NTNU
Kunnskap for en bedre verden

DEPARTMENT OF ELECTRIC ENERGY

TET4510 - SPECIALIZATION PROJECT

Short-Term mFRR Activation Direction Forecasting at 15-Minute Resolution

Author:
Haakon Nygård Hellebust

Date

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

The increasing share of weather-dependent renewable energy sources has reduced the inertia of modern power systems and made short-term system imbalances more frequent and difficult to predict [1]. To maintain frequency quality, Nordic transmision system operators (TSO) rely on on various balancing resources, among which manual Frequency Restoration Reserves (mFRR) play a key role. The mFRR energy activation market (mFRR EAM) enables market participants to offer up- and down-regulation reserves that the TSO can activate when sustained imbalances occur [2].

Historically, mFRR activations were manual and scheduled on an hourly basis. On 4 March 2025, the Nordic region introduced a new automatic mFRR activation platform with a 15-minute resolution [3]. This reform aligns the Nordic system with European balancing-market standards and allows activation signals to more closely follow real-time system conditions. Automated activation of mFRR reserves enables faster response times, making it possible to adress quarter-hourly imbalances effectively [4].

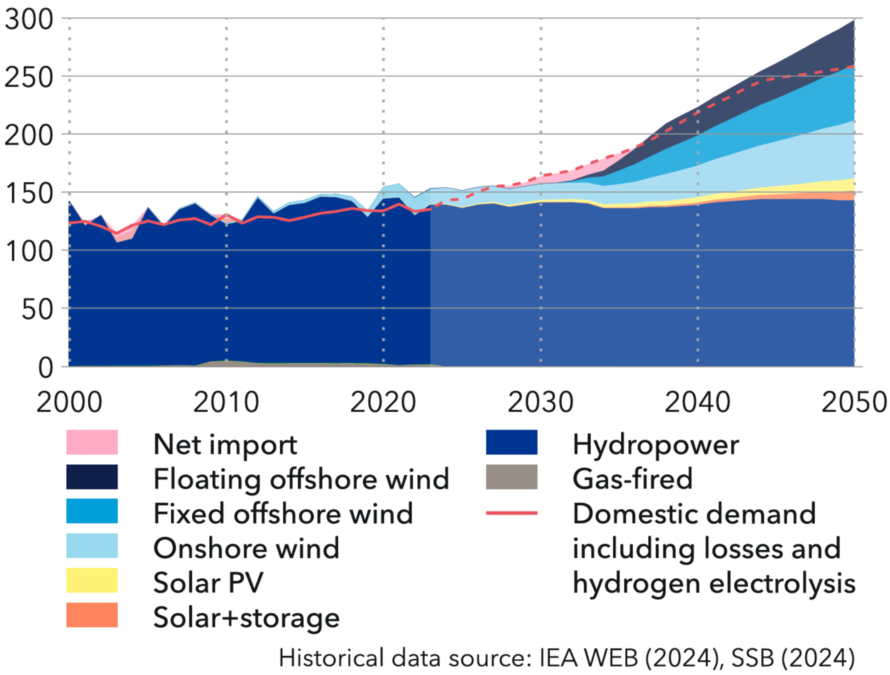


Figure 1: Norway electricity supply by power station and net imports with projections to 2050 [5]

Struggling with what to include in background as I feel like much of interesting background information kind of fits in motivation.

1.2 Motivation

mFRR activations have direct economic implications for market participants. Accurately anticipating whether up- or down-regulation is likely enables participants to choose between committing flexible resources to capacity markets or bidding directly into the activation market, where revenues depend on actual activations. Capacity markets provide stable income, whereas activation markets offer potentially higher but uncertain returns. Being able to accurately assess the likelihood of an activation therefore supports more informed and profitable bidding strategies.

For aggregators of distributed flexible resources, such as electric vehicles, heat pumps, and batteries, predicting mFRR activations is particularly valuable. When commiting flexibility to the mFRR capacity market, the aggregator earns a fixed payment for being available to provide reserves. However, this also means the resources are locked up and cannot be used for other purposes during the committed period. Participation in the activation market locks resources for shorter periods, allowing the aggregator to remain flexible. The guaranteed income from capacity payments must thus be weighed against the opportunity cost of missing out on potential revenues from the activation market. The correct choice depends on the likelihood of activations occurring, which is inherently uncertain.

A reliable activation prediction model can substantially reduce this uncertainty by providing an predictions of whether up- or down-regulation is likely in future intervals. With such information, an aggregator can manage its portfolio more strategically, for example, by reserving flexibility for the activation market only when the probability of activation is sufficiently high, or by safely committing to the capacity market when activations appear unlikely. Accurate predictions therefore enable more efficient utilization of distributed resources and support more economically rational market participation.

The recent shift to 15-minute resolution further increases the relevance of such models. While higher time resolutions offer greater responsiveness, it also creates a shorter decision window and increases sensitivity to rapidly changing system conditions. This motivates data-driven prediction methods as an approach to support market bidding strategies. An interesting application is to use the predictions to inform reinforcement learning-based bidding strategies for aggregators of distributed flexible resources. Reinforcement learning agents can learn to make optimal decisions, but require a reliable model of the environment to base their strategies on. A well-performing mFRR activation prediction model could serve as this model of the environment, enabling the development of near-optimal bidding strategies that adapt to predicted market conditions. **Is this too much reinforcement learning talk for an introduction?**

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? **Change this I suppose?**

1.4 Outline

This report is structured as follows: Chapter 2 provides an overview of the Nordic balancing markets and related work. Chapter 3 reviews existing literature on imbalance and activation forecasting and identifies the research gaps addressed in this study. Chapter 4 details the modelling methodology, including data preprocessing, feature engineering, and model evaluation and selection. Chapter 5 presents the results, while Chapter 6 discusses implications, limitations, and directions for future work.

2 Theory

2.1 Electricity Balancing Market

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.

2.1.1 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, I will make my own figure.

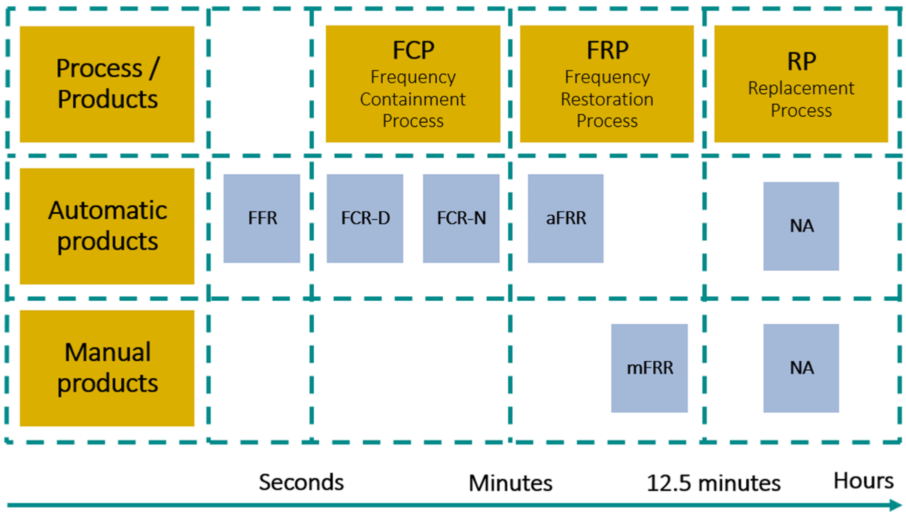


Figure 2: The Nordic balancing market hierarchy, illustrating the different reserve types and their activation times [2].

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 [6]. Figure 3 illustrates roughly the activation times and the interplay between the different reserve types in the Nordic balancing market.

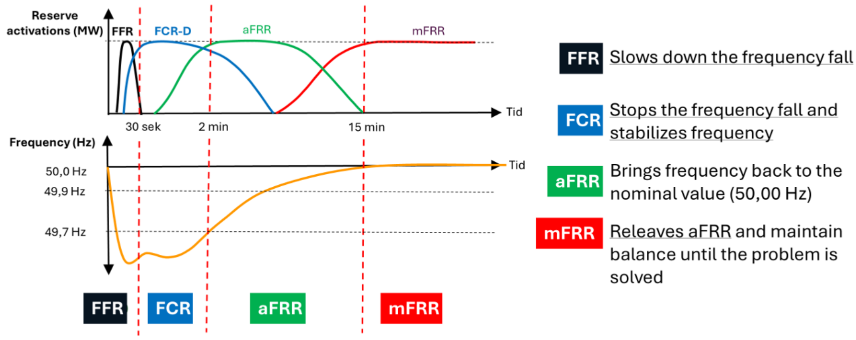


Figure 3: Illustration of the different reserve types and their activation times [2].

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.

2.1.2 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.

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.

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 [7].

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.

Imbalance Prices

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.

2.2 Machine Learning Theory

Machine learning (ML) is a subset of artificial intelligence (AI) that focuses on developing algorithms and statistical models that enable computers to perform specific tasks without explicit instructions. Instead, these algorithms learn from data, identifying patterns and making decisions based on the information provided. In this project, machine learning techniques are employed to predict the occurrence of mFRR activations in NO1 based on historical data and relevant system state indicators.

2.2.1 Supervised Learning for Time-Dependent Classification

Supervised learning is a machine learning paradigm where models are trained to map input data to specific outputs based on example input-output pairs. In this project, the input data consists of information about the current state relevant to the occurence mFRR activations. The output is a ternary label indicating whether an up-regulation, down-regulation, or no activation occurs. Thus, the output variable is categorical, making this a classification task, where the goal is to assign input data points to one of several predefined classes. However, the problem at hand is not a standard supervised learning task, as the data is ordered and time-dependent. This is often referred to as *event prediction*, where *events* are defined as nontrivial occurrences in specific locations and time [8]. In this project, location is fixed to the NO1 bidding zone, while time is discretized into quarter-hourly intervals.

The chronological structure has wideranging implications for model training and evaluation. Standard supervised learning techniques often assume that data points are independent and identically distributed (i.i.d.), which is not the case for time-dependent data. Temporal dependencies must be accounted for, as past events can influence future outcomes. This calls for specialized model and data engineering techniques that can effectively capture and leverage these relationships. Perhaps most importantly, the evaluation methodology must respect the temporal order of the data to avoid data leakage and ensure that the model's performance is assessed in a realistic manner. The model would, for instance, produce over-optimistic results if it has access to future data points when making predictions for a given time step.

2.2.2 Class Imbalance

Perhaps not introduce this yet? It is quite an important topic, though, as it introduces one of the biggest challenges. mFRR up activations are particularly rare events in the NO1 bidding zone compared to down activations and non-activations. This *class imbalance* presents a significant challenge for model training and evaluation. Most ML algorithms are designed to optimize overall accuracy, which can lead to models that are biased towards the majority class, since correctly predicting the majority class contributes more to overall performance. If the minority class is important, as is the case with mFRR up activations, special considerations must be taken to ensure that the model learns to effectively recognize and predict these rare events.

