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 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 March 4th 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 to imbalances [4]. This change also means that market participants must now make bidding decisions at a finer temporal granularity, calling for improved short-term forecasting.

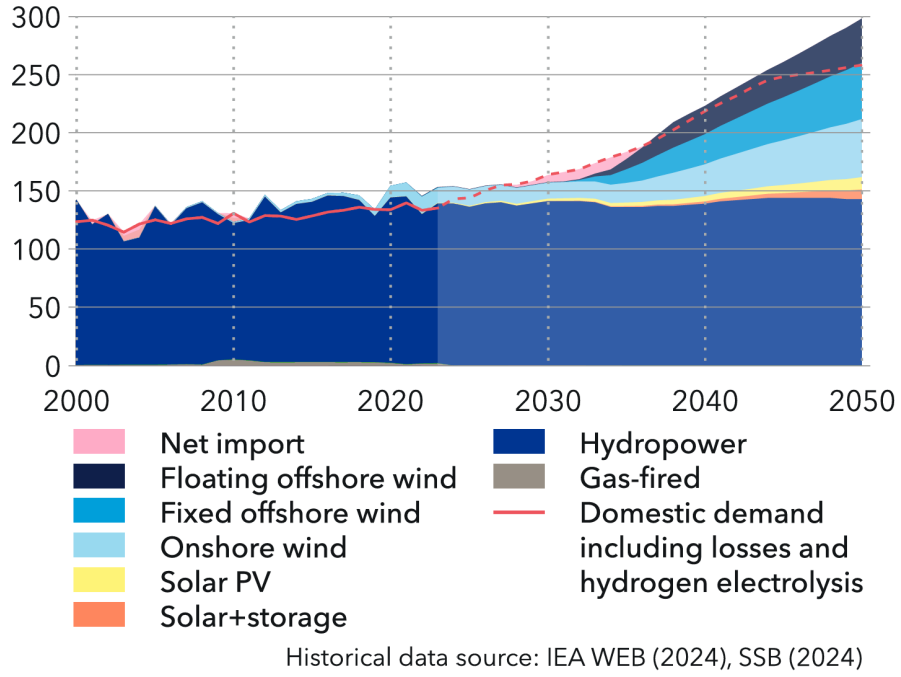


Figure 1: Norway electricity supply by power station and net imports with projections to 2050 [5]

1.2 Motivation

mFRR activations have direct economic implications for market participants. Capacity markets provide stable income, whereas energy activation markets offer higher but uncertain returns. Anticipating whether mFRR up- or down-regulation is likely therefore affects how flexible resources should be allocated between capacity commitments and activation market participation.

This trade-off is particularly relevant for aggregators managing distributed flexible assets such as electric vehicles, heat pumps, and batteries. Committing flexibility to the mFRR capacity market secures a fixed payment but binds resources for the contracted period. In contrast, participation in the activation market preserves greater operational flexibility, as resources are only constrained during actual activations. Aggregators must therefore balance guaranteed capacity revenues against the expected value of future activation opportunities, under substantial uncertainty.

The transition to a 15-minute market time unit represents not only a structural change in market operation but also an opportunity for more informative short-term decision support. Finer temporal resolution allows activation signals to better reflect rapidly evolving system conditions, creating richer and more frequent market signals. At the same time, this increased granularity elevates the importance of forecasts, as market participants must make decisions over shorter horizons and under greater sensitivity to short-term system dynamics. As a result, short-term activation forecasting becomes both more relevant and potentially more valuable for participants seeking to adapt their bidding strategies to real-time market participation.

Reinforcement learning motivation here?

1.3 Research Question

The central research question of this study is: To what extent can short-term mFRR activation direction (up-regulation, down-regulation, or no activation) be predicted at 15-minute resolution using only information available to market participants at decision time?

1.4 Outline

This report is structured as follows: Chapter 2 provides an overview of the Nordic balancing markets and machine learning concepts. Chapter 3 reviews existing literature on imbalance and activation forecasting and identifies the research gaps addressed in this study. Chapter 4 details the project's 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.

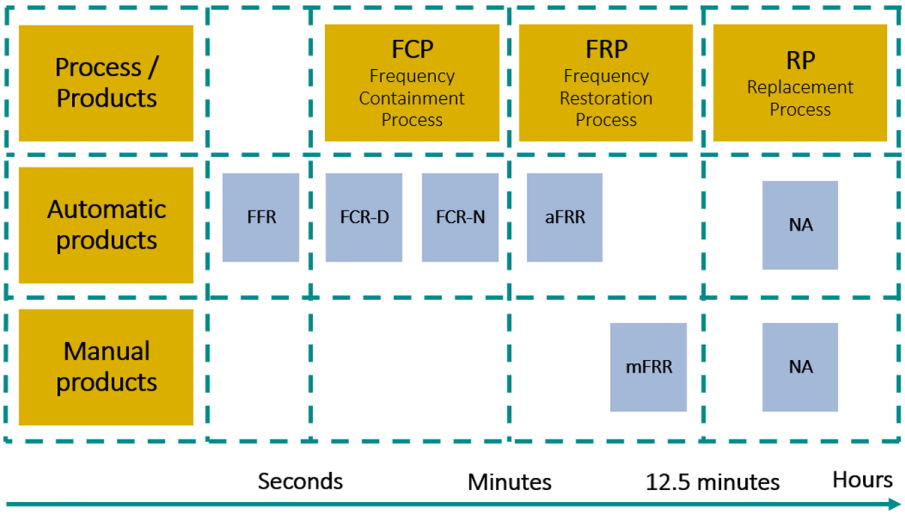


Figure 2: The Nordic balancing market hierarchy, illustrating the different reserve types and their activation times [2].

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.

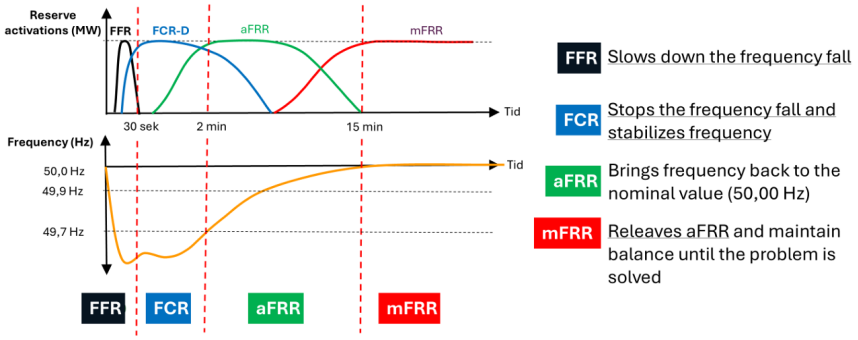


Figure 3: Illustration of the different reserve types and their activation times [2].

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 (CM). 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 (EAM). 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 unrealistic results if it has access to future data points when making predictions for a given time step.

2.2.2 Class Imbalance

Class imbalance is a common challenge in classification tasks, particularly in real-world applications where certain classes occur much more frequently than others [9]. In the context of mFRR activation prediction, 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

Given the characteristics of the problem, tree-based machine learning models are well suited for activation direction prediction. Such models can capture nonlinear relationships and complex interactions between features without requiring strong parametric assumptions. They are also robust to differences in feature scale and can naturally accommodate heterogeneous inputs, including prices, flows, forecasts, and lagged indicators.

