NTNU
Kunnskap for en bedre verden

Department of Electric Energy

TET4510 - Specialization Project

Short-Term mFRR Activation Direction Forecasting

Author:
Haakon Nygård Hellebust

Date 18/12/2025

Abstract

The Nordic power system's transition to automated manual Frequency Restoration Reserve (mFRR) activation at a 15-minute resolution on 4 March 2025 represents a structural shift in balancing market operations and increases the value of short-term forecasting for market participants. While existing literature has largely focused on forecasting activation volumes or prices, often from a transmission system operator perspective, less attention has been paid to predicting mFRR activation direction under the information constraints faced by market participants. For demand-side aggregators of flexible resources, activation direction (up-regulation, down-regulation, or no activation) constitutes the most critical short-term signal, as it determines the likelihood of activation independently of price or volume magnitude.

This study investigates the feasibility of short-term forecasting of mFRR energy activation direction at a 15-minute resolution using only data available to market participants at gate closure. Focusing on the Norwegian NO1 bidding zone, a supervised multi-class classification framework is developed using extensive feature engineering across market prices, cross-zonal flows, production and load forecasts, and temporal persistence indicators. Tree-based machine-learning models are developed and tuned using the open-source AutoML framework AutoGluon, and evaluated against a persistence-based naïve baseline using temporally consistent training, validation, and test splits that respect real-time information availability.

The results demonstrate that the majority of predictive power for mFRR activation direction arises from short-term temporal activation persistence. The findings indicate that explicit direction forecasting is both practically relevant for participant decision-making as correctly predicted activation directions entail meaningful economic value, and methodologically valuable as an intermediate step toward conditional forecasting of activation volumes and prices.

By providing one of the first participant-feasible, 15-minute resolution studies of mFRR activation direction under the new Nordic market design, this work contributes empirical evidence on short-term predictability in automated balancing markets and establishes a foundation for future conditional and probabilistic activation forecasting frameworks.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

The increasing share of weather-dependent renewable energy sources, particularly wind power as illustrated in Figure 1, has fundamentally altered the dynamic behavior of modern power systems. Unlike conventional synchronous generation, these resources contribute little or no rotational inertia, reducing the system's ability to passively resist frequency deviations following disturbances. As a result, frequency deviations now evolve more rapidly and with greater amplitude, making short-term system imbalances both more frequent and more difficult to predict [1]. These challenges are further exacerbated by forecast uncertainty in renewable generation and load, which can lead to rapid and unforeseen deviations from scheduled production and consumption.

To maintain frequency quality and operational security under these conditions, Nordic transmission system operators (TSOs) rely on a hierarchy of balancing resources designed to operate at different time scales. While fast-acting reserves such as FCR and aFRR address immediate frequency deviations, manual Frequency Restoration Reserves (mFRR) play a key role in correcting sustained imbalances and restoring the system to a stable operating point. The mFRR energy activation market (mFRR EAM) enables market participants to offer up- and down-regulation reserves that the TSO can activate when such imbalances persist [2].

Historically, mFRR activations in the Nordic region were manual and scheduled on an hourly basis. Under this structure, activation signals were constrained to hourly blocks, limiting the ability of the balancing mechanism to respond efficiently to short-lived imbalances or rapid changes in system conditions. On March 4th 2025, the Nordic region introduced a new automatic mFRR activation platform operating at a 15-minute resolution [3]. This reform represents a significant shift in balancing market design, aligning the Nordic system with European standards and facilitating closer integration with pan-European balancing initiatives.

The transition to automated 15-minute mFRR activation allows balancing actions to more closely track real-time system dynamics, improving the precision and timeliness of imbalance correction. Automated activation enables faster response times and reduces structural imbalances arising from temporal aggregation [4]. At the same time, the finer temporal resolution fundamentally changes the decision-making environment for market participants. Bidding decisions must now be made over shorter horizons and with increased sensitivity to rapidly evolving system conditions, magnifying the importance of reliable short-term forecasting and situational awareness in the mFRR energy activation market.

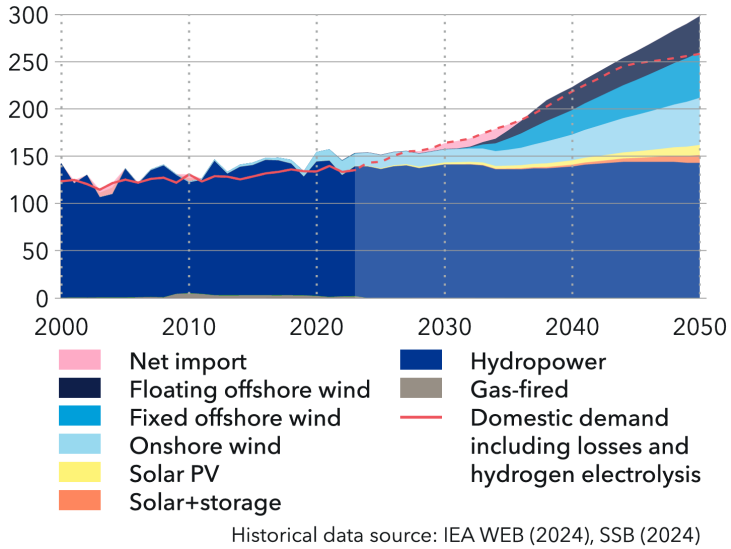


Figure 1: Norway electricity supply by power station and net imports with projections to 2050 [5].

1.2 Motivation

mFRR activations have direct economic implications for market participants. While capacity markets provide stable income, 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. Being able to forecast when no activation is likely to occur is equally important, as capacity market revenues can then be prioritized without risking missed activation opportunities. 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 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 short-term 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.

