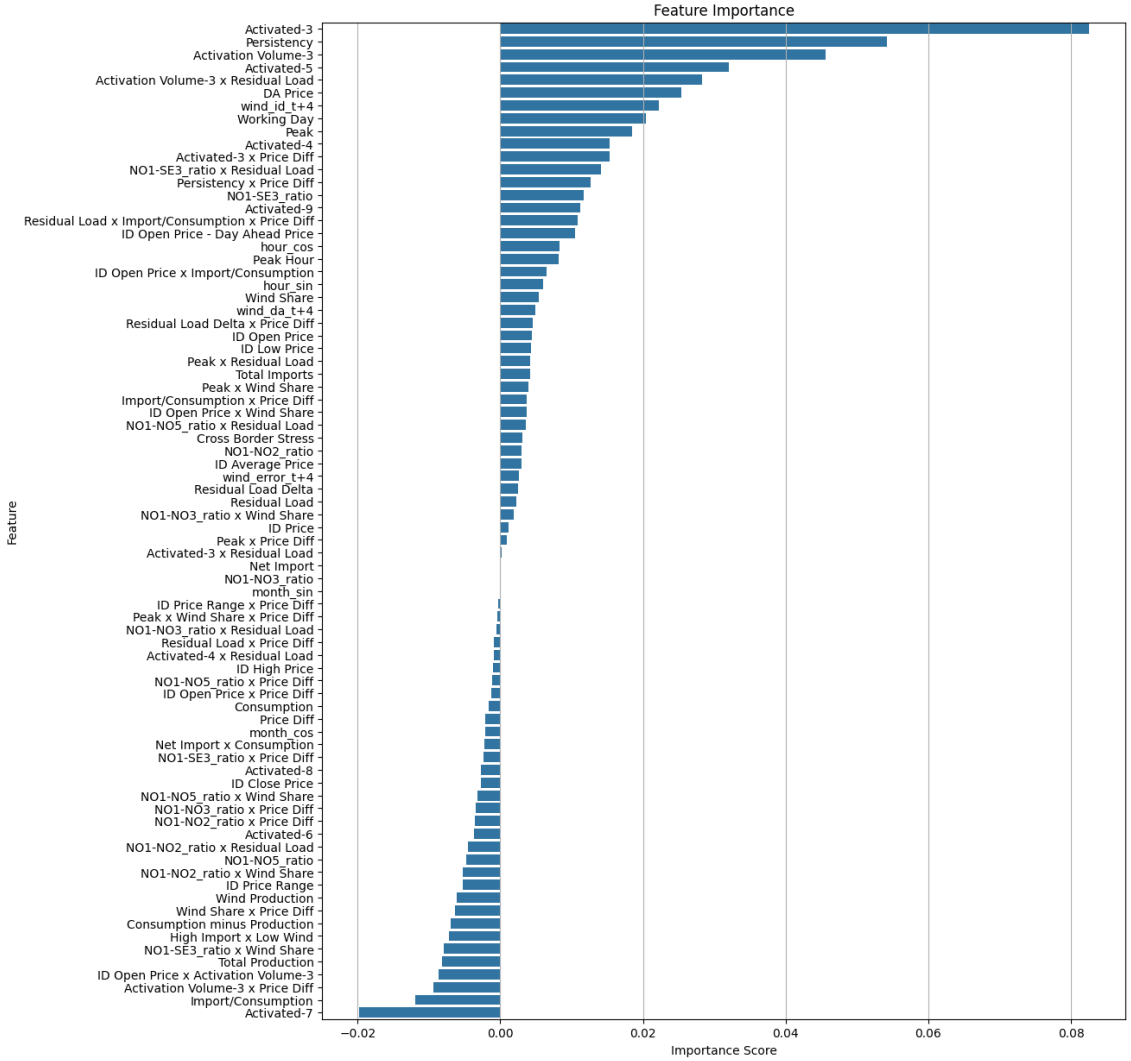


Figure 12: Feature importance for Extra Trees model trained on 2025 dataset with highest F1-score.

Figure 12 shows the feature importance for the Extra Trees model. An important concept to note is that feature importance is independent of the chosen threshold. The feature importance indicates which features the model contribute the most to the underlying probability estimates.

Threshold adjustments merely shift the decision boundary without altering the relative importance of the features. The top three features are in this particular model all lag features, specifically persistency and singular activation lag features. This suggests that the model relies heavily on historical activation patterns to make its predictions. Although one of the goals was to reduce the reliance on lag features, it is still important to catch on to activation trends. Other important features include day-ahead price, intraday wind forecasts, time-related features such as peak hour and working day indicators, and various interaction features. All these features likely contribute to capturing the complex dynamics influencing activation events.