Importantly, tree-based models integrate effectively with feature-engineering approaches that encode temporal information through lagged values and persistence measures. Rather than modeling sequential dynamics explicitly, these models infer temporal structure indirectly from engineered features that summarize recent system behavior. This makes them particularly suitable for settings where direct access to high-resolution real-time data is limited.

3 Literature Review

This chapter reviews the literature relevant to balancing-market forecasting. It first outlines the unique characteristics of balancing activation markets and the challenges they present. Existing methodologies for handling these challenges are then presented, before finally identifying specific research gaps that this study and future work can address.

3.1 mFRR Energy Activation Market Characteristics

The mFRR energy activation market distinguishes itself from the capacity market 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 [10]. 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. [10] 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 note, 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 [11] they report that mFRR capacity demand has increased and will continue to 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 4 March 2025, an endeavour aimed at enhancing market efficiency and integrating renewable energy sources more effectively [12]. 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 [13].

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 [13]. Their findings imply that imbalances are now corrected more accurately and efficiently on shorter 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

For mFRR EAM market participants, the most relevant target is often the activated energy volumes. This target is crucial, as it directly informs bidding strategies and operational decisions. However, activated volumes are conditional on direction because bids must be placed in the actually activated direction to be eligible for activation. In practice, the data are also strongly zero-inflated, meaning most intervals have no activation, and the distribution of non-zero volumes is different in up- and down-regulating events. When direction is not modelled explicitly, the target distribution becomes a mixture of up-, down-, and no-activation regimes, which can inflate apparent noise and reduce the usefulness of forecasts for decision-making. Consequently, modelling activation-direction uncertainty explicitly is an important component of balancing-market forecasting.

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 characterized by transition probabilities.

The following sections review these modelling families, beginning with direct 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. [14] 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 [15]. 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 [16] 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 [17] examine imbalance-volume forecasting from a market-participant-oriented standpoint. They also develop a LSTM model covering several Nordic bidding zones. Their analysis yields three relevant insights. First, balancing volumes are relatively autocorrelated: 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. [18] 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 [19]. 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 depend on activation direction, meaning that models that do not explicitly model activation direction uncertainty must implicitly learn it from noisy, zero-inflated data. This can degrade forecast quality, especially around directional switches, and it can encourage regression models to predict values close to zero most of the time. These limitations motivate work that treats activation uncertainty as a modelling concern. The subsequent sections review how existing literature has addressed activation uncertainty.

3.2.2 Scenario-Based Activation Models

Scenarios are often used to handle uncertainty in optimization problems. Possible future outcomes 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 [20] and stochastic scheduling or bidding frameworks [21].

In [22], 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.

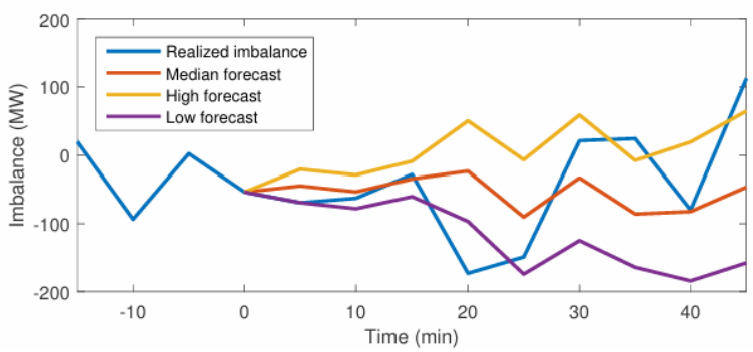


Figure 4: Imbalance forecast scenarios.

Figure 4 is reproduced from Håberg and Doorman [22].

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, as outlined in this literature review, are challenging to develop. Overall, 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.

extbfConfused about this section. Feels like it doesn't fit here as it does not directly address activation direction uncertainty. Might have misunderstood?

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 to analyze and model the Nordic balancing markets [23]. In this study, *regulation states* (up, down, none) are sampled based on historical frequencies, before activation volumes are drawn from a modelled distribution conditional 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 unlikely to be representative of short-term activation behaviour. 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.

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) [24] apply chance-constrained optimization to the problem of reserve dimensioning in a multi-area power system. Here, uncertainty described by scenarios is 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 [22], only modelled implicitly through the sign of the continuous imbalance scenarios.

Browell (2018) [25] 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

Pavić et al. (2023) argue that deterministic reserve activation models inaccurately represent 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 [26]. 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 are gathered and probability distributions are constructed as in Figure 5, representing the likelihood of different activation ratios. Relevant statistics, such as the mean, max, and quantiles, are used as inputs for their models.

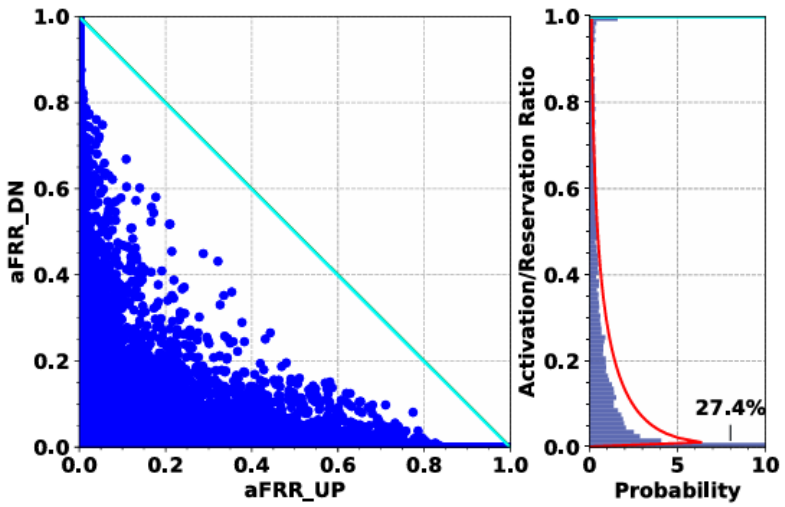


Figure 5: Activation ratio uncertainty ranges for aFRR up.

Figure 5 is reproduced from Pavíc et al. [26].

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. Pavić et al. operate in the context of FCR and aFRR reserves, where there is always an activated imbalance in one direction [26]. 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 [27]. In this study, they separate relevant work into two families: models explicitly modelling balancing state and those modelling 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 modelling. 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 modelled with separate transition matrices, while longer durations used a common matrix [28].

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 [29], only distinguishing between activation and no activation, thereby discounting direction. This model distinguishes between up- and down regulations in a separate price- or volume process.

One-hour-ahead predictions were shown to be quite accurate, predicting correctly 63% and 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 [30].

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 [31] examine the general price formation in intraday and mFRR markets. Among other explorations, they develop logistic regression and ANN (Artificial Neural Network) models to make day-ahead predictions for 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 day-ahead activation direction. In conclusion, they suggest that closer-to-real-time predictions could likely provide insights to market participants.

Porras (2025) [32] 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.