Most of the literature on balancing and activation forecasting focuses on direct prediction of activation volumes and prices [6, 7, 8, 9, 10]. Activation volumes are relevant to both TSOs and market participants because they quantify the system's balancing need and the available market opportunities. Prices are also important to both groups: TSOs care about the cost of balancing and market efficiency, while participants care because prices determine settlement outcomes and profitability.

For many demand-side aggregators of flexible distributed resources (e.g., EVs, heat pumps, batteries), however, the most critical short-term question is often *not* the exact price level or total regulation volume, but whether the system will be in a up-regulation, down-regulation, or no-activation state. Since the marginal costs for demand-side flexible aggregators are small, and thus they can bid low into the mFRR EAM, they are likely to be activated whenever there is a need for regulation in their direction. In this sense, *activation direction* becomes the primary decision variable for flexible demand-side aggregators, while regulation prices remain important, but secondary variables that refine expected revenues. The importance of this perspective increases as aggregators play a larger role in providing system flexibility under higher renewable penetration [1, 11].

Much of the existing work on mFRR activation volumes take the TSO perspective [7, 10]. Due to the difference in data access and decision contexts between TSOs and market participants, one must be cautious when applying TSO-focused findings to participant-relevant processes. Market participants are constrained to using only information available at gate closure, which may exclude certain real-time system signals accessible to TSOs. This availability concern motivates a participant-centric approach to activation direction forecasting, ensuring that developed models and insights are directly applicable to the decision-making contexts of mFRR EAM market participants.

Activation direction foresight is, in addition to its direct relevance for demand-side aggregators, also a useful intermediate step towards full volume and price forecasting. Activation volume and price targets are noisy and difficult to predict directly. Additionally, because mFRR activation volumes are zero-inflated, regression-based models tend to struggle as it must learn to predict both the occurrence and the magnitude of activations simultaneously. An important insight arises from this: **activation volumes and prices are conditional on activation direction.** If the direction of activation is known in advance, volume and price forecasting can be simplified to a conditional regression problem, where activation volumes and prices are only predicted given that

an activation in a specific direction is likely to occur. The zero-inflation problem is thus avoided, as the activation direction model effectively acts as a filter that separates zero from non-zero activations.

1.3 Research Questions

This project is motivated by the need for improved short-term activation direction foresight among market participants, particularly demand-side aggregators of flexible resources. The transition to an automated 15-minute mFRR market creates both the need and the opportunity for enhanced short-term forecasting capabilities. By focusing on activation direction prediction, this study aims to provide insights that are directly relevant for select market participants, and lays the groundwork for more accurate volume and price forecasting in future work. While these are specific motivations, the methodological insights gained from this study may also have broader applicability to other market participant-relevant processes. The likelihoods of up-regulation, down-regulation, and no activation can, for instance, be useful inputs for grander bidding strategy frameworks that seek to optimize participant revenues in various reserve and balancing markets under uncertainty. Reinforced learning-based bidding strategies could potentially leverage activation direction forecasts as part of their state representation to improve decision-making. The research questions for this study are:

RQ1: To what extent can mFRR energy activation direction (up-regulation, down-regulation, or no activation) be predicted at a 15-minute resolution using only information available to market participants at gate closure under the Nordic automated mFRR market design?

RQ2: Which categories of participant-feasible features—particularly temporal persistence, prices, cross-zonal flows, and production and load forecasts—contribute most to the short-term predictability of mFRR activation direction?

RQ3: Does explicit forecasting of mFRR activation direction provide meaningful value over simple persistence-based heuristics, and can it serve as a robust intermediate step toward conditional forecasting of activation volumes and prices?

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. Finally, Chapter 7 provides concluding remarks summarizing the key findings of the study.

2 Theory

2.1 Electricity Balancing Market

In the Nordics, the electricity balancing markets are part of a joint Transmission System Operator (TSO) project called the Nordic Balancing Model (NBM) [12]. These markets, operated by the respective TSOs (Statnett in Norway), are designed to ensure the stability of the power grid by managing supply and demand imbalances in real-time [13]. These markets facilitate the procurement of balancing services, which are necessary to maintain the equilibrium between electricity supply and demand.

2.1.1 Nordic Balancing Market Structure

In the NBM, the balancing hierarchy consists of several layers, each serving a specific purpose in maintaining grid stability. Figure 2 illustrates the market products and their required activation times. The fastest reserves are at the top of the hierarchy, with slower reserves at the bottom.

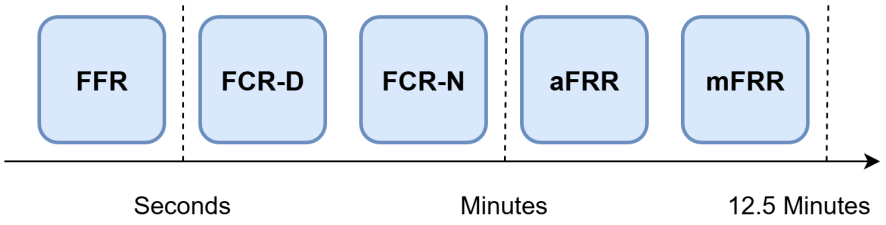


Figure 2: The Nordic balancing market response time requirement hierarchy.

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

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.

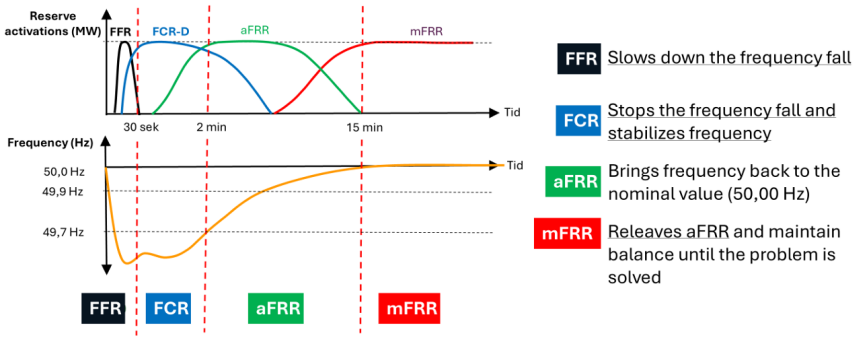


Figure 3: Illustration of the different reserve types and their activation times [2].