2.2.3 Tree-Based Models

Random Forest / Extra Trees

Gradient Boosting

3 Literature Review

This chapter reviews literature relevant to balancing market forcasting, first outlining the unique characteristics of balancing activation markets and the challenges they present. Then existing methodologies for handling these challenges are presented, before finally identifying specific research gaps that this study and future work can 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 in most balancing markets handled through 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 [9]. These market characteristics make it lucrative for participants to predict activation events accurately, as successful predictions can lead to significant financial gains.

In 2022, Klæboe et al. [9] 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 relevant and profitable. In Svenska Kraftnät's balancing market outlook 2030 [10] they present that the mFRR capacity demand has and will steadily increase. The report also suggests that since the automated mFRR EAM is only an intermediate step for connecting to MARI (Manually Activated Reserves Initiative), which is an upcoming European-wide mFRR market, further increases in mFRR demand are to be expected. This suggests that predicting mFRR activations will become increasingly important for market participants seeking to optimize their market strategies.

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 [11]. Under the previous hourly structure, activation signals were constrained to coarse discrete time blocks. Thus, short-lived imbalances or, for instance, rapid ramps in renewable generation could not be reflected optimally in activation decisions. moving to a 15-minute resolution reduces this discretization effect [12].

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 [12]. Their findings imply that imbalances are now corrected more accurately and efficiently on shorteer time scales, making activation patterns more sensitive to rapid system changes. Consequently, the dynamics of the mFRR energy activation market have become more granular and potentially more volatile, increasing the relevance, but also the difficulty, of short-term activation forecasting.

3.2 Activation Uncertainty in Balancing Market Forecasting

Most forecasting studies in the balancing-market literature focus on predicting continuous system variables such as imbalance volumes or imbalance prices. These quantities are natural targets for system operators, who must minimize balancing costs and anticipate system stress. However, imbalance volumes are inherently conditional on the discrete activation direction—upward, downward, or none—because volume magnitudes reflect both the sign and size of the underlying imbalance. When activation direction is not modelled explicitly, directional uncertainty becomes embedded in the volume forecast itself, contributing to noisy forecasts.

To address this challenge, the literature proposes various modelling approaches for representing activation uncertainty. These approaches differ in how uncertainty is represented. For clarity, these methods are grouped into six families: (i) scenario-based activation models, which simulate possible future imbalance trajectories; (ii) activation-ratio or expected-activation models, which derive expected activation ratios from historical data; (iii) activation-probability and chance-constraint models, which enforce reliability requirements based on probabilistic imbalance representations; (iv) activation-range or interval-uncertainty models, which define bounded sets of feasible activation magnitudes; (v) regressor-based activation models, including machine-learning methods that explicitly predict activation direction; and (vi) Markov activation models, which represent activation direction as a stochastic process represented by transition probabilities.

The following sections review these modelling families, beginning with imbalance-volume forecasting studies before assessing how each uncertainty-modelling approach handles, or fails to handle, the discrete up/down/none activation decision relevant for mFRR energy markets.

3.2.1 Direct Imbalance Volume Forecasting

At the system-operator end of the spectrum, a substantial body of literature focuses on point forecasting of continuous imbalance volumes, with the primary purpose of improving TSO operational decisions. Singh et al. [13] exemplify this class of work through a regression-based model for short-term imbalance forecasting in Belgium. They argue that increasing renewable variability, combined with the 15-minute activation window, necessitates accurate short-horizon forecasts to allow TSOs to anticipate system deviations more effectively. Their best-performing model reduces balancing costs by 44.51% relative to TSO benchmarks, driven by reductions in energy-not-supplied, excess energy, and correction costs.

Related work in the Nordic region highlights similar challenges. Edling and Azarang (2025) forecast short-term mFRR activation volumes across the four Swedish bidding zones using LSTM models [14]. Their results reveal strong geographical heterogeneity in predictability: SE2 exhibits comparatively high accuracy, while SE3 and SE4 show limited predictability due to the prevalence of zero-activation intervals. This distinction highlights an important structural feature of the Nordic system: regions with frequent imbalances do not necessarily experience frequent activations. Because Sweden's flexible hydropower capacity is concentrated in SE1 and SE2, the TSO often activates reserves there even when imbalances originate in other zones, provided network constraints permit it. Earlier results by Overmaat [15] confirm that SE1 and SE2 historically provide the majority of balancing energy on short and medium time scales.

While Edling and Azarang approach the problem from a TSO perspective, Backe et al. in the KoBas project [16] examine imbalance-volume forecasting from a market-participant-oriented standpoint. Additionally, they develop a LSTM model, including several Nordic bidding zones. Their analysis yields three relevant insights. First, balancing volumes are relatively auto-correlated: past imbalances contain predictive information about short-term imbalances. Second, they note that forecast accuracy could likely be improved by incorporating weather-related variables. Third, they stress that zero-regulation dominates the dataset, meaning the model must infer relatively infrequent activation events from a mostly inactive baseline. This is an inherent limitation when direction is not modelled explicitly.

Plakas et al. [17] focus on market-participant perspectives as well, proposing 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.

In a similar vein, Bankefors (2024) applies linear machine-learning-based time series models (ARIMAX, SARIMAX) to predict signed mFRR activation volumes [18]. Bankefors concludes that while imbalance and activation volume forecasting is challenging, the models showed promise in implicitly predicting activation direction. He suggests that future work could explore classification-based approaches to directly predict activation direction rather than inferring it from volume forecasts.

Together, these studies highlight a structural limitation of direct imbalance volume forecasting: the sign and magnitude of imbalances volumes depend on activation direction, meaning that models that do not explicitly model activation direction uncertainty must implicitly learn it from the noisy data. This can lead to degraded forecast quality, especially around directional switches. While Plakas et al. [17] couple imbalance volume and price forecasting in a two-stage framework, this approach could potentially be extended further by first predicting activation direction, then imbalance volumes, and finally imbalance prices. The subsequent sections discuss how existing literature has addressed the challenge of modeling activation direction uncertainty.

3.2.2 Scenario-Based Activation Models

Scenarios are often used to handle uncertainty in optimization problems. Possible future outucomes are represented as discrete scenarios, each with an associated probability. In the context of balancing markets, scenarios represent possible trajectories of net system imbalance, which implicitly determine required activation volumes. This approach is for example used in reserve dimensioning [19] and stochastic scheduling or bidding frameworks [20].

In [21], Håberg and Doorman model activation uncertainty using three discrete scenarios: high, median, and low, as illustrated in figure 4. The scenarios are represented as forecasted continuous imbalance volumes over a 40-minute horizon. The imbalance forecast scenarios were generated from probability distributions based on historical imbalance data.

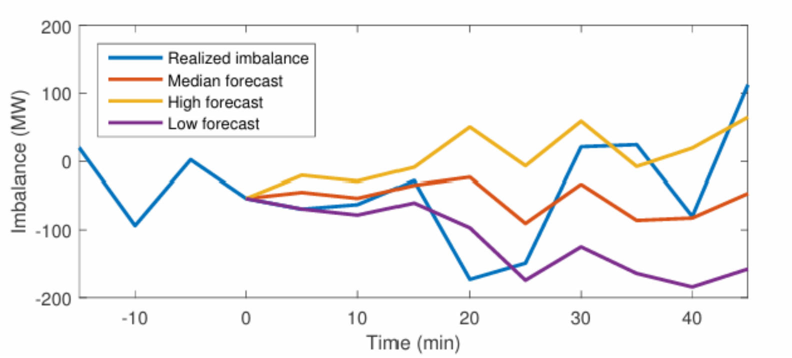


Figure 4: Imbalance forecast scenarios [21].

A key limitation of this modelling family is that activation direction is not modelled explicitly, but arises solely from the sign of the scenario imbalance volumes. Consequently, any uncertainty in direction is entirely dependent on the quality of the scenario-generation process. Thus, they rely on high quality imbalance forecasting models, which is, as outlined in this literature review, a challenging task in itself. All in all, scenario-based approaches seem to be useful only when highly accurate imbalance forecasts are available. Therefore, they do not directly address the challenge of modelling activation direction uncertainty.

Hmm, actually I do not really feel like this section fits here in the structure. I may have misunderstood what you mean by scenario-based approach, but it seems it is not really a modeling of activation uncertainty, activation direction is simply implied by the sign of imbalance scenarios which arise from imbalance forecasting. This example study probably is not relevant as it applies imbalance forecasts, instead of helping create them.

3.2.3 Activation Ratio or Expected Activation

Some studies model activation uncertainty using constant activation ratios or expected activations. A common approach is to estimate the probability of activation in each direction (upward, downward, none) based on historical activation frequencies. Irrmann (2023) applied this method analyze and model the Nordic balancing markets [22]. In this study, *regulation states* (up, down, none) are sampled based on historical frequencies, before activation volumes are drawn from a modelled distribution conditonal on the sampled state. This approach decouples the discrete activation decision from the continuous volume forecasting, allowing for more targeted modelling of each component. However, the activation probabilities are statically estimated from historical frequencies. Although such estimations may be adequate over longer time horizons, they are likely to perform poorly in the short term, as activation patterns are highly non-stationary and dependent on the current system state, requiring more adaptive methods. Irrmann somewhat addressed this by estimating separate probabilities for each month of the year, but this coarse temporal segmentation is unlikely to capture the full dynamics of activation behaviour.

3.2.4 Activation Probability and Chance Constraints

Chance constraints are a mathematical optimization technique used to handle uncertainty by ensuring that certain constraints are satisfied with a specified probability. Papavasiliou et al. (2022) [23] apply chance-constrained optimization to the problem of reserve dimensioning in a multi-area power system. Here, uncertainty described by scenarios are revealed in the form of continuous imbalances. The chance constraints impose reliability limits for up and downward reserves, ensuring that the procured reserves can cover imbalances with a certain probability. This is an application of chance constraints to balancing markets, but it is geared towards TSO reserve dimensioning rather than participant-side activation forecasting. Additionally, activation direction is, similar to Håberg and Doorman's scenario-based approach [21], only modelled implicitly through the sign of the continuous imbalance scenarios.

Browell (2018) [24] develops risk constrained trading strategies for stochastic generators in the UK balancing market. In this study, *system length*, i.e. the net imbalance direction, is modelled probabilistically by a logistic regression model. Then, chance, or risk, constraints are imposed to ensure that trading strategies meet certain performance criteria with high probability. This study is tailored particularly to stochastic generators, whose production uncertainty directly influences their balancing market participation. Thus, the method and results will not generalize perfectly to other applications, but this paper marks an early and important attempt to explicitly model activation/imbalance direction uncertainty.