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/a-FRR)	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 (day-ahead)	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 transitioned to a 15-minute mFRR market time unit (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 conditional on activation direction, which is not modelled explicitly in most studies. Results are therefore affected by directional uncertainty, making predictions noisier. In addition, activation data are often zero-inflated, so a single direct regression model may be biased toward predicting near-zero values. This study therefore argues that explicit direction modelling is useful for market participants by itself, and that it provides a natural foundation for subsequent conditional forecasting of activation magnitudes (and ultimately prices) once direction has been resolved.

In other words, a promising participant-oriented approach is to first model the discrete activation direction (up/down/none) under strict information constraints, and then model activation volumes conditional on the predicted direction. The literature contains elements of this idea (e.g., sequential frameworks for related targets), but it remains unclear how well such a direction-first approach performs for Nordic mFRR at the new 15-minute resolution using only participant-feasible inputs.

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 [32], 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.

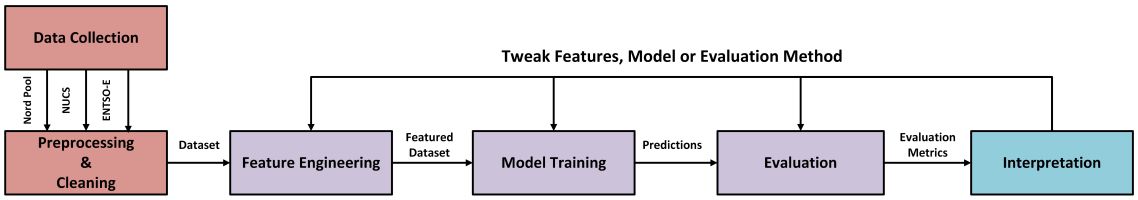


Figure 6: Overview of the methodological approach used in this project.

Methodological specifics are detailed in the subsequent sections, which follow the logical structure of the workflow illustrated in Figure 6. Each section elaborates on one stage of the approach, including data sources, preprocessing, feature construction, and model evaluation. While the methodology was implemented programmatically, the focus of the following sections is on the underlying methodological choices, assumptions, and constraints that govern the modeling process rather than on implementation details.

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 [33]. 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) [34]. 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 real-time data from the NUCS API for potential future real-time applications (**Appendix?**).

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.

mFRR CM. Capacity market data for mFRR was hard to come by, but eventually, NUCS was found to provide historical mFRR capacity market prices and volumes for the NO1 bidding zone through their API. As discussed in the literature review, Svedlindh and Yngvesson [31] found that

mFRR capacity market data was their most important feature for predicting mFRR activations. This study conducted day-ahead predictions, however, where capacity market data is more relevant. In a closer-to-real-time setting, the importance of capacity market data may be reduced, as the system state is better known closer to real-time. Still, mFRR capacity market data can provide valuable insights into the expected balancing needs 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 [35].

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 [36]. 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.

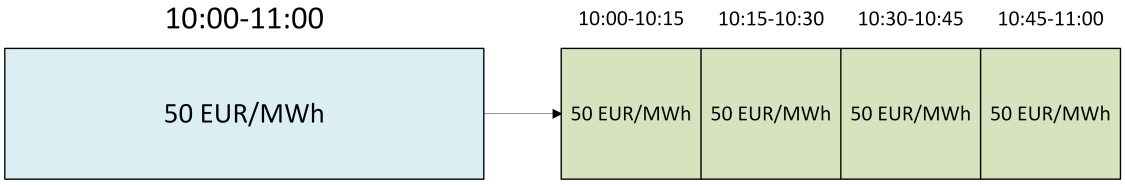


Figure 7: Visualization of resampling from 1-hour to 15-minute resolution using forward filling.

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 [37].

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

Activation Direction 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 class imbalance is a critical consideration for the prediction task, which in general makes it more challenging for models to accurately predict the minority class [9]. The predominance of no-activation contextualizes the zero-inflated nature of the imbalance volumes, as discussed in the literature review.

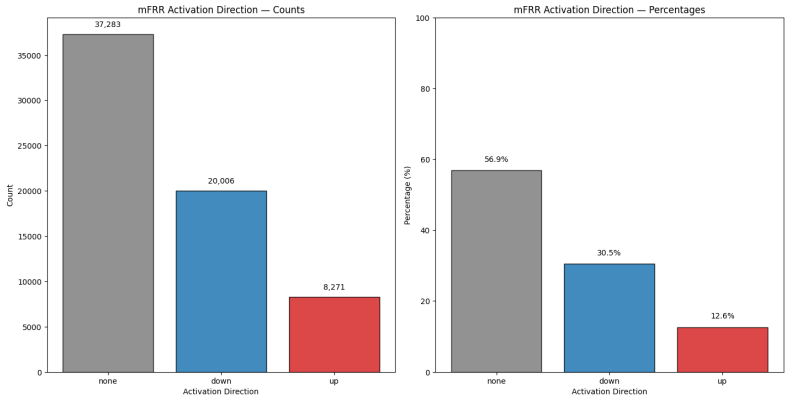


Figure 8: mFRR activation direction distribution.

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.**

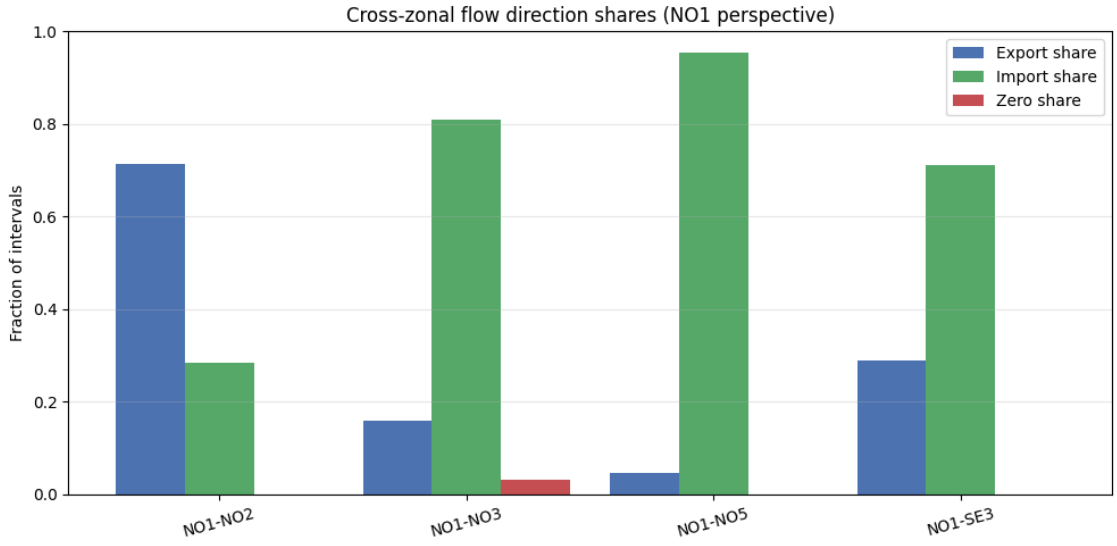
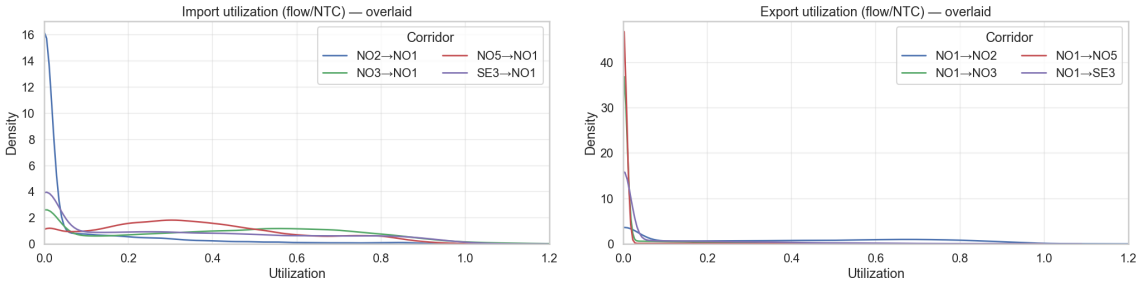


Figure 9: Cross-zonal flow distributions for the NO1 bidding zone.