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]. Balance Service Providers (BSP), market participants that provide balancing services [15], can offer their capacities to the grid operator through the Nordic aFRR capacity market or in the national 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 [16].

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.

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 it a classification task, where the goal is to assign input data points to one of several predefined classes.

Supervised learning is a machine learning paradigm in which models are trained to learn a mapping from input variables to target outputs based on labeled example pairs. When the target variable takes values from a set of predefined discrete categories, the task is formulated as a classification problem, where the objective is to assign each input instance to one of the several predefined classes. However, standard supervised learning is not applicable in this project, 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 [17].

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 [18]. 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, 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.

2.2.4 AutoGluon

AutoGluon-Tabular is an open-source AutoML framework developed by Amazon [19]. The framework is designed to simplify the process of building and deploying machine learning models for tabular data. AutoGluon-Tabular has the ability to greatly simplify data preprocessing, feature

engineering, model selection, and hyperparameter tuning through automation. It succeeeds by *ensembling* multiple models and *stacking* them in multiple layers. Experiments has shown that such multi-layer combination of many models offers better use of allocated traning time than seeking out a single best model [19].

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. It is thus a pay-as-produced market, where participants are compensated based on the volume of energy they deliver when activated by the TSO [20]. 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 [21]. These market characteristics make it lucrative for participants to predict activation events accurately, as successful predictions can lead to significant financial gains.

In 2022, day-ahead market bidding strategies incorporating balancing market predictions were analyzed in the context of the Nordic market [21]. The gains from incorporating balancing market predictions into day-ahead bidding strategies were found to be near-zero. It was noted, however, that the need for balancing services will increase in the future, and that such strategies will therefore become more relevant and profitable. In a balancing market outlook for 2030, mFRR capacity demands was reported to have increased significantly in recent years and will continue to increase [22]. 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 [23], 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 [24]. 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 [25].

In a recent study examining the effects of increased temporal resolution on balancing activation patterns, it was found that increased resolution significantly reduces structural imbalances and achieves about 60% of the possible reduction in total balancing, compared to a 5-minute resolution ideal [25]. The 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 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 seven families, where the first two families only model activation direction implicitly, and the remaining five address activation direction uncertainty explicitly:

1. Direct imbalance or activation volume forecasting models, which produce point or probabilistic predictions of continuous imbalance or activation volumes;
2. Scenario-based activation models, which simulate discrete sets of possible future imbalance trajectories;
3. Activation-ratio or expected-activation models, which derive expected activation ratios from historical data;
4. Activation-probability and chance-constraint models, which enforce reliability requirements based on probabilistic imbalance representations;
5. Activation-range or interval-uncertainty models, which define bounded sets of feasible activation magnitudes;
6. Markov activation models, which represent activation direction as a stochastic process characterized by transition probabilities; and
7. Regressor-based activation direction classification models, including machine-learning methods that explicitly predict activation direction (upward, downward, or none).

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 and Activation Volume Forecasting

First, it is important to distinguish between *imbalance* and *activation* volumes. Imbalance refers to the net difference between generation and consumption in the power system, while activation volumes refer to the specific amounts of energy that the TSO activates in reserve and balancing markets to correct these imbalances. In this literature review, both imbalance and activation volume forecasting studies are covered, though activation volume studies are more directly relevant for market participants. Imbalance forecasting studies are included because they often address similar challenges and can provide transferable insights.

At the system-operator end of the spectrum, a substantial body of literature focuses on point forecasting of continuous system *imbalance* volumes, with the primary purpose of improving TSO operational decisions. In a 2025 study, this class of work is exemplified through a regression-based model for short-term imbalance forecasting in Belgium [10]. It is argued 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. In a recent study on machine-learning based mFRR forecasting, short-term mFRR *activation* volume forecasting across the four Swedish bidding zones using LSTM models was conducted [26] . The 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 confirm that SE1 and SE2 historically provide the majority of balancing energy on short and medium time scales [27].

From a market-participant perspective, activation volume forecasting has been studied using LSTM models across several Nordic bidding zones [6]. The analysis yields three relevant insights. First, balancing volumes are relatively autocorrelated: past activations contain predictive information about short-term activations. Second, forecast accuracy could likely be improved by incorporating weather-related variables. Third, 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.**

A recent study adopts a market-participant perspective, proposing a two-stage probabilistic framework for sequentially forecasting imbalance volumes and prices in the Greek balancing market [8]. The first stage employs quantile regression to generate probabilistic forecasts of system imbalances, while the second stage uses these quantiles to predict imbalance prices. The findings indicate that system imbalance volumes are critical predictors of imbalance prices, underscoring the strong dependence of prices on system state. By extending from imbalance-volume point forecasts to price forecasting, this line of work illustrates how imbalance forecasts can be translated into more directly actionable price information for balancing-market participants.

Linear machine-learning-based time series models (ARIMAX, SARIMAX) have also been applied to predict signed mFRR activation volumes in Sweden [28]. The study concludes that while 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 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 [29] and stochastic scheduling or bidding frameworks [30].

In 2017, a study modeled uncertain imbalance volumes using three discrete scenarios: high, median, and low, as illustrated in Figure 4 [31]. 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.