3.2.5 Activation Uncertainty Ranges

Pavíc et al. (2023) argue that deterministic reserve activation models inccurately represent the activation uncertainty. Thus, they present a stochastic model, but more interestingly, they also propose a robust electric vehicle aggregator scheduling model using uncertain bounded activation ranges [25]. They use *reserve activation* (RA) as input for activation uncertainty, which is defined as the ratio of activated reserve energy to the accepted reserve capacity. Their analysis is limited to 30-minute FCR and aFRR reserve data for 2018. Activation data is gathered and probability distributions are constructed as figure 5, representing the likelihood of different activation ratios. Relevant statistics, like mean, max and quantiles, are used as inputs for their models.

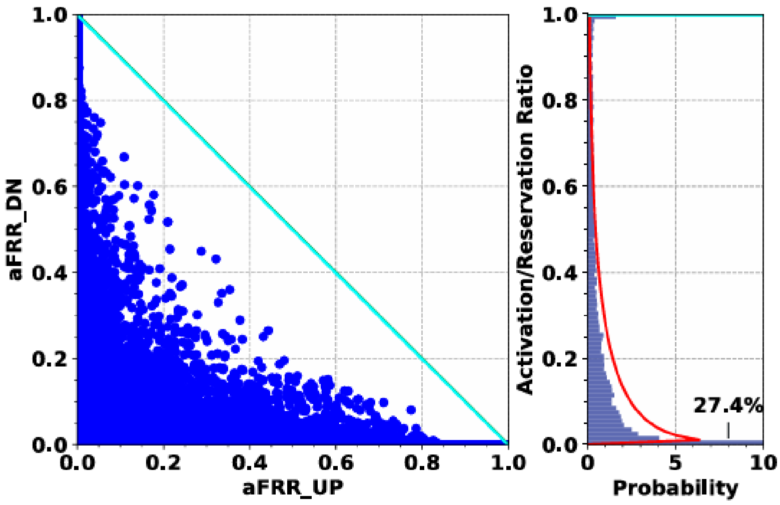


Figure 5: Activation ratio uncertainty ranges for aFRR up [25].

Using activation ranges to represent uncertainty is an interesting approach, as it directly models the fraction of accepted reserves that are likely to be activated. This is very useful information for flexible demand-side aggregators, who must decide how much capacity to offer based on expected activations. The ranges imply worst-case, best-case, and expected activation scenarios, which can be used to inform robust models attempting to remain feasible under uncertainty. Such uncertainty sets may be more appropriate than probabilistic and deterministic frameworks for flexible demand-side market participants, who must ensure feasibility at all costs. Pavic et al. operate in the context of FCR and aFRR reserves, where there is always an activated imbalance in one direction [25]. mFRR balancing, on the other hand, often requires no activation at all. This is not a problem per se, but it would skew the activation ratio distributions significantly, as a large probability mass would be located at zero activation.

3.2.6 Markov Activation Models

Klæboe et al. benchmarked time series based forecasting models for electricity balancing markets in 2015 [26]. In this study, they separate relevant work into two families: models explicitly modeling balancing state and those modeling it implicitly. Various implicit and explicit activation direction models have been discussed extensively in this literature review, but Klæboe et al. highlight Markov models as a particularly interesting approach for explicit balance state modeling. The study refers to work by Olsson and Söder, who used a non-time-homogeneous Markov model, with varying transition probabilities depending on balancing state durations. Balancing states of durations 0-5 hours were modeled with separate transition matrices, while longer durations used a common matrix [27].

Another approach is to construct different transition matrices for different hours in the day. This allows the model to capture patterns such as higher transition probabilities during the day compared to night. Klæboe et al. also tested a model based on the work of Croston [28], only distinguishing between activation and no activation, thereby discounting direction. This model distinguishes between up- and down regulations in a sepearate price- or volume process.

One-hour ahead predictions were shown to be quite accurate, predicting correctly 63% amd 73% of the time for duration dependent and hour specific models, respectively. However, the models struggled with longer horizons, with accuracy dropping to around 30% at day-ahead. The Croston-based model, when benchmarked on regulation vs no-regulation, achieved around 59% accuracy at one-hour ahead, notably lower than the other models, but outperformed them at day-ahead prediction. Another finding was the struggle to predict direct transitions between upward and downward activations, as these events have low transition rates. The duration dependent Markov model is an interesting approach, and one that is used as inspiration for the model development in this study. It captures persistence in activation patterns, which are known to be important [29].

3.2.7 Activation Direction Classification Models

Whereas imbalance forecasting estimates continuous system imbalance magnitudes, activation-direction forecasting seeks to predict the discrete TSO decision to activate upward, downward, or no mFRR energy. Activation direction is directly relevant for market participants because bids must be placed in the correct direction to be eligible for activation. With the Nordic system's transition to 15-minute activation intervals, short-term direction forecasting has become more important. However, the academic literature that treats direction as a primary target remains sparse.

Svedlindh and Yngveson [30] examine 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 promising results. They find that mFRR capacity market prices and procured volumes are informative predictors of activation direction.

Porras (2025) [31] applies an XGBoost two-stage model to sequentially forecast activation direction and imbalance prices at hourly resolution in SE2. The study demonstrates the potential of tree-based methods, but it also exposes two practical limitations for participant-oriented forecasting: (i) the model operates at hourly resolution, and it remains unclear how well the approach would perform at the new 15-minute resolution; and (ii) its most important predictor is "balance-direction at $t - 0$" that appears to be unavailable to market participants at the time of bidding. This represents a form of feature leakage (use of variables that would not be accessible in real decision-making) and likely inflates performance. **This is correct, right?**

In summary, regressor-based approaches show promise for explicit direction forecasting, but current studies leave questions yet to be answered: Can direction be predicted reliably at 15-minute resolution with only available information? And can sparse up/down events be predicted with useful precision? The present study addresses these questions by evaluating multiple classifier families under strict participant-feasible information constraints and at the updated 15-minute resolution.

3.3 Literature Synthesis

Table 1 summarizes the studies reviewed in this literature review that are most relevant to balancing-market forecasting and uncertainty modelling. Literature is organized by authors, target variable, temporal resolution, and uncertainty-modelling approach.

Table 1: Overview of key forecasting and uncertainty-modelling studies

Study	Target Variable	Resolution	Uncertainty Model
Singh et al. (2025)	Imbalance volume	15-min	Point forecast (regression)
Edling & Azarang (2025)	mFRR activation volume	1-hour	ML point forecast (LSTM)
Backe et al. (KoBas) (2023)	Imbalance volume	1-hour	Probabilistic LSTM; implicit direction
Plakas et al. (2025)	Imbalance volume and price	1-hour	Probabilistic (quantile regression)
Bankefors (2024)	Signed activation volume	1-hour	Linear ML time-series; implicit direction
Irrmann (2023)	Direction and volumes	1-hour	Expected activation ratios; sampled regulation states
Papavasiliou et al. (2022)	Reserve sufficiency	Scenario-based	Chance constraints on imbalance scenarios
Browell (2018)	System length (direction)	1-hour	Probabilistic logistic model with risk constraints
Pavić et al. (2023)	Reserve activation (RA ratio)	30-min (FCR/aFRR)	Activation ranges / bounded uncertainty sets
Klæboe et al. (2015)	Balancing state (up-/down/none)	1-hour	Markov transition probabilities (duration/hour-specific)
Svedlindh & Yngveson (2025)	Activation direction	1-hour	Classification (logistic, ANN)
Porras (2025)	Activation direction and price	1-hour	Classification (XGBoost); sequential predictive model

3.4 Research Gap

Despite notable progress and coverage of balancing-market forecasting, a couple of gaps emerge from the literature. The simplest observation is that most studies were conducted before the Nordic system's transition to 15-minute mFRR MTU in March 2025. As discussed, this change alters the market dynamics and activation patterns. While hourly forecasts remain relevant, it remains unclear how well existing models perform at the higher resolution.

The literature focuses predominantly on continuous imbalance volumes or prices as target variables. These quantities are of great importance, but they are conditonal on activation direction, which is not modelled explicitly in most studies. Results are therefore affected by directional uncertainty, making predictions noisier. This study argues that explicit direction modelling is useful for market participants by itself, but also as a stepping stone towards improved volume and price forecasting.

A further consideration is that not all studies robustly address the data availability constraints faced by market participants. Some studies incorporate TSO-only data, while others do not explicitly evaluate whether their chosen features would be accessible to market participants at decision time. Porras' study [31], for instance, predicts activation direction directly, but relies on not-yet-available features, likely inflating performance. This study emphasizes strict adherence to participant-feasible information sets to ensure practical relevance.

This study aims to fill these gaps by systematically evaluating machine learning-based activation-direction forecasting in the NO1 bidding zone at 15-minute resolution, using only features available to market participants.

4 Methodology

4.1 Overview of Methodological Approach

The methodological approach adopted in this project is visualized in Figure 6. The first step, data collection, involves gathering relevant datasets from various sources, including Nord Pool, NUCS, and ENTSO-E. The collected data is then preprocessed and cleaned in the second step. These first steps are uniquely colored red in the figure to indicate that they are mostly one-time efforts required to set up the dataset for further.

The third step, feature engineering, involves creating and selecting relevant (**either introduce features here or make sure it has been introduced**) features, i.e. attributes in the dataset that help the model learn patterns related to mFRR activations. The feature-engineered dataset is subsequently sent to the model training algorithm. Here, machine learning techniques are applied to train classification models using the prepared dataset. Then, the trained models are evaluated with the relevant metrics to assess their predictive performance. These three steps are colored purple, indicating that they are iterative processes that make up the bulk of the work invested in this project.

Evaluation results are interpreted in the last step indicated by the blue box. This step is unique as it only involves deriving insights and conclusions from the previous steps. Methodological feedback loops flow from the interpretation step back to the feature engineering, model training, and evaluation steps. These feedback loops represent the iterative nature of the methodology, where insights gained from interpretation inform further refinements and improvements in the earlier stages of the process.

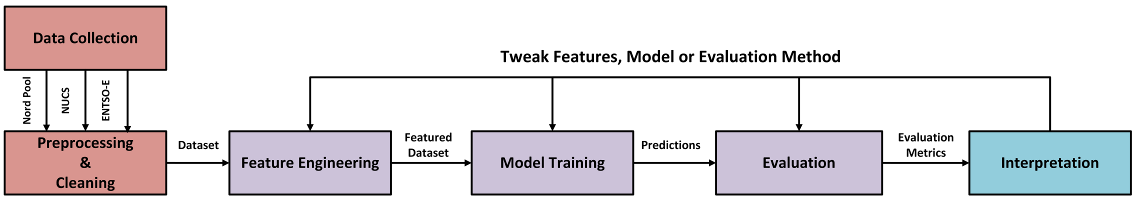


Figure 6: Overview of the methodological approach used in this project.

4.2 Data Sources

This section introduces the data sources used in this project and describes how the data was gathered and preprocessed for further use in model training and evaluation. The gathered data spans a period from January 1, 2024, to December 4, 2025, providing a comprehensive view of NO1 mFRR activation patterns over nearly two years.