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.**

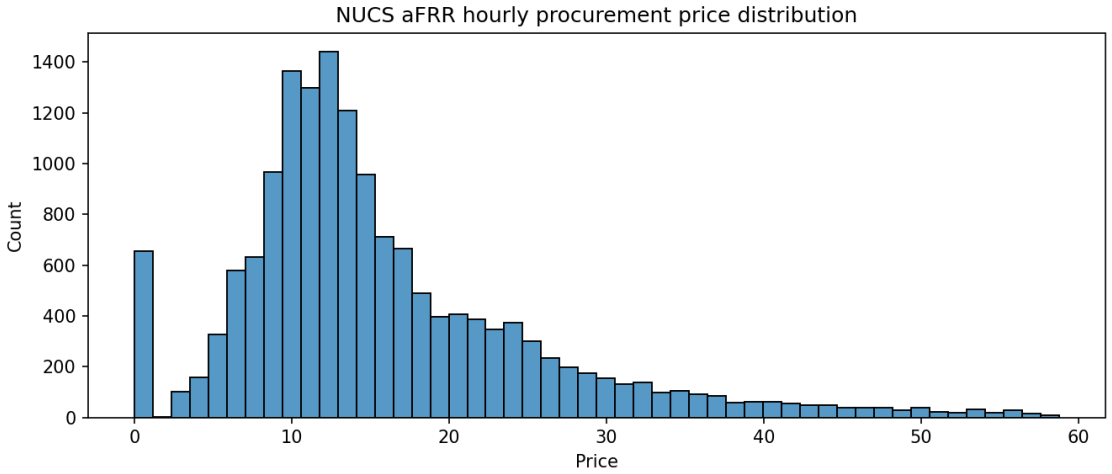


Figure 11: A histogram of hourly aFRR procurement prices for the NO1 bidding zone from NUCS data.

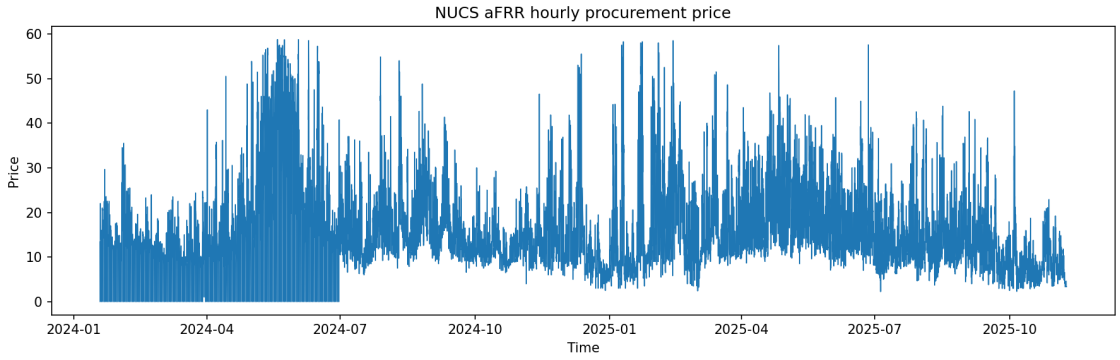


Figure 12: Hourly aFRR procurement prices for the NO1 bidding zone from NUCS data.

Correlation Analysis

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 a target variable, e.g., 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 imbalance volumes are known to be highly auto-correlated [17], even delayed activation data can provide valuable information about current trends. Figure 13 illustrates the time restrictions on feature availability for predicting mFRR activations at time $t+4$ based on data available at time t. Most real-time-relevant information is delayed by an hour, while accepted balance market activation volume and price are available at $t-1$, making it the most recent activation-related data point accessible at prediction time.

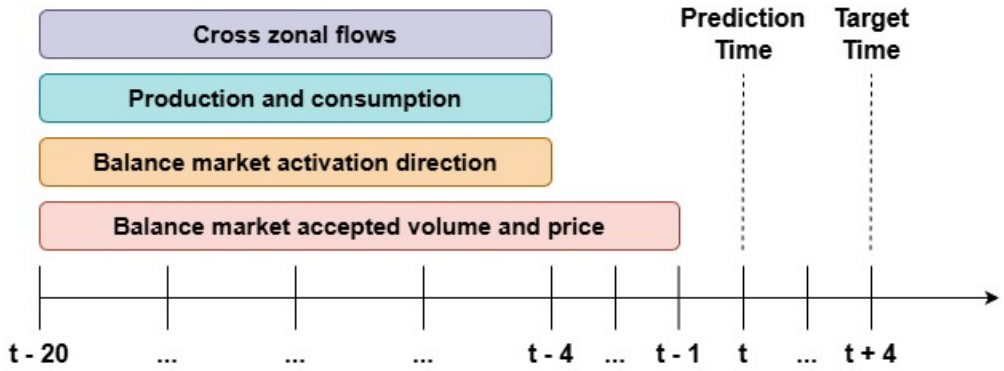


Figure 13: Illustration of time restrictions on feature availability for predicting mFRR activations at time $t+4$ based on data available at time t.

An additonal complication is that some data sources are made available only on certain times of the day. mFRR capacity market data for the following day, for instance, is published after the capacity market gate closure at 23:00 (**check**) each day. At 01:00 the next day, capacity market data is available for the entire day ahead, but at 22:00 the same day, only data for the next two hours is available. This data is thus not only available for target time $t + 4$. This complicates feature engineering, as the model must be able to handle features that are only partially available depending on the time of day. Table 3 summarizes the release times of various market datasets relative to the delivery day D. In this project, these timing constraints are not taken into account when engineering features, but they should be considered in future work to ensure that models can operate under real-world data availability conditions. **I can implement this, but i dont know if time allows it.**

Table 3: Market data release and availability times relative to delivery day D [38].

Market / Dataset	**Release Time**	**Relative to Delivery**
mFRR Capacity Market	23:00	$D-1$
Day-Ahead Auction	12:00	$D-1$
Intraday Auction 1 (IDA1)	15:00	$D-1$
Intraday Auction 2 (IDA2)	22:00	$D-1$
Intraday Auction 3 (IDA3)	10:00	D

Actually activation: t-2, production/flows t - 2, consumption t-3? Also is this the optimal ordering of sections?

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.