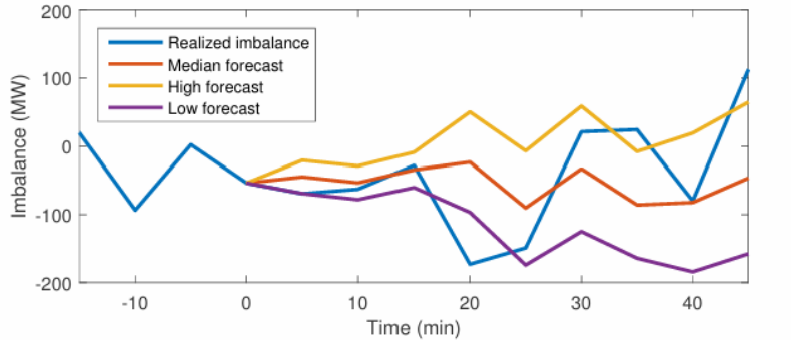


Figure 4: Imbalance forecast scenarios [31].

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.

3.2.3 Activation Uncertainty Modelling Approaches

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. This method was applied in 2023 to analyze and model the Nordic balancing markets [32]. 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.

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. A 2022 study applied chance-constrained optimization to the problem of reserve dimensioning in a multi-area power system [33]. 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 the scenario-based approach [31], only modelled implicitly through the sign of the continuous imbalance scenarios.

From the UK balancing market context, a notable study emerged, developing risk-constrained trading strategies for stochastic generators [34]. 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.

Activation Uncertainty Ranges. In a 2023 study, it was argued that deterministic reserve activation models inaccurately represent activation uncertainty. Thus, a stochastic model was presented, but more interestingly, a robust electric vehicle aggregator scheduling model using uncertain bounded activation ranges was proposed [35]. They use *reserve activation* (RA) as input for activation uncertainty, 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. 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.

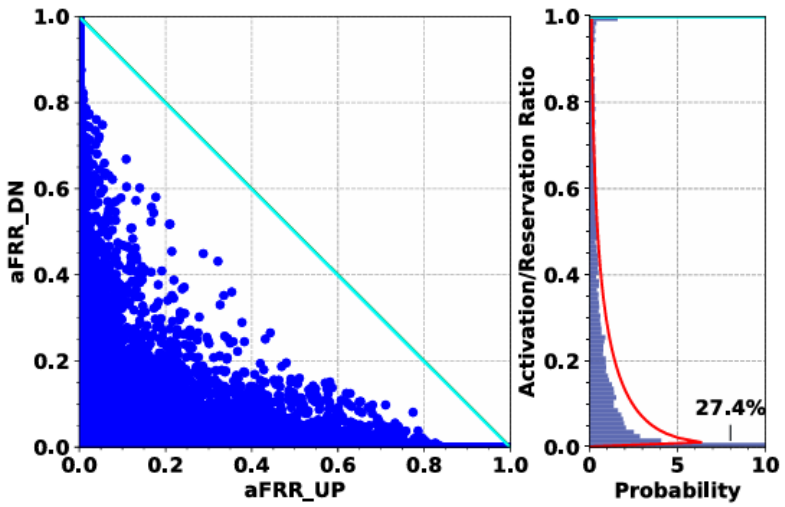


Figure 5: Activation ratio uncertainty ranges for aFRR up[35].

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. The reviewed study operates in the context of FCR and aFRR reserves, where there is always an activated imbalance in one direction [35]. 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.

Markov Activation Models. In 2015, time-series-based forecasting models for electricity balancing markets were benchmarked [36]. In this study, relevant work is separated 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 here Markov models are highlighted as a particularly interesting approach for explicit balance-state modelling. The study refers to 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 [37].

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. A model distinguishing only between activation and no activation, thereby discounting direction, was also tested, [38]. 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 second 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 [39].

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. 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.

In a recent study, the general price formation in intraday and mFRR markets was examined [40]. Among other explorations, logistic regression and ANN (Artificial Neural Network) models were developed 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, the results were promising. 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.

In a slightly different approach to balancing market forecasting, a two-stage model with sequential forecasts of activation direction (XGBoost) and prices at hourly resolution in SE2 was proposed [41]. 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.

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	Year	Target Variable	Resolution	Uncertainty Model
[10]	2025	Imbalance volume	15-min	Point forecast (regression)
[26]	2025	mFRR activation volume	1-hour	ML point forecast (LSTM)
[6]	2023	Balancing volume	1-hour	Probabilistic LSTM; implicit direction
[8]	2025	Imbalance volume and price	1-hour	Probabilistic (quantile regression)
[28]	2024	Signed activation volume	1-hour	Linear ML time-series; implicit direction
[32]	2023	Activation direction and volumes	1-hour	Expected activation ratios; sampled regulation states
[33]	2022	Reserve dimensioning	Scenario-based	Chance constraints on imbalance scenarios
[34]	2018	System length (direction)	1-hour	Probabilistic logistic model with risk constraints
[42]	2023	Reserve activation (RA ratio)	30-min (FCR/a-FRR)	Activation ranges / bounded uncertainty sets
[21]	2015	Balancing state (up/down/none)	1-hour	Markov transition probabilities (duration/hour-specific)
[40]	2025	Activation direction (day-ahead)	1-hour	Classification (logistic, ANN)
[41]	2025	Activation direction and price	1-hour	Classification (XGBoost); sequential predictive model

In summary, the reviewed literature shows steady progress in forecasting imbalances, activation volumes, and (more recently) activation direction, using methods ranging from regression and time-series models to probabilistic frameworks, scenario-based representations, bounded uncertainty sets, Markov state models, and modern classifiers. Across these families, a recurring theme is the difficulty posed by zero-inflation and regime switching between up, down, and no activation—effects that often reduce performance when direction is only handled implicitly. Recent classification-based studies indicate that explicit direction modelling can be feasible and informative, but evidence is still largely based on hourly resolution and, in some cases, on information that may not be available to market participants at decision time.