8.2 The 2024-2025 Dataset

Some extra preprocessing is necessary when including data from 2024, as data often come in yearly batches. Separate CSV (Comma Separated Values) files for 2024 and 2025 are, therefore, merged into a single dataset before further handling. This dataset was used for most of the training and evaluation process, as it provides more data for the models to learn from, potentially leading to better generalization and performance. This comes at the cost of slightly inconsistent data. mFRR activation data, for instance, transitioned from hourly to 15-minute resolution in mid-2024 (**specify date perhaps**). This introduces inconsistency and noise into the dataset, which could affect model performance. However, the benefits of having a larger dataset likely outweigh the this drawback.

8.2.1 CatBoost Models

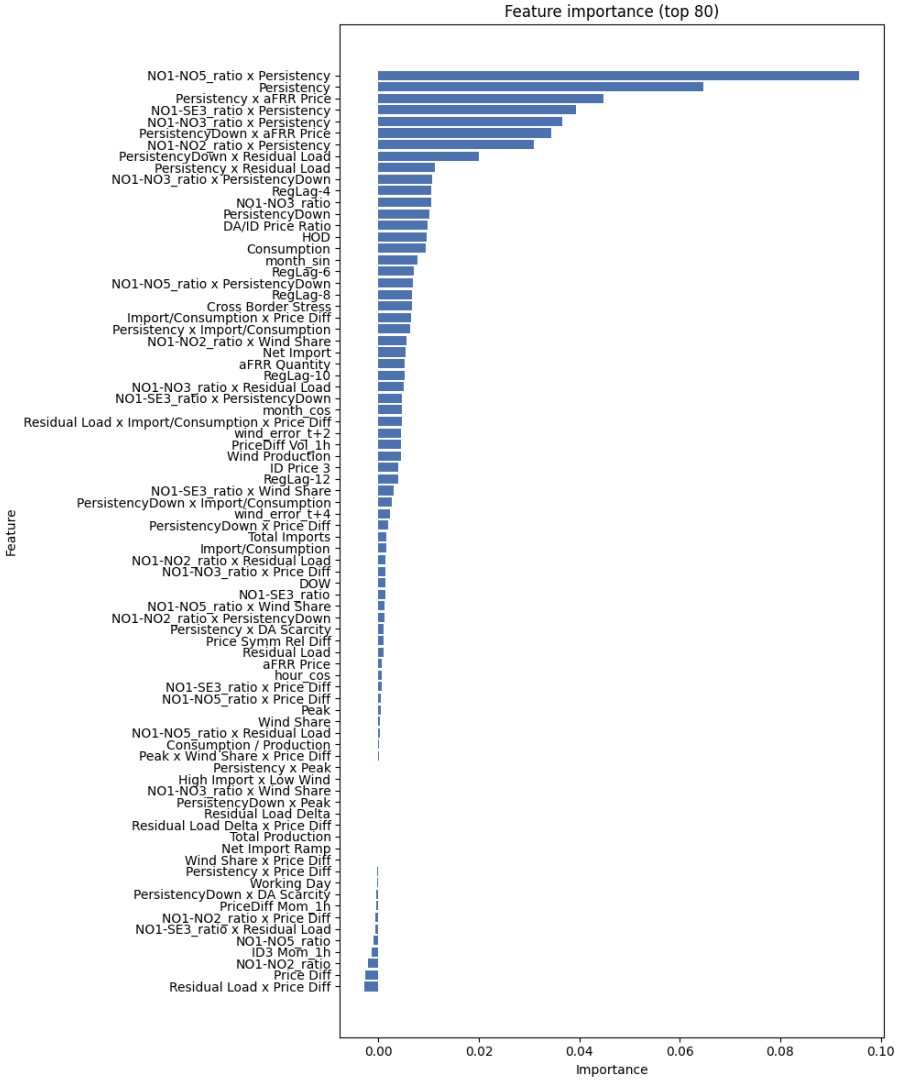


Figure 13: Feature importance for CatBoost model trained on 2024-2025 dataset after hyperparameter tuning.

I probably want a kind of sweep over performances on different hyperparamter combinations.