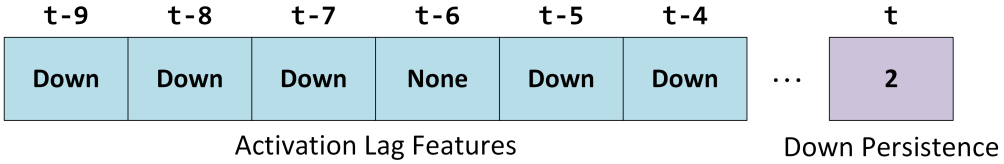


Figure 14: Illustration of down persistence feature calculation based on lagged activation features. Here, the down persistence at time t is 2, as there have been down-activations in the two most recent intervals ($t-4$ and $t-5$), before an interval with no activation at $t-6$.

Persistence feature capture consecutive activation trends effectively, but they do have limitations. Singular lagged activation features 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.

4.4.2 Cross-zonal flow features

Cross-zonal flow features capture information about electricity flows between different zones or regions in the power grid- in this case, in and out of the NO1 bidding zone. These flows can indicate the grid's stress level and influence mFRR activations. For instance, high inflows into NO1 may signal increased demand or generation shortages, potentially causing upregulating activations. Conversely, high outflows may indicate surplus generation, potentially causing downregulating activations. It is unlikely, though, that cross-zonal flow features alone can predict activations. Combining them with available transfer capacity provides a picture of how close the grid is to its operational limits. For instance, if the inflow into NO1 is close to the maximum available transfer capacity, only a small margin remains for additional inflows, which could increase the likelihood of upregulating activations. Such situations often occur in zones that are short, i.e. zones where consumption exceeds production. In such cases, the grid operator may need to activate expensive mFRRreserves 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. There is thus value in transforming such features to be on a similar scale [39]. 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 $\mathrm{ATC}_i(t)$ be the available transfer capacity for that interconnection at time t. The capacity-normalized flow is then

$$F^i_{\mathrm{ratio}}(t) = \frac{F_i(t)}{\mathrm{ATC}_i(t)}$$

Values of $F^i_{\mathrm{ratio}}(t)$ 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.

4.4.3 Temporal features

Temporal features capture time-related patterns in the data. These features help the model un derstand 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 [40]. Hourly features are, for instance, encoded as:

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

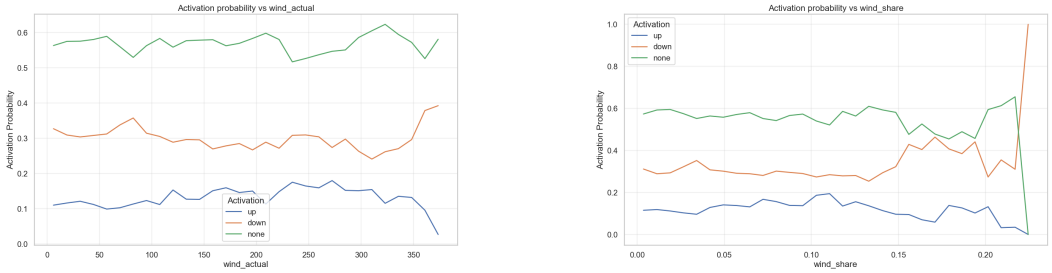
Monthly features are encoded similarly, using 12 as the divisor instead of 24.

4.4.4 Price features

Price features capture information about the various electricity market prices. The mFRR ac tivaiton 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.

4.4.5 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 15: Distributions of mFRR activations as functions of realized wind production and wind share.

4.4.6 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.

4.4.7 Interaction features

Interaction features are created by combining two or more existing features to capture complex relationships. Many such interactions were experimented with, mostly by subtracting, multiplying, or dividing pairs of features that were theoretically expected to have meaningful interactions. Price features were, for instance, combined by calculating price spreads or ratios between different market prices (day-ahead, intraday, aFRR, mFRR). Consumption and production features were also combined to create features that capture net load or supply-demand imbalances. Many more are possible, but only the most promising interaction features were included in the final feature set to make feature analysis more interpretable.

4.4.8 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

4.5 Evaluation Framework

Classification problems are often evaluated using accuracy, precision, recall, and F1-score. These metrics are defined as follows [41]:

- **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 [42].

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 16. 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 16: 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. 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 persistence-based trends. If a model has no such pattern recognition capabilities, it is of little use.

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. The transition metric evaluates the model's ability to predict *transitions* between classes. This is measured by looking at all sequential time interval pairs $t-4$ and $t+4$ with different class labels. If the model predicted the class label for time interval $t+4$ correctly, it is counted as a successfully predicted transition. This is a crucial metric as it indicates its non-persistence predictive capabilities.

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 reliably evaluating machine learning models. In this study, the dataset is partitioned into training, validation, and test sets in chronological order, with proportions of 60%, 20%, and 20%, respectively (see Figure 17). This temporal split is essential to avoid data leakage and to ensure that model performance is assessed on truly unseen future data. The training set is used to fit the models, the validation set is used for tuning and model selection during development, and the test set is held out until the end for an unbiased final evaluation of predictive performance.

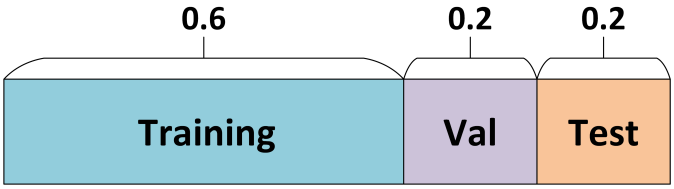


Figure 17: Temporal data splitting into training, validation, and test sets.

Some variation of cross-validation has the potential provide more generalizable results, as singular validation splits may not fully capture the variability in time series data. Porras (2025) applied a time series cross-validation, ensuring that the models were always trained on past data and validated on future data [32]. An analysis of this is presented in the results chapter.

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. **Section not finished.**

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 [43]. 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

Decision thresholds are the thresholds at which a model assigns class labels based on predicted probabilities. Although classification models output discrete class labels, they do so based on underlying predicted probabilities for each class. These thresholds can be adjusted to optimize certain performance metrics, such as precision, recall, or F1-scores. This is particularly important in imbalanced classification problems, where the default threshold (usually 0.5) may underpredict the minority class [44]. Figure 18 illustrates a typical precision-recall curve for a binary classification problem (up-activation vs. not up-activation), showing how precision and recall vary with different classification thresholds. The red dot indicates the point (threshold of 0.24) where the F1-score is maximized, representing an optimal balance between precision and recall. It is important that threshold tuning is not performed on the test set, as this would lead to overfitting and an overly optimistic estimate of model performance. Instead, threshold tuning should be carried out using the validation set, as is done in this project, ensuring that the test set remains a completely unseen dataset for final model evaluation.