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 select market participants on its own, and that it provides a natural foundation for subsequent conditional forecasting of continuous activation quantities or other types of analysis and decision-makiing processes.

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 [41], 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 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.

Performance is difficult to evaluate in a vacuum. A *naive* baseline model was therefore developed to provide a reference point for assessing the performance of more sophisticated models. The naive model assumes the target $t + 4$ to equal the most recent known activation direction, i.e. at $t - 4$. This approach captures the persistence often observed in mFRR activations, where future activations tend to correlate with recent activations. The naive model's performance is evaluated using the same metrics as the more complex models, allowing for direct comparison.

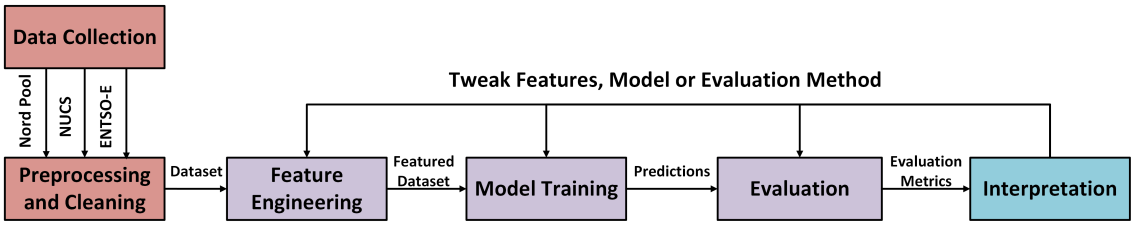


Figure 6: Overview of the methodological approach used in this project.

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 [43]. 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.

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) [44]. 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.

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 [40] 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 [45].

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 [46]. 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 [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.

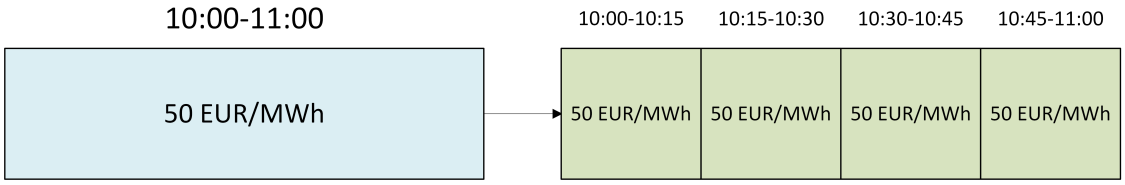


Figure 7: Visualization of resampling from 1-hour to 15-minute resolution using forward filling.

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 [47].

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 (2024-2025). 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 [18]. The predominance of no-activation contextualizes the zero-inflated nature of the imbalance volumes, as discussed in the literature review. The classification formulation adopted in this project makes zero-volume intervals a discrete target class, reducing noise from volume magnitude variations within activation classes.

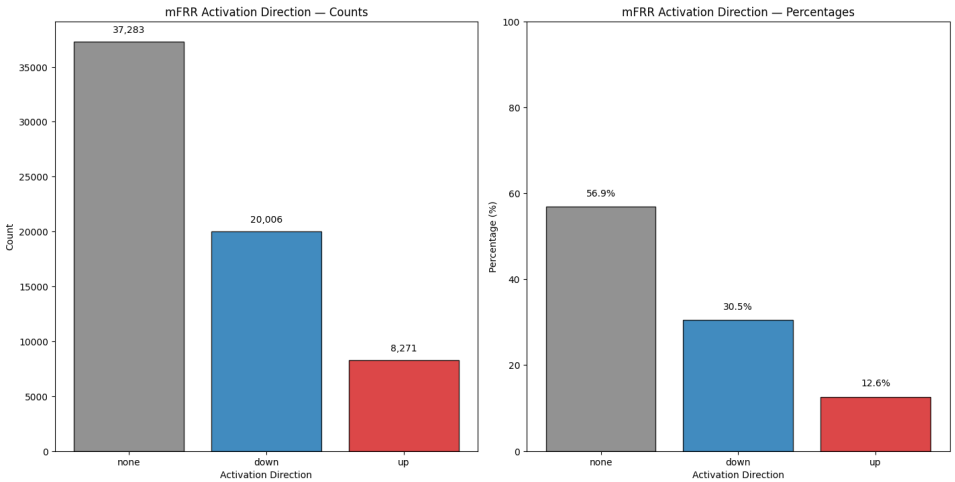


Figure 8: mFRR activation direction distribution.

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.

Cross-zonal Flows. 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.