4.2.1 Nord Pool

Nord Pool provides market and system data in the Nordic region. Data was downloaded manually through the Nord Pool data portal in yearly chunks [32]. A 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 obtained. This is not an issue for this project and research as real-time data is not required for model training and evaluation. 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. **This is not really relevant right here, but it is an important point to make somewhere.**

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.

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.

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.

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.

4.2.2 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) [33]. This allows for automated data retrieval, which is especially useful for real-time applications. This project does not have access to comprehensive real-time data, and the NUCS API was thus only used to gather historical data for the training and evaluation of the models. An algorithm was developed, however, to automatically retrieve up-to-date data from the NUCS API for potential future real-time applications.

aFRR data. aFRR data is not available through Nord Pool, but through the NUCS API, historical aFRR procurement prices and volumes for the NO1 bidding zone are accessible. The data is available at an hourly resolution, with separate values for up- and down-regulation. This data provides insights into the amount of balancing that is expected to be needed in the system.

4.2.3 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 [34].

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.

4.3 Data Preprocessing and Analysis

4.3.1 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.

4.3.2 Resampling, Imputation, and Merging

Nordic power market data is transitioning from hourly to 15-minute resolution. However, many datasets are still only available at an hourly resolution, and some datasets have mixed resolutions over the past years. Day ahead price data, for instance, transitioned to 15-minute market time units (MTU) on September 30th 2025 [35]. Consequently, data before this date is at an hourly resolution, while data after this date is at a 15-minute resolution. mFRR activation data transitioned to 15-minute MTU on March 4th 2025 [3], but Nord Pool has updated their historical data to be at a 15-minute resolution for the entire dataset period.

To ensure consistency across all datasets, all data with hourly resolution data subsets are resampled to a 15-minute resolution. This is done using the built-in Pandas `resample()` function with forward filling. Forward filling entails propagating the last valid observation forward to fill gaps. For instance, if the day-ahead price between 10:00 and 11:00 is 50 EUR/MWh, then after resampling, the price for 10:00, 10:15, 10:30, and 10:45 will all be set to 50 EUR/MWh, as illustrated in Figure 7. Thus, the data still remains constant within the hour, but is now available at the desired 15-minute resolution.

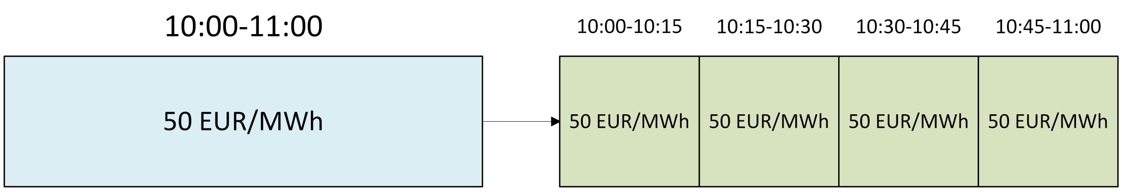


Figure 7: Visualization of resampling from 1-hour to 15-minute resolution using forward filling.

The handling, or *imputation*, of missing values, is an important step in data preprocessing. In time-dependent data, simply removing the rows with missing values is problematic, as it would break the time continuity. Most of the datasets used in this project do not have significant issues with missing values, but for the few that do, interpolation or forward/backward filling techniques are used to estimate the missing values based on surrounding data points. Interpolation is often preferred, as it can provide smoother estimates, as it considers both previous and subsequent data points. Backward filling is used when missing values are at the beginning of the dataset, as there are no previous data points to reference. Otherwise, forward filling is used as the default method, as it maintains the most recent known value, thus preventing time leakage from future data points.

After resampling and handling missing values, the various datasets are merged into a single dataset.

A successful merge requires that all datasets share a common time index format and resolution. Datasets from different sources often differ in time zone and how time stamps are encoded. Therefore, all time stamps are converted to a common time zone (CET/CEST) and format (Pandas `to_datetime()` function) before merging. The final merged dataset contains all data aligned at a 15-minute resolution, ready for feature engineering and model training.

4.3.3 Exploratory Data Analysis

Target Class Imbalance. Figure 8 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 [36].

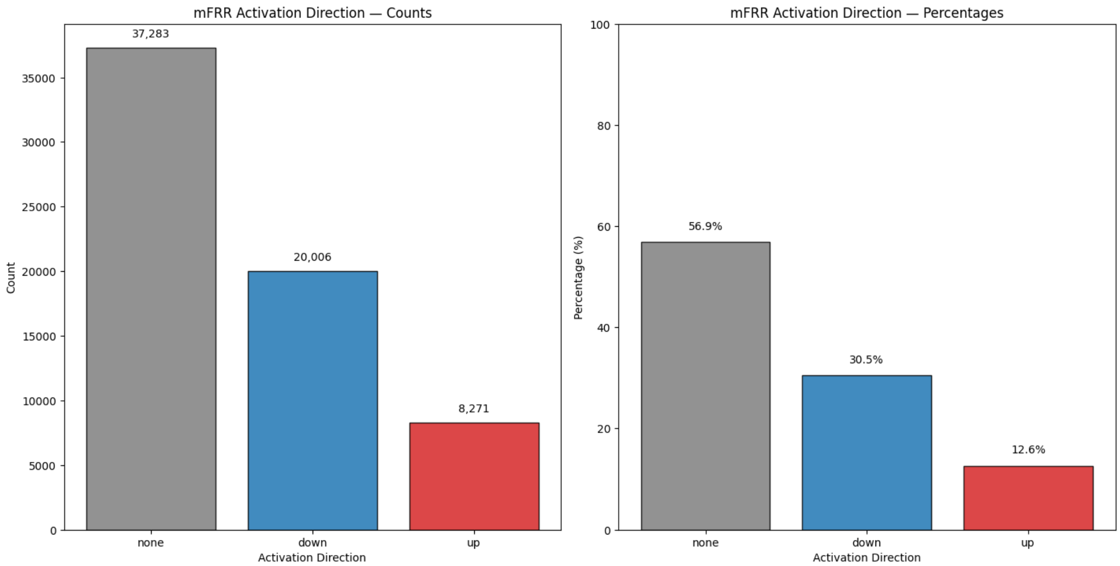


Figure 8: mFRR activation distribution.

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.

Temporal Activation Patterns

Feature Distributions. Figure 9 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.**

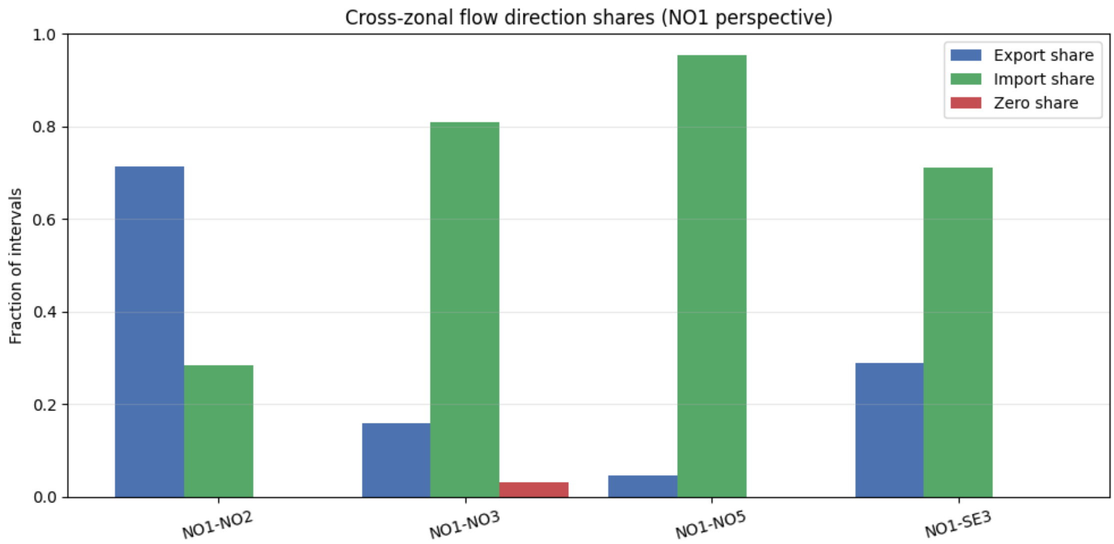
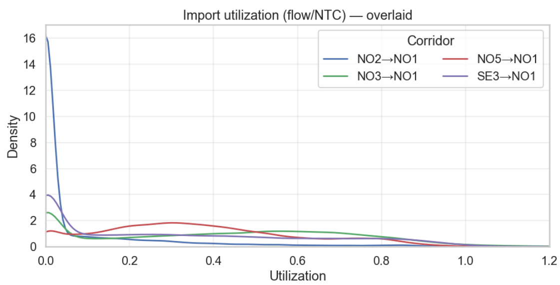
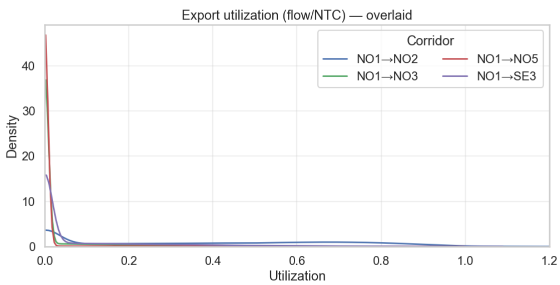


Figure 9: Cross-zonal flow distributions for the NO1 bidding zone.

Figure 10a and 10b 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 10: Cross-zonal flow utilizations for the NO1 bidding zone calculated as the ratio between actual flow and NTC capacity.

Production May or may not include production data EDA here.

Table 2: 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

aFRR. Figure 11 and 12 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 pre-July 2024 data, as this period contains many missing values. The problem seeems to be intermittent missing values rather than large gaps of missing data, indicated by the time series plot. Interpolation is thus a suitable imputation method, as it can estimate the missing values based on surrounding data points. This may miss out on some extreme price spikes, but is still expected to provide a reasonable estimate. The data appears complete after this date. **Short discussion on price levels and patterns here would be good. ALSO be specific on up/down.**

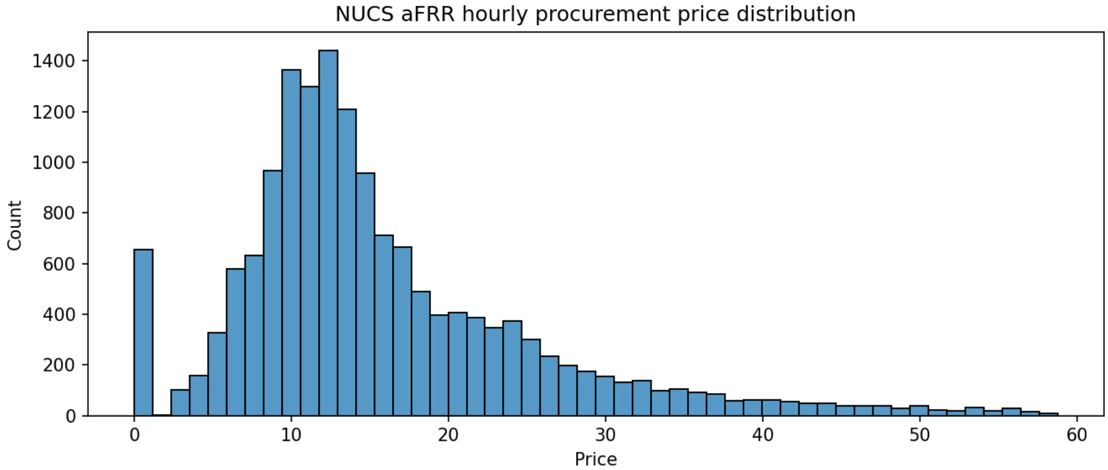


Figure 11: A histogram of hourly aFRR procurement prices for the NO1 bidding zone from NUCS data.