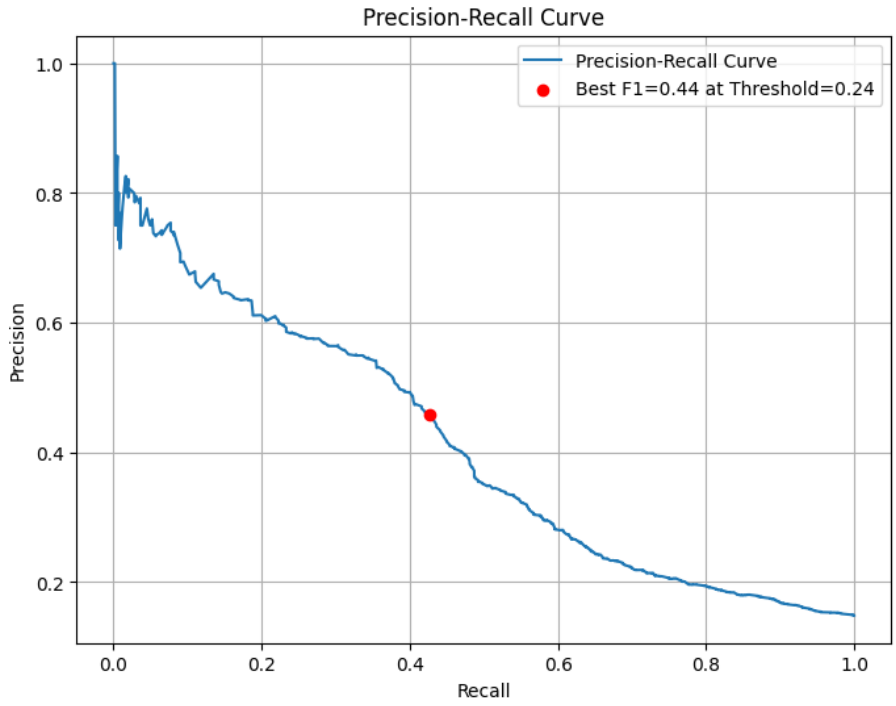


Figure 18: Typical binary Precision-Recall Curve with F1-score maximization point indicated.

When expanding from binary to three-class classification (up, down, none), there are more degrees of freedom when it comes to adjusting decision thresholds. However, the main challenge remains the same: the model tends to underpredict the minority class, which in this case is the "up" activation class. Thus, the "up" class is post-tuned on the validation set. In a multi-class setting, it was observed that the model hesitated even more in predicting the minority class, often achieving recall scores below 0.1 for the "up" class. An automatic F1-macro score maximization procedure by method of up-weighting the "up" class probabilities was therefore implemented to mitigate this issue. Thus, the "up" class is predicted more often to the extent that the overall F1-macro score is maximized on the validation set.

The rationale behind this method is that the model picks up on up-activation patterns to some extent, but as these patterns occur infrequently, the model is not confident enough to predict the "up" class often enough. By up-weighting the predicted probabilities for the "up" class, the model is encouraged to predict "up" more frequently, thereby increasing recall for this class. This has the to be detrimental to other class predictions, but if the trade-off is optimized for F1-macro score, the overall performance across all classes can be improved. There is no guarantee that this tuning will generalize onto the test set, but it provides a systematic way to address the underprediction of the minority class based on validation set performance.

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 [17]. 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 4 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 19, 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 4: Classification report for the naive last-observed-class baseline model.

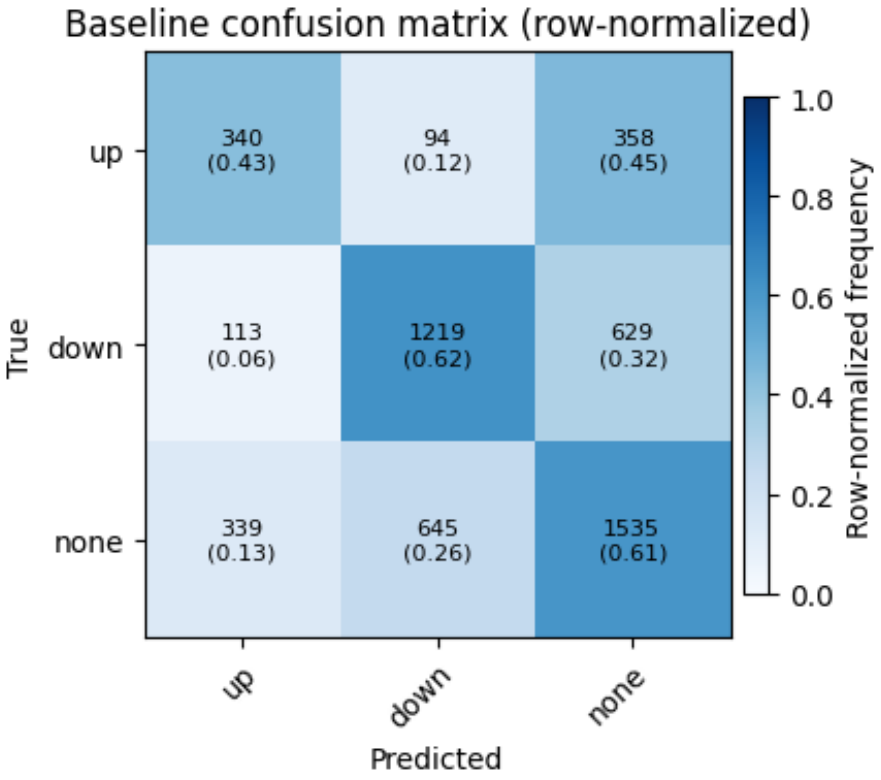


Figure 19: Confusion matrix (row-normalized) for naive model on post-March 4th 2025 dataset.

5.2 Machine Learning Model Results

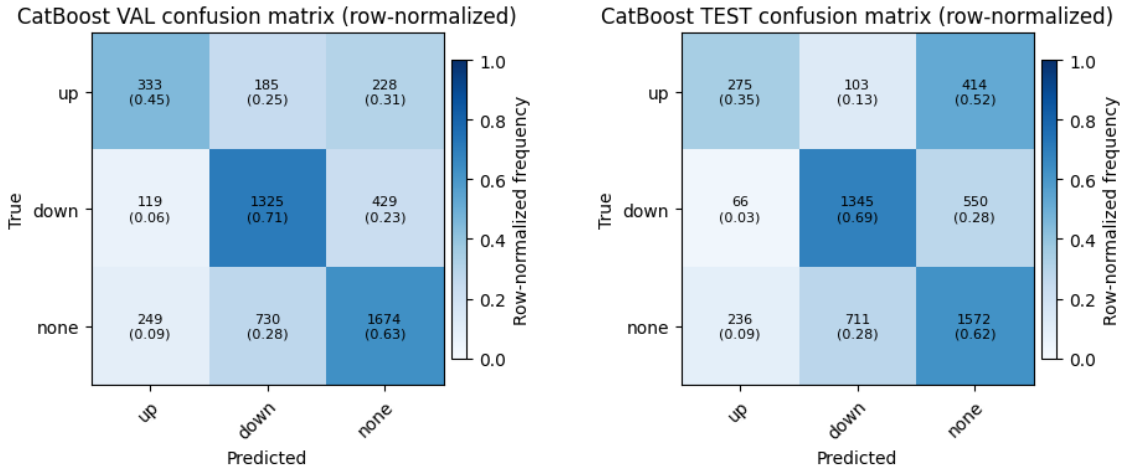
Models were trained using the open-source AutoML framework AutoGluon-Tabular [45]. Much of the appeal of AutoGluon lies in its method of ensembling multiple models and stacking them in multiple layers. Thus, individual models do not need to be considered in isolation, as the ensemble model often outperforms 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 Evaluation

Table 5 and Figure 20 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, validation and 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. Comparing the baseline and CatBoost test set confusion matrices underscore this point as they barely differ. 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 5: 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 20: 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 21 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 $t-4$, thus attempting to predict transitions. 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.