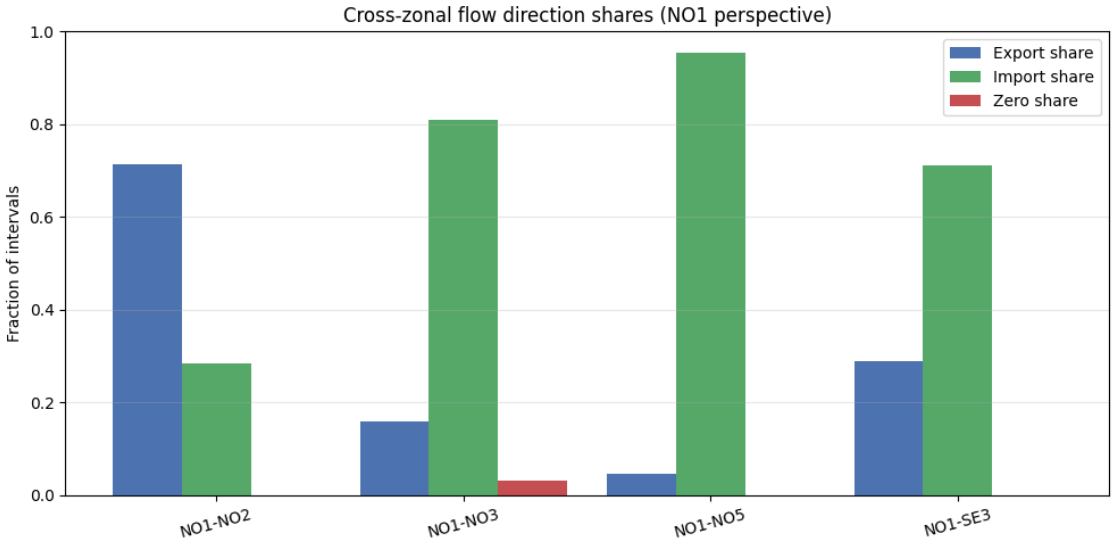
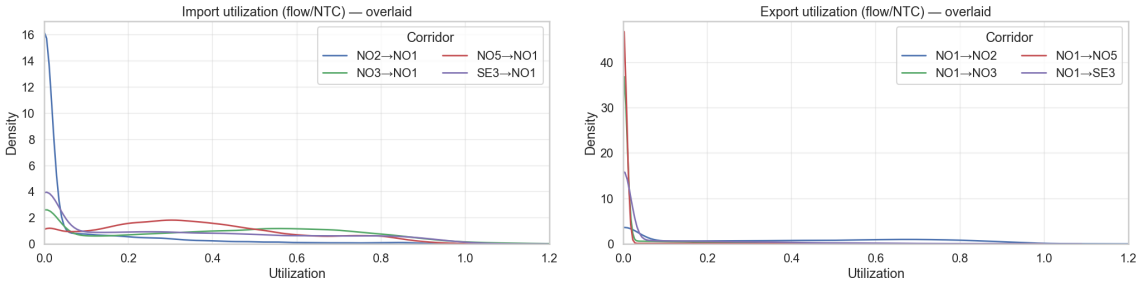


Figure 9: Cross-zonal flow distributions for the NO1 bidding zone.

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. Table 2 summarizes key statistics for NO1 wind-related features in the dataset (from 2024 to 2025). The statistics include mean, standard deviation, minimum, various percentiles (P10, P50, P90), maximum, and the number of observations for each feature. The features encompass day-ahead and intraday wind forecasts, actual wind production, forecast revisions, forecast errors, and the share of wind in total production. The mean wind production revision (from intraday to day-ahead) is 12.59 MW, with a standard deviation of 15.27 MW. When compared with the mean actual production of 120.38 MW, this indicates that wind errors are quite significant. This is a positive sign for the relevance of wind-related features in predicting mFRR activations, as intraday-day-ahead discrepencies can lead to imbalances requiring mFRR activations.

While wind-related variables capture an important weather-driven source of uncertainty, additional meteorological features (e.g., temperature, precipitation, wind speed/gusts, and pressure forecasts) could plausibly improve performance by explaining variation in both demand and renewable production. Such features were not included in this study due to scope and data integration constraints, and are left as a natural extension for future work.

Table 2: Summary statistics for wind-related features (2024–2025, NO1)

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

aFRR. Figure 11 and 12 show a histogram and time series plot of the hourly NO1 aFRR "up" 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. This phenomenon is consistent for both up- and down-regulation prices and volumes.

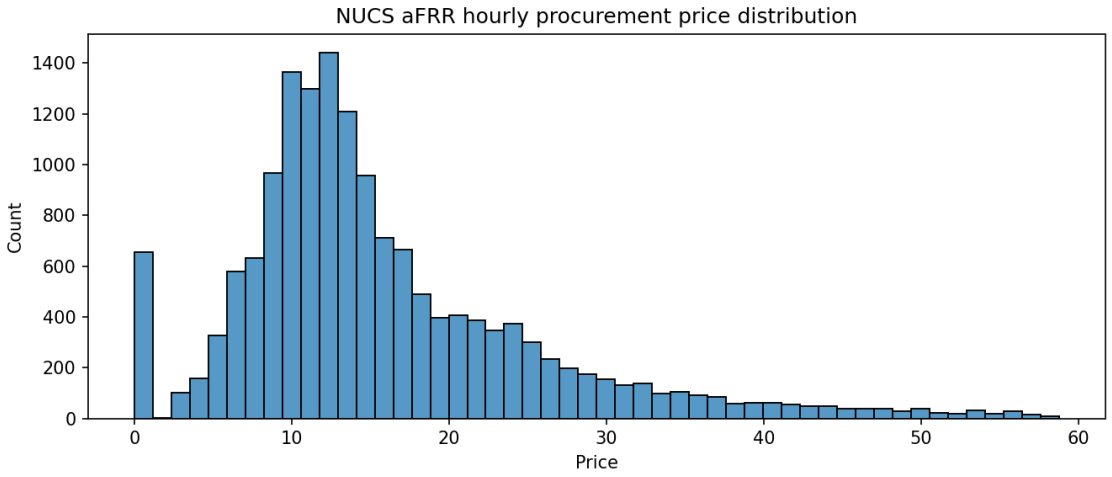


Figure 11: A histogram of hourly aFRR "up" procurement prices for the NO1 bidding zone from NUCS data.