Metrics table

	val_f1_macro	val_accuracy	test_f1_macro	test_accuracy
0	0.633974	0.714051	0.59821	0.649165

Val classification report

	precision	recall	f1-score	support
down	0.73	0.71	0.72	4957
none	0.74	0.76	0.75	6274
up	0.44	0.42	0.43	918
accuracy			0.71	12149
macro avg	0.64	0.63	0.63	12149
weighted avg	0.71	0.71	0.71	12149

Test classification report

	precision	recall	f1-score	support
down	0.72	0.62	0.67	5112
none	0.63	0.73	0.68	5603
up	0.49	0.41	0.45	1436
accuracy			0.65	12151
macro avg	0.61	0.59	0.60	12151
weighted avg	0.65	0.65	0.65	12151

Figure 14: Performance metrics for CatBoost model on 2024-2025 dataset at different thresholds.

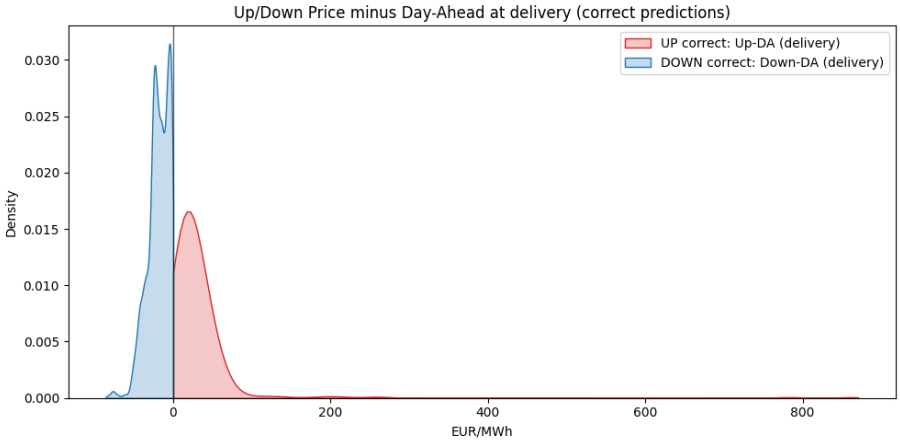


Figure 15: Up/down price minus day-ahead price distribution for CatBoost model trained on 2024-2025 dataset. quick_multiclass_cat_hpo

Figure 16 shows the performance metrics for a CatBoost model trained on data from March 4th, 2025. The metrics are evaluated at specific dataset subsets determined by how confident the model is in its predictions. For instance, at a confidence threshold of 0.6, only predictions where the model's predicted probability for the chosen class is at least 0.6 are considered. This approach allows for an analysis of how the model's performance varies with its confidence level. As the confidence threshold increases, accuracy steadily improves. This is expected, as higher confidence predictions should generally be more reliable. F1-macro score dips, however, at higher thresholds, especially beyond 0.7. The most likely reason for this is that the model rarely is very confident in predicting the less frequent classes (up- and down-activations). Most of these predictions are no-activation predictions, resulting in good accruacy (0.778 in this case). Assume that all of these predictions are no-activation predictions (**can probably check this quickly**). Then, 22.2% of the predictions are false negatives for the up- and down-activation classes, leading to low recall (0 in this case) and thus low F1-score for these classes. The overall F1-macro score, being the average of the F1-scores for all classes, consequently drops as well.

	threshold	coverage	acc	f1_macro
0	0.4	0.982168	0.683801	0.648802
1	0.5	0.869384	0.711508	0.655303
2	0.6	0.650533	0.742613	0.657654
3	0.7	0.331172	0.751748	0.613787
4	0.8	0.077351	0.778443	0.291807
5	0.9	0.000000	NaN	NaN

Figure 16: Performance metrics for CatBoost model on 2025 March 4th dataset at different confidence thresholds.

Bibliography

[1] X. Cai, N. Zhang, E. Du, Z. An, N. Wei and C. Kang, 'Low Inertia Power System Planning Considering Frequency Quality Under High Penetration of Renewable Energy', *IEEE Transactions on Power Systems*, vol. PP, pp. 1–12, Jan. 2023. DOI: 10.1109/TPWRS.2023.3302515

[2] ENTSO-e, *Noric Balancing Philosophy ENTSOE*, 2024. Accessed: 13th Nov. 2025.

[3] *Energy Transition Outlook Norway 2024*, https://www.norskindustri.no/siteassets/dokumenter/rapporter-og-brosjyrer/energy-transition-norway/energy-transition-norway-2024.pdf. Accessed: 3rd Dec. 2025.