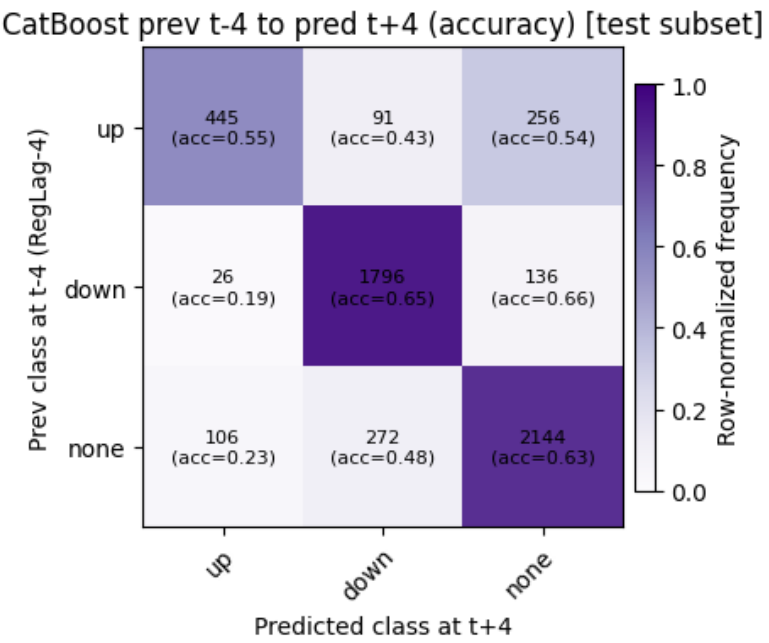


Figure 21: Confusion matrix (row-normalized) for CatBoost model on post-March 4th 2025 dataset.

5.2.2 Feature Importance

Figure 22 shows the top 40 most important features based on their feature importance scores from the CatBoost model trained on the post-March 4th 2025 dataset. There is seemingly a diverse range of features contributing to the model's predictions. Lag features, particularly the nearest regulation direction lag features ($t-4$, $t-5$, $t-6$), stand out as the most important. Persistence features are included in this particular feature set, and they also rank highly. These are expected to be the most decisive features as they explain the bulk of the auto-correlation patterns in the data.

Day-ahead price-related features distinguish themselves as the most important non-persistence features. Features such as PriceUp $-$ DA, PriceDown $-$ DA, and PriceUp/DA rank highly. These features represent the relationship between the day-ahead price and the $(t-1)$ up/down regulation prices. This suggests that the model leverages recent discrepency between day-ahead and balancing prices to inform its predictions. This makese sense, since big differences between regulation prices to day-ahead prices differences indicate big imbalances. The Up-Down Price Skew feature also ranks highly, further emphasizing the importance of recent price-related features.

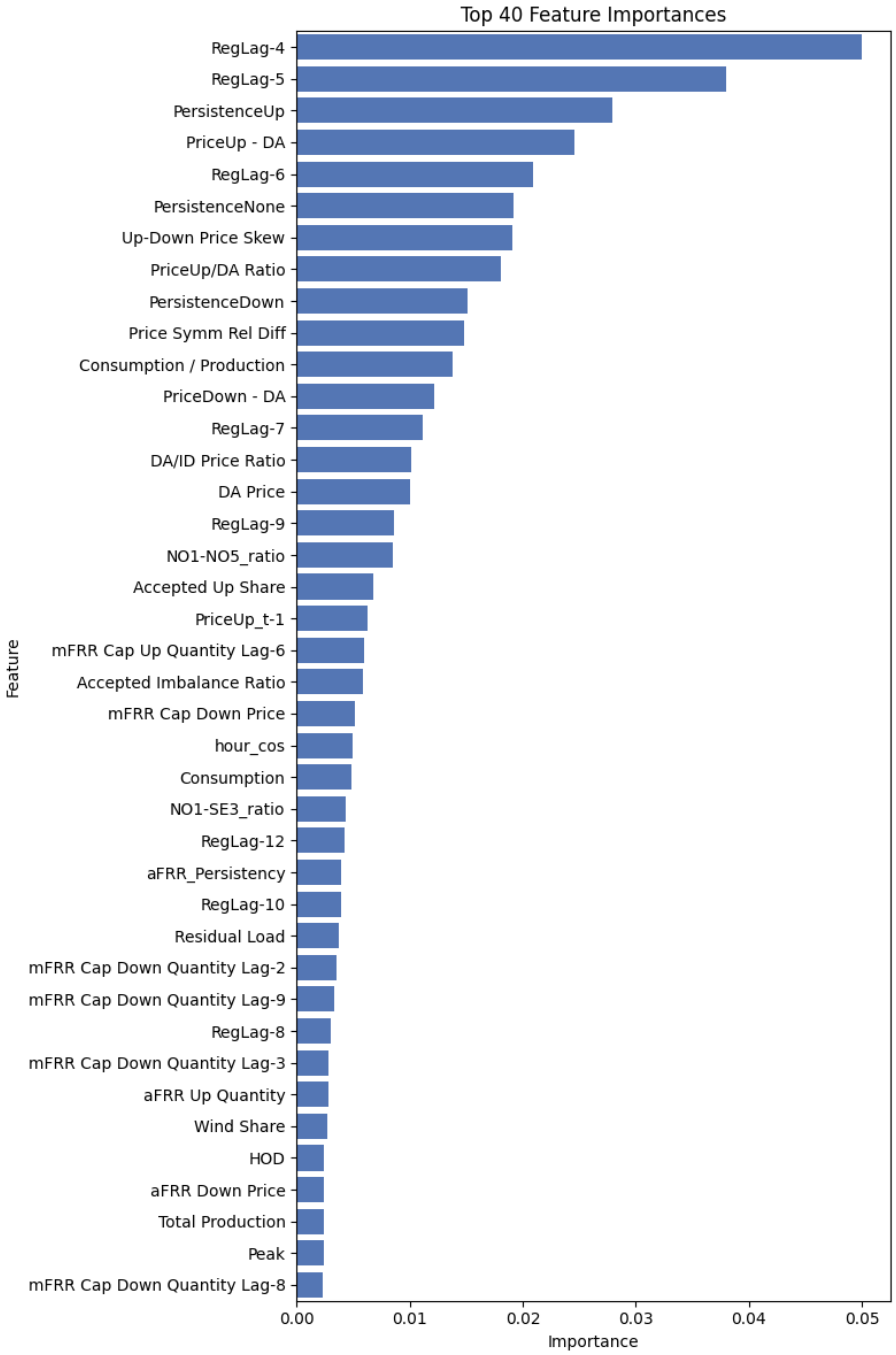


Figure 22: Feature importance for the CatBoost model trained on the post-March 4th 2025 dataset.

5.2.3 Price Difference Distribution

Market participants often consider the relationship between day-ahead prices and regulation prices when making bidding decisions. Figure 23 visualizes and Table 6 summarizes the distribution of the day-ahead to regulation price differences (PriceUp − DA and PriceDown − DA) for correctly predicted up and down-activations, respectively. This visualization helps to understand how much market participants stand to gain from the correctly predicted activation times. There is, of course, the problem of actually being activated, as the activated volume is miniscule compared to the accepted volume. Separate bidding models is required to improve activation likelihood, which is outside the scope of this project.