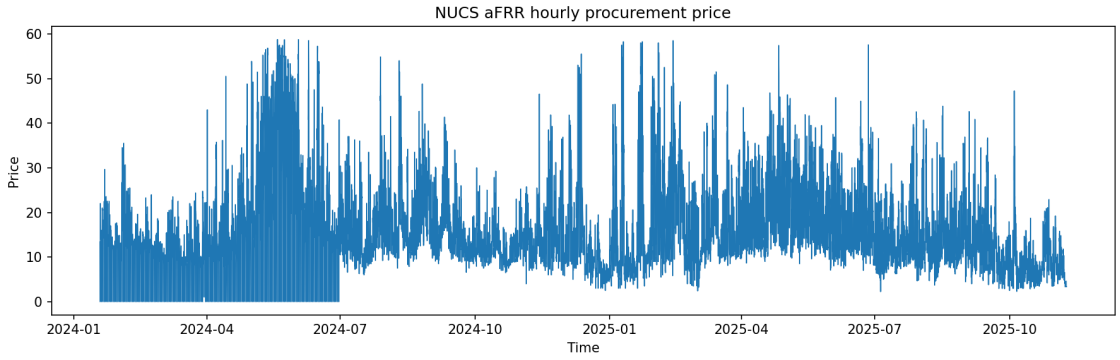


Figure 12: Hourly aFRR "up" procurement prices for the NO1 bidding zone from NUCS data.

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 a target variable, e.g., income. In this project, each data point represents a specific timestamp 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.

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 delivery 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 autocorrelated [6], 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.

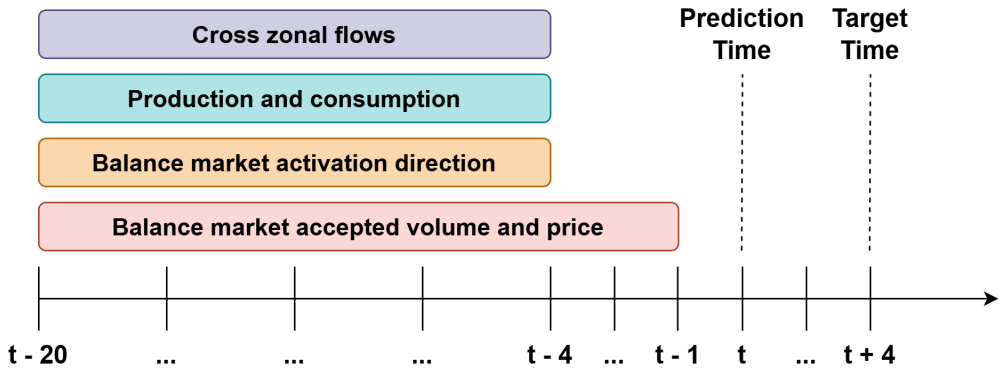


Figure 13: Illustration of time restrictions on feature availability for predicting mFRR activations at time $t + 4$ based on data available at time t.

An additional complication is that some data sources are made available only at 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 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 the 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 gate closure times of various market datasets relative to the delivery day D. Market gate closure may not align perfectly with actual market participant availability and release times, but they provide a good approximation. In this project, these data release 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.

Table 3: Market gate closures and availability times relative to delivery day D [48].

Market / Dataset	Gate Closure Time	Relative to Delivery
mFRR Capacity Market	23:00	$D-1$
Day-Ahead Auction	12:45	$D-1$
Intraday Auction 1 (IDA1)	15:00	$D-1$
Intraday Auction 2 (IDA2)	22:00	$D-1$
Intraday Auction 3 (IDA3)	10:00	D

During the course of this project, it seems that many data sources and their availabilities have changed slightly. Balance market activation data is, for instance, now available for $t-3$ instead of $t-4$. Such changes should be taken into account in future work, but it is unlikely that they will have a major impact on the overall findings of this study.

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 behavior 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 14. 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.

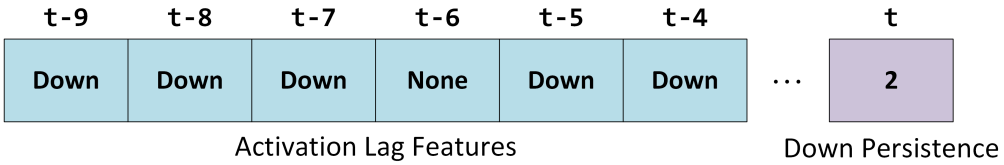


Figure 14: Illustration of down-persistence feature calculation based on lagged activation features. Here, the down persistence at time t is 2, as there have been down-activations in the two most recent intervals ($t-4$ and $t-5$), before an interval with no activation at $t-6$.

Persistence features 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 kinds 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 [49]. 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 understand how mFRR activations vary with time, such as daily or weekly cycles. Basic temporal features include hour of the day, day of the week, and month of the year. These features allow the model to learn patterns related to specific times. mFRR activations could, for instance, be caused by completely different factors during peak hours on weekdays compared to off-peak hours on weekends. Temporal features like these are most often represented using cyclical encoding to reflect their periodic nature. For example, 1 AM and 11 PM are close in time, even though their numerical representations (1 and 23) are far apart. Cyclical encoding uses sine and cosine transformations to capture this periodicity [50]. Hourly features are, for instance, encoded as:

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

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 activation market is closely linked to other electricity markets, such as the day-ahead market, the intraday market, and the aFRR market. While most prices may not directly impact activations, by constructing features that capture relationships between prices, the model may infer system stress levels that could lead to mFRR activations. Large discrepancies 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. Table 4 summarizes the engineered price features and their underlying intuition.

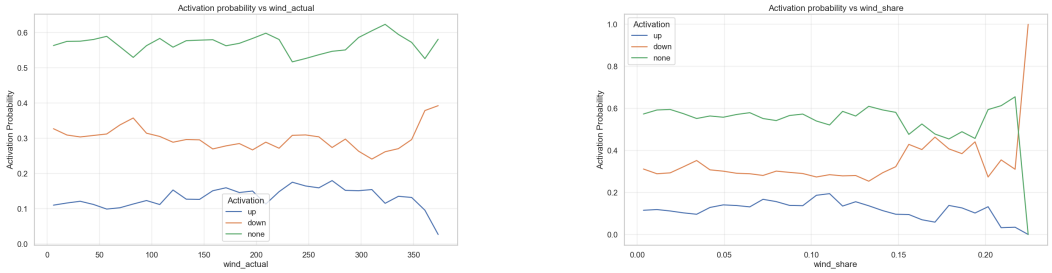
Table 4: Overview of engineered price features and their intuition.