[4] *Confirmation of mFRR EAM go live March 4th 2025*, https://www.statnett.no/en/for-stakeholders-in-the-power-industry/news-for-the-power-industry/confirmation-of-mfrr-eam-go-live-march-4th-2025/, Oct. 2025. Accessed: 22nd Nov. 2025.

[5] *Raske frekvensreserver - FFR*, https://www.statnett.no/for-aktorer-i-kraftbransjen/systemansvaret/kraftmarkedet/reservemarkeder/ffr/, Nov. 2025. Accessed: 13th Nov. 2025.

[6] Statnett, *Vilkår for mFRR aktiveringsmarked*, Jan. 2024.

[7] G. Klæboe, J. Braathen, A. L. Eriksrud and S.-E. Fleten, 'Day-ahead market bidding taking the balancing power market into account', *TOP*, vol. 30, no. 3, pp. 683–703, Oct. 2022, ISSN: 1134-5764, 1863-8279. DOI: 10.1007/s11750-022-00645-1 Accessed: 1st Dec. 2025.

[8] 'Balancing market outlook 2030',

[9] *Transition to 15-minute Market Time Unit (MTU)*, https://www.nordpoolgroup.com/en/trading/transition-to-15-minute-market-time-unit-mtu/. Accessed: 2nd Dec. 2025.

[10] V. V. Kallset and H. Farahmand, 'Improving Balancing Activation Through Continuous-Time Optimization and Increased Market Time-Resolution', in *2025 21st International Conference on the European Energy Market (EEM)*, May 2025, pp. 1–6. DOI: 10.1109/EEM64765.2025.11050190 Accessed: 2nd Dec. 2025.

[11] C. Singh, S. Sreekumar and T. Malakar, 'A novel dynamic imbalance volume forecasting model for balancing market optimization', *Electrical Engineering*, vol. 107, no. 12, pp. 15 375–15 392, Dec. 2025, ISSN: 1432-0487. DOI: 10.1007/s00202-025-03331-0 Accessed: 2nd Dec. 2025.

[12] K. Plakas, N. Andriopoulos, D. Papadaskalopoulos, A. Birbas, E. Housos and I. Moraitis, 'Prediction of Imbalance Prices Through Gradient Boosting Algorithms: An Application to the Greek Balancing Market', *IEEE Access*, vol. 13, pp. 103 968–103 981, 2025, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2025.3580274 Accessed: 3rd Dec. 2025.

[13] D. Azarang and C. Edling, 'Machine Learning-Based Prediction and Key Drivers of mFRR Activations',

[14] E. R. A. Overmaat, 'Balancing Contributions in the Nordic Electricity System',

[15] T. Svedlindh and K. Yngvesson, *Price Formation and Forecasting Models in the Electricity Market : An Analysis of the Intraday and mFRR Markets*. 2025. Accessed: 29th Nov. 2025.

[16] R. C. Porras, 'Short-Term Forecasting of mFRR Activation Direction and Imbalance Price using XGBoost',

[17] *Power Market Data*, https://www.nordpoolgroup.com/en/services/power-market-data-services/. Accessed: 22nd Nov. 2025.

[18] *Class Imbalance Problem - an overview — ScienceDirect Topics*, https://www.sciencedirect.com/topics/computer-science/class-imbalance-problem. Accessed: 26th Nov. 2025.

[19] *Static Content — Nordic Unavailability Collection System*, https://www.nucs.net/content/static_content/Static%20content/data%20repository/DataRepositoryGuide.html. Accessed: 22nd Nov. 2025.

[20] *Data & Standardisation*, https://www.entsoe.eu/data/. Accessed: 22nd Nov. 2025.

[21] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Springer Science & Business Media, Dec. 2012, ISBN: 978-1-4615-5689-3.

[22] J. Brownlee, *Basic Feature Engineering With Time Series Data in Python*, Dec. 2016. Accessed: 22nd Nov. 2025.

[23] H. Pelletier, *Cyclical Encoding: An Alternative to One-Hot Encoding for Time Series Features*, May 2024. Accessed: 22nd Nov. 2025.

[24] *Classification: Accuracy, recall, precision, and related metrics — Machine Learning*, https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall. Accessed: 22nd Nov. 2025.

[25] A. V. Dorogush, V. Ershov and A. Gulin, *CatBoost: Gradient boosting with categorical features support*, Oct. 2018. DOI: 10.48550/arXiv.1810.11363 arXiv: 1810.11363 [cs]. Accessed: 22nd Nov. 2025.