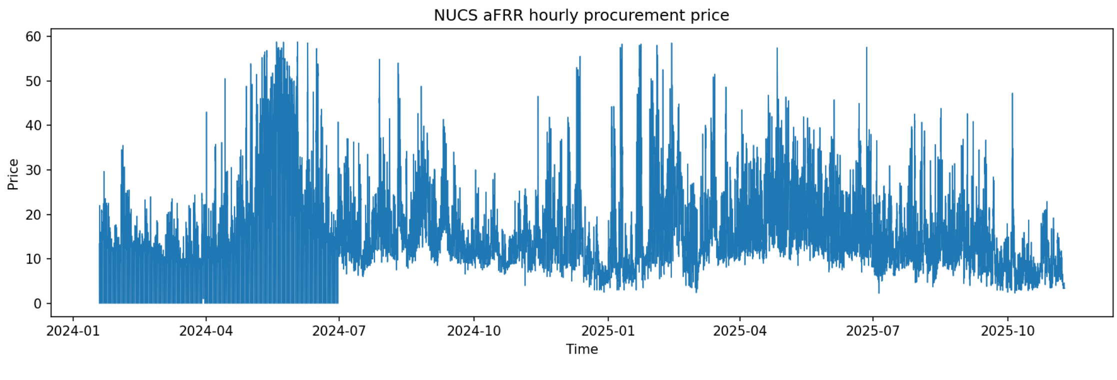


Figure 12: Hourly aFRR procurement prices for the NO1 bidding zone from NUCS data.

Correlation Analysis

Information Availability

4.4 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* [37]. 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 **??** to ensure that only real-time available features are used.

4.4.1 Time Restrictions and Data Availability

Short-term activation prediction is constrained by the limited availability of recent system data at prediction time. mFRR EAM bidding for a specific MTU closes 45 minutes prior, so at time t we can only act on predictions for the interval $t+4$ and later. This restriction causes a great amount of uncertainty. Even the best approximation of the current system state is several 15-minute intervals old at the target delievery time.

Many data sources have reporting delays, meaning that the most recent data points are not yet published at prediction time. For example, mFRR EAM activation data is available with a delay of approximately one hour. This means that at time t, the most recent activation data available is from time $t - 4$. However, since mFRR EAM activations are known to be highly autocorrelated ([16]**source or maybe not here?**), so even delayed activation data can provide valuable information about current trends.

Time delay figure?

Additional challenges arise from data sources published at lower resolutions resolutions. For example, aFRR activation data is released hourly, causing new information to arrive in batches rather than continuously. Depending on the prediction time within the hour, some recent lagged values may not yet exist, leading to inconsistencies in which features are available in real time.

Similar delays affect other system indicators such as consumption, production, and wind forecasts, many of which are published with reporting lags. As a consequence, the prediction task must rely primarily on features that are available at or before time t, including forecasts and lagged values that are consistently observable. These constraints shape both the feature engineering process and the achievable model performance, as the model cannot learn from information that would not be accessible to market participants in practice.

4.4.2 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 [38]. By including lag features, the model can leverage historical information to make more informed predictions about future mFRR activations. This becomes increasingly important when real-time data is limited, as lag features can provide context about recent system states and patterns. d **Lag feature figure?**

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 $t-4$ and earlier for predicting activations at time $t+4$. As a result, lag features for upregulating and downregulating activations are created for time steps $t-4$, $t-5$, $t-6$, ..., $t-9$. These features are useful by themselves, but they also serve as a basis for creating other features that capture activation trends more effectively.

Persistence (activation streak length) Lag features indicate what happened in the most recent intervals, but they do not express whether the system has been in a sustained activation phase. For the model to understand such trends, it would need to look at several lagged activation features simultaneously and infer whether there has been a streak of up- or down-activations. To capture this behaviour succinctly, a set of persistence features is included. These measure how long the latest sequence of up-, down-, or no-activations has lasted, based only on the activation data that are available at bid time. The idea is straightforward: if down-activations occurred in several consecutive intervals leading up to and including $t-4$, the down-persistence value reflects the length of that streak, as visualized in figure 13. The same applies for up-activations and no-activations. These persistence features aim to capture the tendency for mFRR activations to occur in clusters, as periods of system stress often lead to repeated activations. By condensing this pattern into a single value for each direction, the model is given a clearer representation of ongoing activation dynamics than lagged indicators alone can provide.

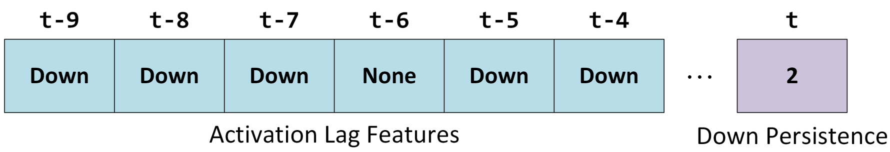


Figure 13: Illustration of down persistence feature calculation based on lagged activation features. Here, the down persistence at time t is 2, as there have been down-activations in the two most recent intervals ($t-4$ and $t-5$), before an interval with no activation at $t-6$.

Persistence feature capture consecutive activation trends effectively, but they do have limitations. Singular lagged activation featurs are still useful in the case of intermittent activations that do not form long streaks. Additionally, persistence features do not convey the magnitude of recent activations, only their occurrence. How all mFRR activation-related features are used in conjunction will be subject to experimentation and analysis during model development.

4.4.3 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 high-capacity 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 flow on interconnection i at time t, taken as positive when power flows into NO1 and negative when it flows out. Let $ATC_i(t)$ be the available transfer capacity for that interconnection at time t. The capacity-normalized flow is then

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

Values of $F^i_{\text{ratio}}(t)$ close to 1 mean that the inflow is close to the capacity, values close to -1 mean that the outflow is close to the capacity, and values near 0 mean that the net flow is small compared to the available capacity.

4.4.4 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 [39]. Hourly features are, for instance, encoded as:

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

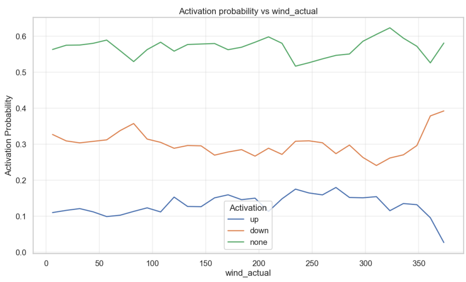
Monthly features can be encoded similarly, using 12 as the divisor instead of 24.

4.4.5 Price features

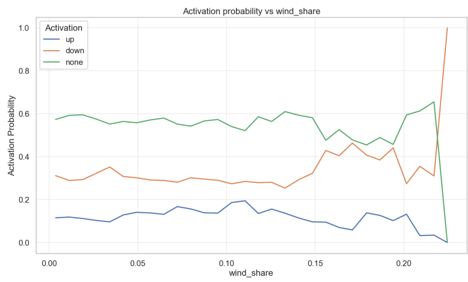
Price features capture information about the various electricity market prices. The mFRR activaiton market is closely linked to other electricity makets, such as the day-ahead market, the intraday market, and the aFRR market. Most prices may not have direct impacts on activations, but by crafting features 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.

4.4.6 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 14a and 14b 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 14: Production data distributions.

4.4.7 Load features

Load features capture information about electricity consumption patterns. Absolute consumption magnitude for NO1 is included as a feature, but there are many ways to encode consumption in normalized or relative terms. For instance, consumption can be expressed as a ratio to forecasted consumption, to capture forecast errors. Consumption can also be expressed as a ratio to relevant historical consumption values. For the final feature set, a ratio between current consumption and the average consumption at the same hour throughout the dataset is used as a feature to capture deviations from typical patterns.

4.4.8 Interaction features

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

4.4.9 Feature Correlation

I put this here only to highlight that it might be appropriate to discuss the possibility that many of the included features correlate with eachother: eg. changes in cross-zonal flows may correlate with wind forecast errors etc., and of course the many activation features.

Balancing vs reserve(activaiton vs capacity)

Activation volumes mostly zero, so predicting volumes maybe not good to predict directly. Two stage.

Demand side flexibility, dont care about price, marginal costs zero

First step towards regression.

4.5 Evaluation Framework

Classification problems are often evaluated using accuracy, precision, recall, and F1-score. These metrics are defined as follows [40]:

- **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.

Multiclass evaluation. When the classification is no longer binary, but multiclass (three or more classes) which is the focus of this project, evaluation metrics must be adapted. Common approaches include macro-averaging, micro-averaging, and weighted averaging. The weighted-averaged F1-score is calculated by taking the mean of all per-class F1-scores, weighted by their *support*, i.e., the number of true instances for each class. This approach is appropriate if the class distribution is imbalanced, and it is desired to give more importance to the performance on the more frequent classes. This is not the case in this project, as minority classes are of great importance. Therefore, the macro-averaged F1-score is used, which treats all classes equally regardless of their frequency. Macro-averaged F1-scores are calculated by computing the F1-score for each class independently and then taking the average of these scores. Micro-averaging, which aggregates the contributions of all classes to compute the average metric, is less commonly used for imbalanced datasets. Micro-averaging is analogous to accuracy in binary classification, which is not suitable for this problem due to class imbalance [41].

Confusion matrix. The confusion matrix is a cross tabulation of predicted versus actual class labels. It provides a breakdown of correct and incorrect predictions for each class, aligning correct predictions along the diagonal. What stands out in a confusion matrix is not just the overall accuracy, but also the specific types of errors the model makes. For instance, in a three-class classification problem, the confusion matrix can reveal if the model tends to confuse certain classes more than others. An example confusion matrix for a three-class classification problem is shown in figure 15. In this example, the model performs well on the 'down' class and 'none' class, capturing most of the true instances while making few false predictions. The 'up' class has four correct predictions but also three misclassifications as 'none' and one as 'down', indicating that the model struggles to differentiate the 'up' class from the others.