However, assuming activation, the mean price difference for correctly predicted up-activations is 17.98 EUR/MWh, while for down-activations it is -12.34 EUR/MWh. These price differences represent good opportunities for market participants to profit from their bids. The up-activation price differences are higher in absolute terms than the down-activation differences, but as down-activations are more frequent and predictable, they are the more reliable source of profit. Price differences are likely to vary over time, and as the current dataset is relatively short, these statistics should be interpreted with caution as they may not generalize well to other time periods.

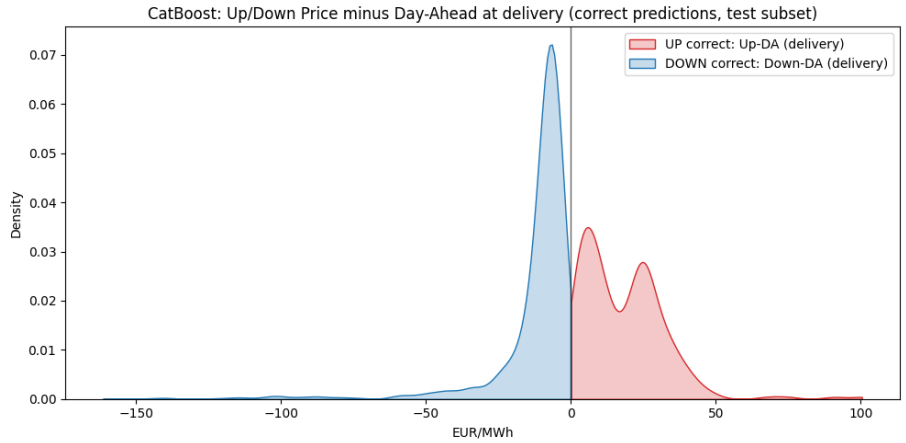


Figure 23: Distribution of Day-Ahead to Regulation Price Differences (PriceUp − DA and PriceDown − DA) across activation classes in the post-March 4th 2025 dataset.

Metric	n	mean	std	p05	p25	p50	p75	p95
Up Price – DA (correct UP, delivery)	275	17.98	14.33	1.81	6.32	15.82	25.82	40.05
Down Price – DA (correct DOWN, delivery)	1343	-12.34	16.78	-39.67	-11.84	-7.66	-5.54	-1.92
UP % change vs DA (correct UP, delivery) [%]	232	25.14	25.85	2.20	7.04	19.80	35.74	62.58

Table 6: Summary statistics for delivery-time spreads, restricted to correctly predicted UP/DOWN cases.

5.2.4 Feature Correlation

Although feature importance analysis indicates contributions from various features, it is important to assess whether these features provide information beyond persistence. It turns out that many features exhibit high correlation with persistence, suggesting that their predictive power may be partially redundant. Figure 24 visualizes an important concept: models trained on only persistence features attribute high importance to few persistence features, while models trained on the full feature set distribute importance more evenly across many features. In practise, this means that even though many features appear important, most of them only convey information that is already captured by persistence features. Thus, features that are thought to only convey redundant information can be pruned without significant loss of information, simplifying the model. Lag featues were initially included up to $t - 20$ (5 hours), indicated by figure 13, but many of these were pruned as the models showed better performance with only the nearest lag features included.

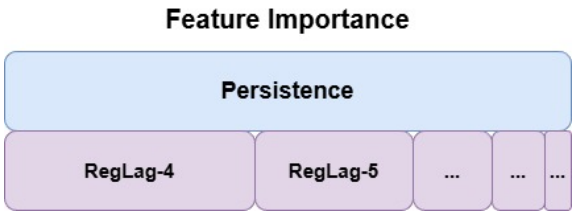


Figure 24: Visualization of feature importance correlation with persistence fatures.

6 Discussion

This section discusses the main findings of the study, their practical implications, and limitations that affect how the results should be interpreted.

6.1 Summary of Findings

Across the evaluated models and feature sets, the results indicate that short-term mFRR activation behaviour exhibits strong persistence. A naive persistence baseline is therefore difficult to beat in macro-averaged classification performance, especially under the constraint that the prediction must be made at bid close (i.e., without access to the most recent activation observations).

At the same time, the feature importance analysis suggests that multiple market and system variables contribute incremental signal beyond pure persistence, particularly in periods where activation patterns are less stable.

6.2 Why Performance Plateaus

Several factors can explain why performance gains over a persistence baseline are limited: (i) class imbalance, where up-activations are rare; (ii) mFRR activations exhibit strong auto-correlation; and (iii) information constraints at bid close, which remove the most informative recent observations.

6.3 Limitations

The study is limited by data availability and resolution transitions during the sample period. Some inputs are only available at hourly resolution and must be resampled. In addition, structural changes in the Nordic market (including the transition to 15-minute MTU) can introduce non-stationarities that reduce the transferability of patterns learned from earlier periods. Lack of data.

6.4 Future Work

Future work can focus on (a) improving minority-class performance through rebalancing and cost-sensitive learning [46], (b) incorporating richer real-time signals (where available) to reduce the information gap at bid close, and (c)

explicitly modelling direction uncertainty as a first-stage classification problem, potentially followed by conditional modelling of expected volumes.

7 Conclusion

This project investigated short-horizon prediction of mFRR activations in NO1 under participant-relevant timing constraints.

The results show that activation dynamics are strongly auto-correlated, making a persistence-based baseline highly competitive. More complex models can add incremental value, but improvements are constrained by class imbalance, and limited information available at bid close.

Overall, the findings suggest that modelling activation direction explicitly is a practical framing for market participants, and that future improvements are most likely to come from better handling of rare up-activation events and from integrating higher-quality near-real-time signals.

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

[10] 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.

[11] 'Balancing market outlook 2030',

[12] *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.

[13] 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.

[14] 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.

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

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

[17] 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.

[18] 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.

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

[20] 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.

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

[22] 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.

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

[24] 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.

[25] *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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

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

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

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

[36] *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.

[37] S. M. Ribeiro and C. L. Castro, 'Missing Data in Time Series: A Review of Imputation Methods and Case Study', *Learning and Nonlinear Models*, vol. 20, no. 1, pp. 31–46, Oct. 2022, ISSN: 16762789. DOI: 10.21528/lnlm-vol20-no1-art3 Accessed: 13th Nov. 2025.

[38] *Intraday Auctions on SIDC*, https://www.entsoe.eu/network_codes/cacm/implementation/ida/. Accessed: 15th Dec. 2025.

[39] *Numerical data: Normalization — Machine Learning*, https://developers.google.com/machine-learning/crash-course/numerical-data/normalization. Accessed: 15th Dec. 2025.

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

[41] *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.

[42] 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.

[43] 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.

[44] *3.3. Tuning the decision threshold for class prediction*, https://scikit-learn/stable/modules/classification_threshold.html. Accessed: 14th Dec. 2025.

[45] 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.

[46] *2. Cost-sensitive learning — Reproducible Machine Learning for Credit Card Fraud detection - Practical handbook*, https://fraud-detection-handbook.github.io/fraud-detection-handbook/Chapter_6_ImbalancedLearning/CostSensitive.html. Accessed: 15th Dec. 2025.