		Predicted Classification		
	Classes	**Down**	**None**	**Up**
Actual Classification	**Down**	**10**	**2**	**1**
	None	**3**	**12**	**3**
	Up	**1**	**1**	**4**

Figure 15: Example of a confusion matrix for a three-class classification problem.

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.

4.5.1 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 internal hyperparameter tuning and model evaluation during development, while the test set is reserved for the final evaluation of the model's performance.

Explain more.

4.5.2 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.**

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.

CatBoost CatBoost is a gradient boosting algorithm that was found to perform well with particular feature sets and hyperparameter configurations [42]. Random Forest and Extra Trees were generally favoured throughout the project, but CatBoost provided competitive results in some scenarios. Since

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.

4.5.3 Adjusting Classification Thresholds

REDUNDANT AS WE ONLY CONSIDER 3-CLASS CLASSIFICATION. 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 16. 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.

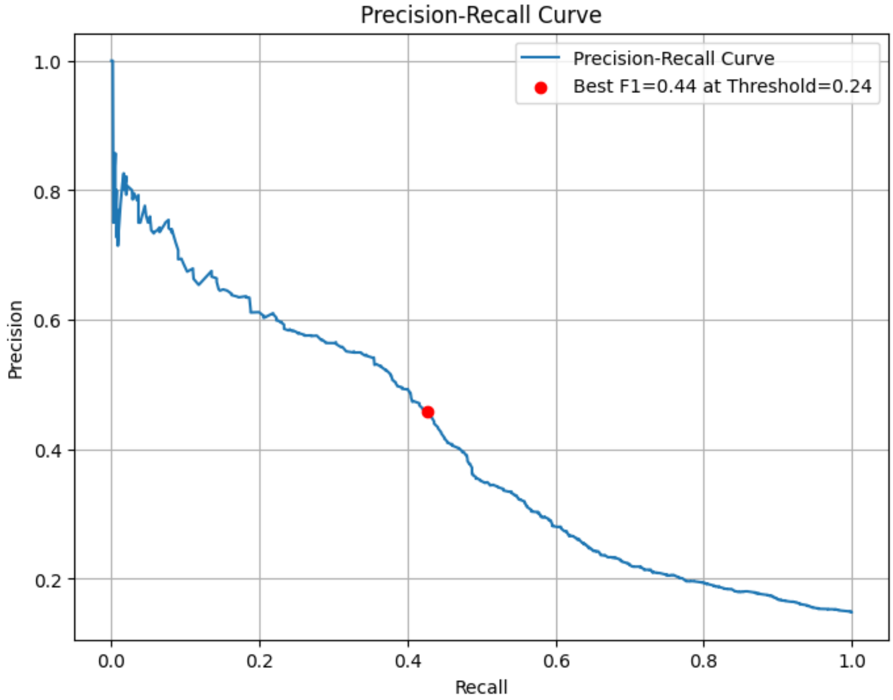


Figure 16: Typical Precision-Recall Curve with F1-score maximization point indicated.

4.5.4 Decision-bias tuning

After extending the prediction task from binary (up vs. none) to ternary (up, down, none), the model performance on the "up" class dropped substantially. This was expected, as discussed in Section 4.3.3, because "up" activations are rare and therefore provide the model with far fewer examples to learn from. As a result, the model became overly conservative and almost always favoured predicting the majority classes.

To address this issue, a simple post-processing method—here referred to as *decision-bias tuning*—was applied. The idea is similar in spirit to adjusting the decision threshold in binary classification, but adapted to the three-class setting. Instead of changing the model itself, we gently shift its decision boundary so that it becomes more willing to predict the "up" class.

The procedure is straightforward. After the model outputs its class probabilities, the predicted probability for the "up" class is multiplied by a bias factor (greater than one). This does not change the underlying model or its learned parameters; it only adjusts how the final class label is chosen. If the model was previously hesitating to predict "up" even when it assigned a moderate probability to it, this adjustment makes such predictions more likely.

The strength of the bias is selected using the validation dataset. Several candidate bias levels are tested, and for each of them we compute the resulting F1-score for the "up" class by keeping the other classes unchanged. The bias factor that gives the best validation performance is then used when making predictions on the test set. This tuning process helps correct the model's natural tendency to favour the majority classes and provides a controlled way of improving recall and overall F1-score for the under-represented "up" class, without degrading the performance on the remaining classes more than necessary.

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-macro score, we ensure that the trade-off between precision and recall across all classes is optimized for practical performance.

4.5.5 AutoGluon

5 Model Results

During the project's course, various models were developed and trained on differing datasets and feature sets. Much experimentation went into determining the best combination of factors to optimize practical performance and utility. Model iterations are evaluated based on practical metrics as well as comparisons to a naive baseline model. **Should naive model be mentioned earlier?** This model serves as a benchmark, providing a reference point against which more complex models can be compared.

Two dataset timeframes were primarily used for model training and evaluation: a post-March 4th 2025 dataset and a combined 2024-2025 dataset. The two-year dataset provides more data for training, potentially improving model performance and generalization. However, it also introduces some inconsistencies due to changes in data resolution over time, and the March 4th 2025 mFRR market transition [3]. The shorter dataset avoids these issues but offers less data for training. Both datasets have their advantages and drawbacks, and during the project, models were trained and evaluated on both to assess their performance under different conditions.

5.1 Naive Model

The naive model simply predicts that the activation state at time $t + 4$ is the same as at time $t - 4$, representing a strictly persistence-based approach. This approach is inspired by findings in literature discussed in Section 3, which highlight the auto-correlated nature of activation patterns [16]. Much of the motivation for this project is to explore whether more sophisticated models can outperform auto-correlation-based baselines by capturing additional relevant information from various features.

Table 3 shows the classification report for the naive model on the post-March 4th 2025 dataset test split. The model performs well, with a F1-macro score of 0.55. One can directly infer from the metrics that, in the test split at least, 62% of down-activations, 61% of no-activations, and 43% of up-activations are persistent from $t-4$ to $t+4$. This does not guarantee that it is persistent between these two time steps, but it is a strong auto-correlation indicator. Similar persistence values were observed in the entire dataset with 66%, 64%, and 39%, respectively. Down and no-activations are more persistent than up-activations, which makes up-activations even more challenging to predict. The "up" class is already the least frequent class, so the model has less data to learn from, and the lower persistence further complicates accurate predictions.

Interestingly, precision and recall are equal for all classes. This can also be seen in the confusion matrix in Figure 17, as it is close to being symmetric and respective rows and columns have similar sums. This phenomenon occurs because the confusion matrix basically represents a transition matrix for the activation states ($t - 4$ to $t + 4$). The symmetry reflects that, for instance, the number of none $\rightarrow$ up transitions is similar to up $\rightarrow$ none transitions over the dataset.

Class	Precision	Recall	F1-score	Support
down	0.62	0.62	0.62	1961
none	0.61	0.61	0.61	2519
up	0.43	0.43	0.43	792
accuracy			0.59	5272
macro avg	0.55	0.55	0.55	5272
weighted avg	0.59	0.59	0.59	5272

Table 3: Classification report for the naive last-observed-class baseline model.

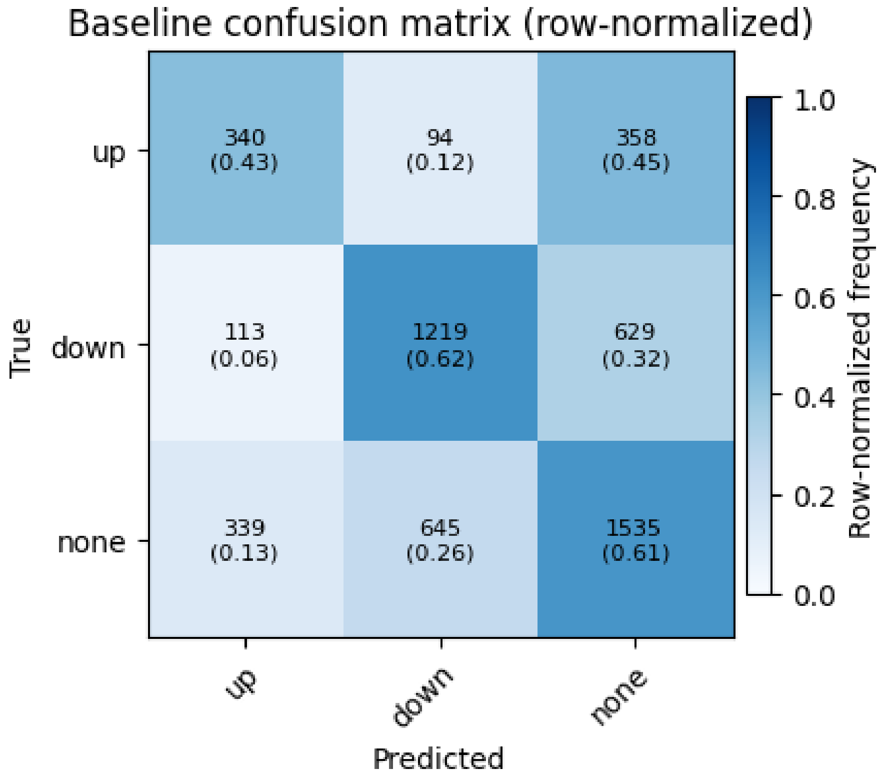


Figure 17: Confusion matrix (row-normalized) for naive model on post-March 4th 2025 dataset.

5.2 Machine Learning Model Results

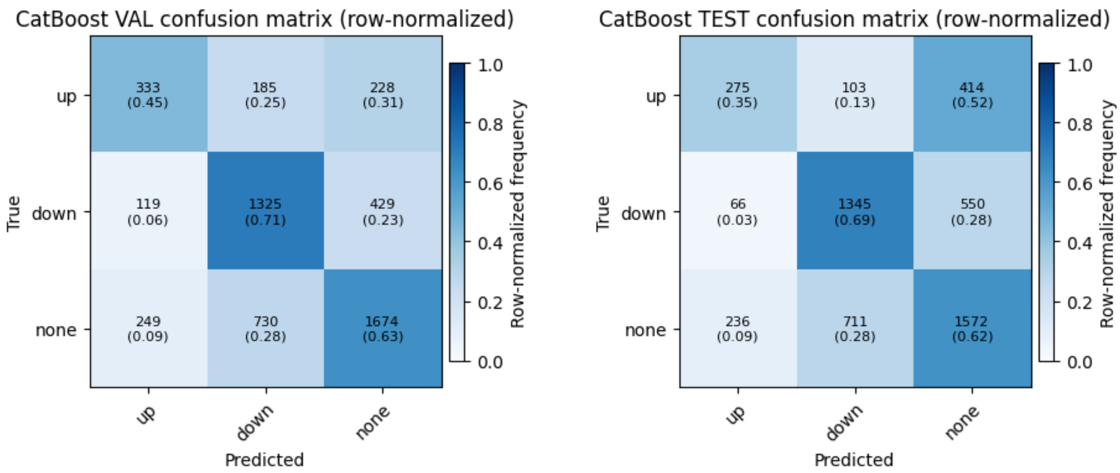
Models were trained using the open-source AutoML framework AutoGluon-Tabular [43]. Much of the appeal of AutoGluon lies in its method of ensembling multiple models and stacking them in multiple layers. Thus, individual models do not need to be considered in isolation, as the ensemble model often outperforms any single model. As a consequence, it is difficult to present results for individual model runs, as each iteration may consist of different features and model parameters. However, model performance remained relatively consistent across different runs, not improving drastically with new feature additions or model tuning. As a matter of fact, in the final model iterations, singular CatBoost models performed consistently on par with complex ensembles. Thus, since singular models are faster to train and evaluate, CatBoost models were used to explore various data and model configurations more extensively. Performance and evaluation of these models are presented in the following sections.

5.2.1 Model Performance

Table 4 and Figure 18 show a classifcation report and confusion matrices for a CatBoost model trained on the final featured post-March 4th 2025 dataset. Although scores improved slightly during model development, test set F1-macro scores consistently remained around 0.5-0.6. Across all model iterations, F1-macro scores never deviated significantly from the naive baseline model. The model does manage to catch notably more down-activations than the naive model, with a recall of 0.69 compared to 0.62. However, up-activation F1-scores are lowered. This highlights the imbalance challenge, and the up-class's lower predictability. No model iteration managed to significantly outperform the naive baseline on up-activation predictions.

Table 4: CatBoost performance on validation and test sets (classes: down, none, up).

Split / Class	Precision	Recall	F1-score	Support
Validation				
down	0.59	0.71	0.64	1873
none	0.72	0.63	0.67	2653
up	0.48	0.45	0.46	746
accuracy			0.63	5272
macro avg	0.59	0.59	0.59	5272
weighted avg	0.64	0.63	0.62	5272
Test				
down	0.62	0.69	0.65	1961
none	0.62	0.62	0.62	2519
up	0.48	0.35	0.40	792
accuracy			0.61	5272
macro avg	0.57	0.55	0.56	5272
weighted avg	0.60	0.61	0.60	5272



(a) Validation set

(b) Test set

Figure 18: Row-normalized confusion matrices for the CatBoost model on the validation and test splits of the post-March 4th 2025 dataset.

Although macro metrics indicate that the model performs only slightly better than the naive baseline, there is evidence suggesting that the model not only relies on persistence patterns. Precision and recall vary across classes, unlike the naive model. Higher recall and same precision for down-activations indicates that the model captures some down-activations that occur non-persistently, while maintaining precision. This is promising as it indicates that the dataset contain useful information beyond persistence patterns. Figure 19 shows a cross-tabulation of predicted classes at $t + 4$ against actual classes at $t - 4$, with accuracies for each entry in parantheses. This visualization essentially shows how many predictions the model makes based on persistence versus non-persistence, and how accurate these predictions are. Most predictions are indeed based on persistence, as indicated by the high values along the diagonal. However, there are cases where the model predicts a different class than the one observed at $t - 4$. It is clear to see that the model struggles most with predicting up-activations, as many of its attempt to predict transitions from none/down to up fail (0.23 and 0.19 accuracy, respectively). However, the model manages to, for instance, catch transitions from down to none with 0.66 accuracy. This is promising, but it is clear that more work is needed to improve the model's ability to predict non-persistent activation events.

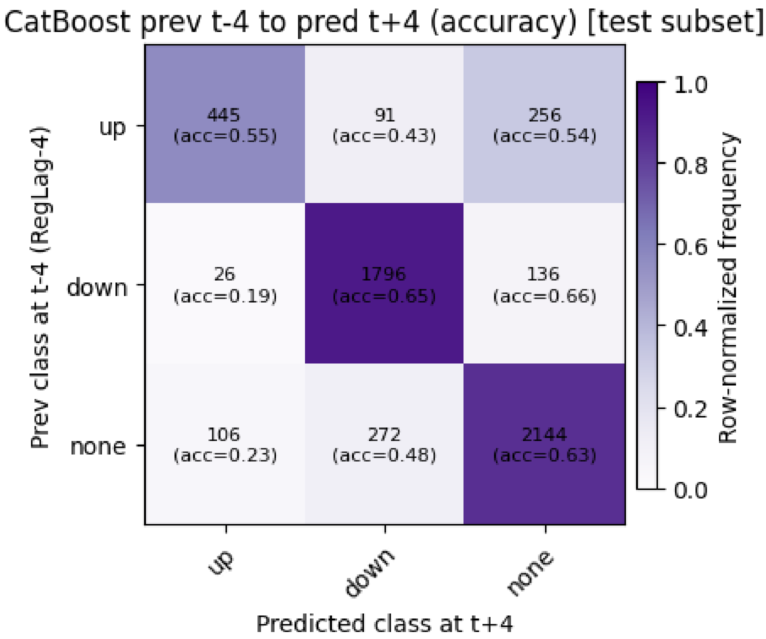


Figure 19: Confusion matrix (row-normalized) for CatBoost model on post-March 4th 2025 dataset.

5.3 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.

5.3.1 Extra Trees

Figure 20 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.

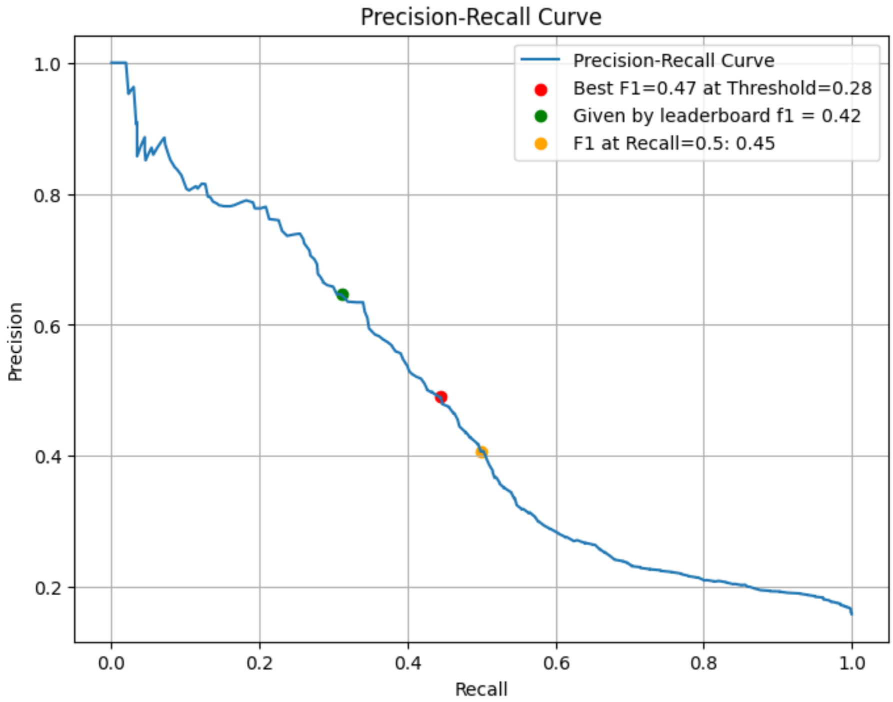


Figure 20: Precision-Recall Curve for Extra Trees model trained on 2025 dataset with highest F1-score.

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

Table 5 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.

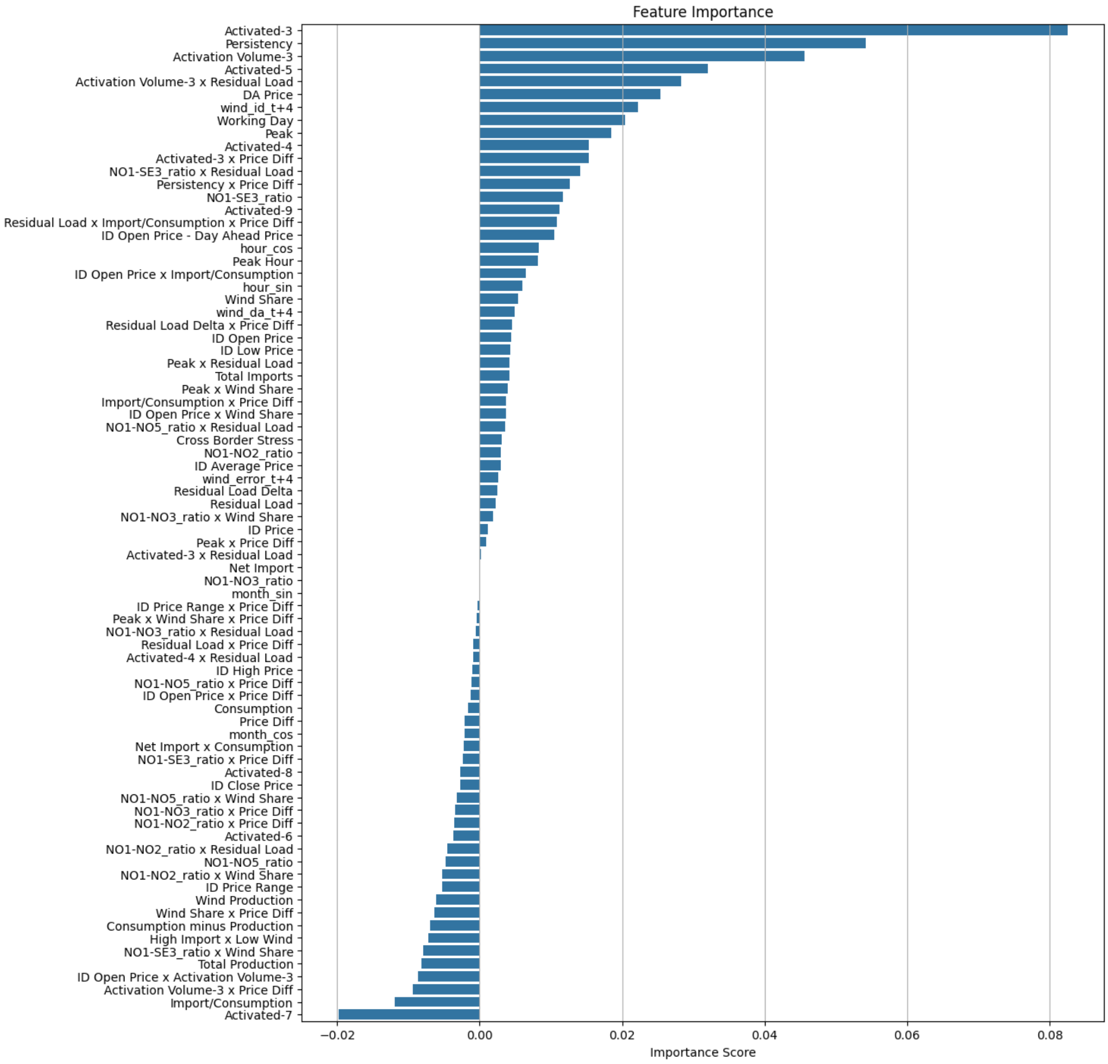


Figure 21: Feature importance for Extra Trees model trained on 2025 dataset with highest F1-score.

Figure 21 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.

5.4 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.

5.4.1 CatBoost Models

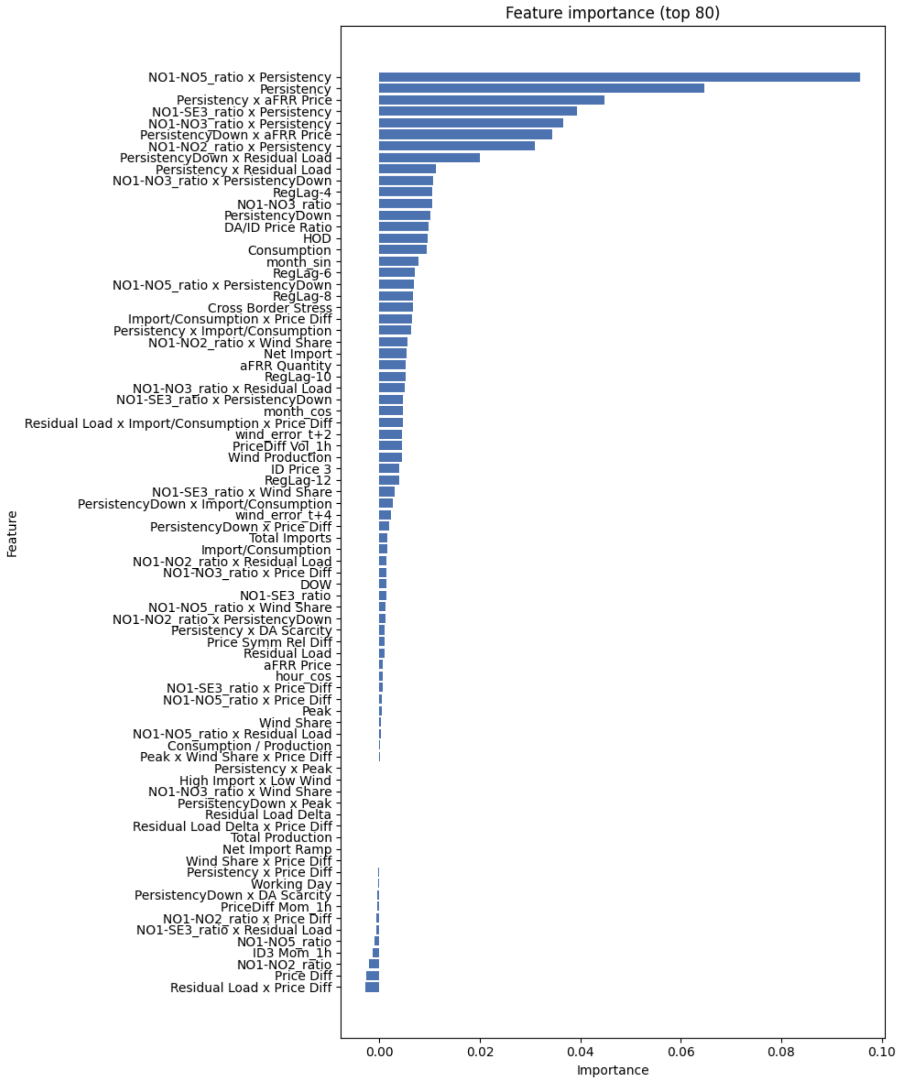


Figure 22: Feature importance for CatBoost model trained on 2024-2025 dataset after hyperparameter tuning.

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

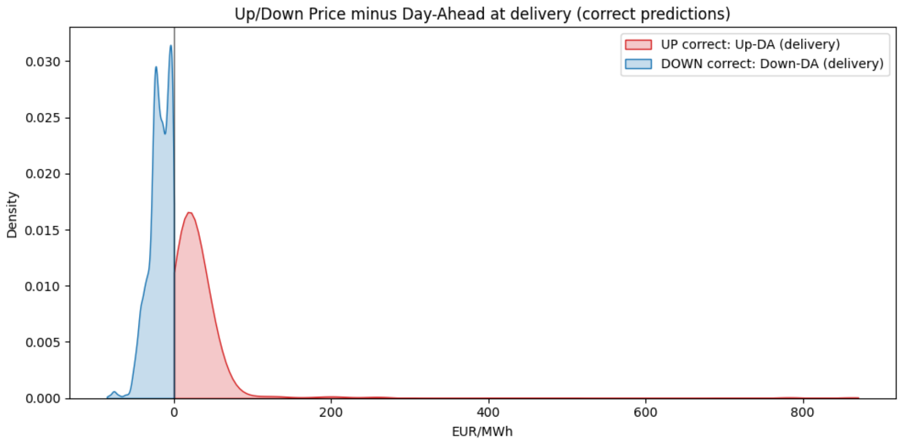


Figure 24: Up/down price minus day-ahead price distribution for CatBoost model trained on 2024-2025 dataset. quick_multiclass_cat_hpo

Figure 25 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 25: Performance metrics for CatBoost model on 2025 March 4th dataset at different confidence thresholds.

5.5 Correlation

I do not know where this will end up in the structure, but talking about correlation between features is key. Analysis has been done on correlation using different methods. Explaining corelation.

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] *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.

[4] *The Nordic power market introduces 15-minute balancing*, https://www.volue.com/news/nordic-power-market-introduces-15min-balancing. Accessed: 9th Dec. 2025.

[5] *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.

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

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

[8] L. Zhao, 'Event Prediction in the Big Data Era: A Systematic Survey', *ACM Computing Surveys*, vol. 54, no. 5, pp. 1–37, Jun. 2022, ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3450287 Accessed: 8th Dec. 2025.

[9] 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.

[10] 'Balancing market outlook 2030',

[11] *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.

[12] 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.

[13] 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.

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

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

[16] S. Backe, S. Riemer-Sørensen, D. A. Bordvik, S. Tiwari and C. A. Andresen, 'Predictions of prices and volumes in the Nordic balancing markets for electricity', in *2023 19th International Conference on the European Energy Market (EEM)*, Jun. 2023, pp. 1–6. DOI: 10.1109/EEM58374.2023.10161961 Accessed: 9th Dec. 2025.

[17] 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.

[18] J. Bankefors, 'Day-ahead modelling of the electricity balancing market',

[19] A. Khodadadi, H. Nordström, R. Eriksson and L. Söder, 'Investigating Reserve Dimensioning Approaches for Multi-Area Reserve Capacity Markets: A Nordic Case Study', Accessed: 9th Dec. 2025.

[20] T. Hagström, *Optimizing Risk-Aware Bidding Strategies for EV Fleets in the 15-Minute Nordic mFRR Market*. 2025. Accessed: 9th Dec. 2025.

[21] M. Håberg and G. Doorman, 'A stochastic mixed integer linear programming formulation for the balancing energy activation problem under uncertainty', in *2017 IEEE Manchester PowerTech*, Jun. 2017, pp. 1–6. DOI: 10.1109/PTC.2017.7980980 Accessed: 9th Dec. 2025.

[22] L. Irrmann, 'Analysis and Modelling of the Balancing Energy Market in the Nordics and Finland', Jun. 2023. Accessed: 11th Dec. 2025.

[23] A. Papavasiliou, A. Bouso, S. Apelfrojd, E. Wik, T. Gueuning and Y. Langer, 'Multi-Area Reserve Dimensioning Using Chance-Constrained Optimization', *IEEE Transactions on Power Systems*, vol. 37, no. 5, pp. 3982–3994, Sep. 2022, ISSN: 0885-8950, 1558-0679. DOI: 10.1109/TPWRS.2021.3133102 Accessed: 11th Dec. 2025.

[24] *Risk Constrained Trading Strategies for Stochastic Generation with a Single-Price Balancing Market*, https://www.mdpi.com/1996-1073/11/6/1345. Accessed: 11th Dec. 2025.

[25] I. Pavić, H. Pandžić and T. Capuder, 'Electric Vehicle Aggregator as an Automatic Reserves Provider Under Uncertain Balancing Energy Procurement', *IEEE Transactions on Power Systems*, vol. 38, no. 1, pp. 396–410, Jan. 2023, ISSN: 1558-0679. DOI: 10.1109/TPWRS.2022.3160195 Accessed: 11th Dec. 2025.

[26] G. Klæboe, A. L. Eriksrud and S.-E. Fleten, 'Benchmarking time series based forecasting models for electricity balancing market prices', *Energy Systems*, vol. 6, no. 1, pp. 43–61, Mar. 2015, ISSN: 1868-3975. DOI: 10.1007/s12667-013-0103-3 Accessed: 12th Dec. 2025.

[27] M. Olsson and L. Soder, 'Modeling Real-Time Balancing Power Market Prices Using Combined SARIMA and Markov Processes', *IEEE Transactions on Power Systems*, vol. 23, no. 2, pp. 443–450, May 2008, ISSN: 0885-8950, 1558-0679. DOI: 10.1109/TPWRS.2008.920046 Accessed: 12th Dec. 2025.

[28] J. D. Croston, 'Forecasting and Stock Control for Intermittent Demands', *Operational Research Quarterly (1970-1977)*, vol. 23, no. 3, pp. 289–303, 1972, ISSN: 0030-3623. DOI: 10.2307/3007885 JSTOR: 3007885. Accessed: 12th Dec. 2025.

[29] S. Backe, S. Riemer-Sørensen, D. A. Bordvik, S. Tiwari and C. A. Andresen, 'Predictions of prices and volumes in the Nordic balancing markets for electricity', in *2023 19th International Conference on the European Energy Market (EEM)*, Jun. 2023, pp. 1–6. DOI: 10.1109/EEM58374.2023.10161961 Accessed: 9th Dec. 2025.

[30] 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.

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

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

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

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

[35] *Transition to 15-minute Market Time Unit (MTU)*, https://www.nordpoolgroup.com/en/trading/transition-to-15-minute-market-time-unit-mtu/. Accessed: 8th Dec. 2025.

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

[37] 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.

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

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

[40] *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.

[41] M. Grandini, E. Bagli and G. Visani, *Metrics for Multi-Class Classification: An Overview*, Aug. 2020. DOI: 10.48550/arXiv.2008.05756 arXiv: 2008.05756 [stat]. Accessed: 9th Dec. 2025.

[42] 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.

[43] N. Erickson et al., *AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data*, Mar. 2020. DOI: 10.48550/arXiv.2003.06505 arXiv: 2003.06505 [stat]. Accessed: 13th Dec. 2025.