Feature name	**Intuition**
Day-ahead price	Baseline wholesale price level for the MTU.
Intraday prices (IDA1, IDA2, IDA3)	Near-real-time price; reflects late system updates.
aFRR up/down price	Balancing cost proxy; indicates system stress.
Lagged up/down mFRR prices	Recent mFRR price trends; activation cost signal.
Up/down to day-ahead spread and ratio	Relative balancing cost vs. baseline price.
DA–ID spread and ratio	Late deviations vs. schedule; forecast error signal.
DA–ID symmetric relative spread	Magnitude of late price changes; system volatility.

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 15a and 15b 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) Wind share 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.

4.5 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 16). 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.

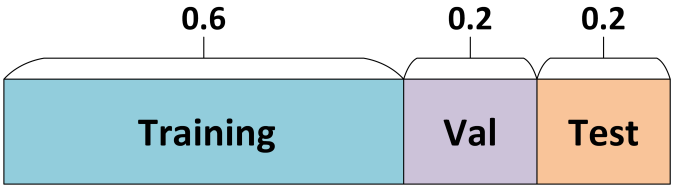


Figure 16: Temporal data splitting into training, validation, and test sets.

Cross-validation could 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 [41]. An analysis of this is presented in the results chapter.

4.6 Evaluation Framework

Classification problems are often evaluated using accuracy, precision, recall, and F1-score. These metrics are defined as follows, based on definitions visualized in Figure 17 [51]:

- **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:
$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

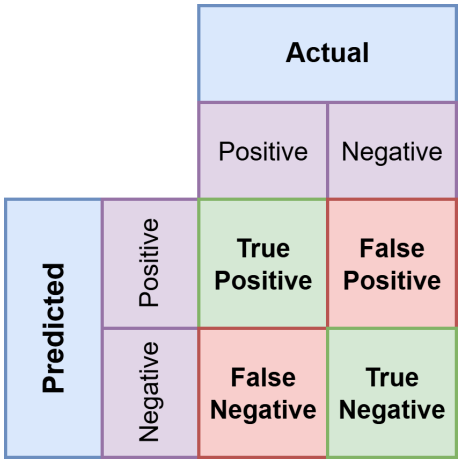


Figure 17: Confusion matrix illustrating true/false positives/negatives.

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 [52].

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 18. 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 18: 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.

4.6.1 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 [53]. Figure 19 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.

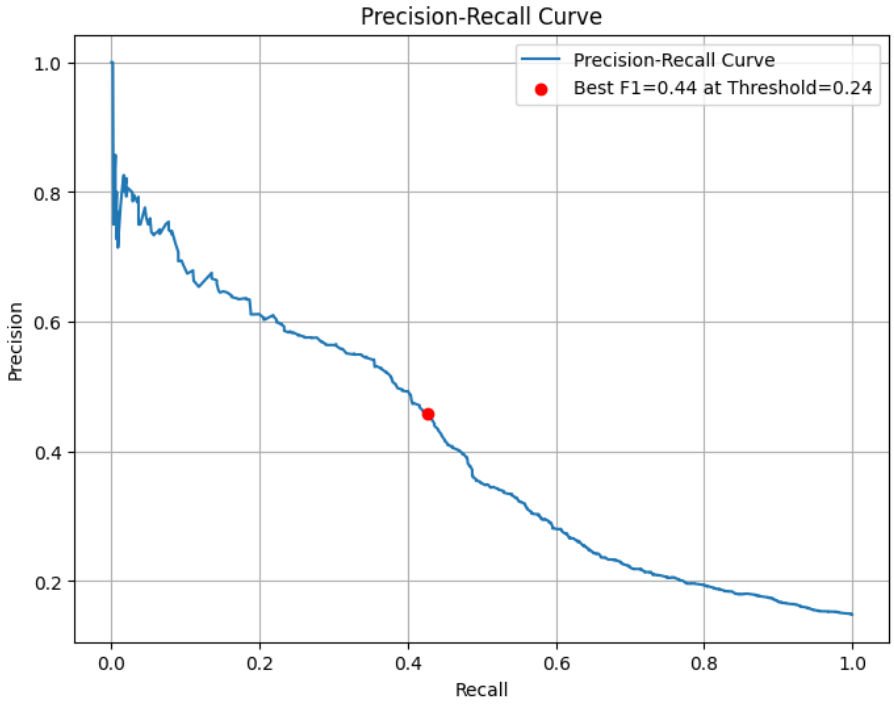


Figure 19: Typical binary Precision-Recall Curve with F1-score maximization point indicated.

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.

4.7 Model Selection

AutoGluon was used as the main machine learning framework for model training and evaluation. When AutoGluon is given a featured dataset, it automatically trains and tunes multiple stacked and ensembled models using different algorithms and hyperparameter settings. It then evaluates the performance of each model on a validation set and selects the best performing model based on a specified metric, such as F1-macro [19]. This automated approach simplified the model selection process for this project, allowing for more efficient experimentation and iteration over different dataset and feature configurations. As such, selection of specific models was not a primary focus, but rather the overall framework and methodology.

5 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. 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 [6]. Much of the motivation for this project is to explore whether more sophisticated models are able to outperform auto-correlation-based baselines by capturing additional relevant information from various features.

Table 5 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 persistence is upheld inbetween intervalsq, but it is a strong auto-correlation indicator. Similar persistence values were observed when evaluated on the entire dataset with 66%, 64%, and 39%, respectively. Down and no-activations are more persistent than up-activations, further increasing the challenge of predicting up-activations. 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 20, 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 5: Classification report for the naive last-observed-class baseline model.

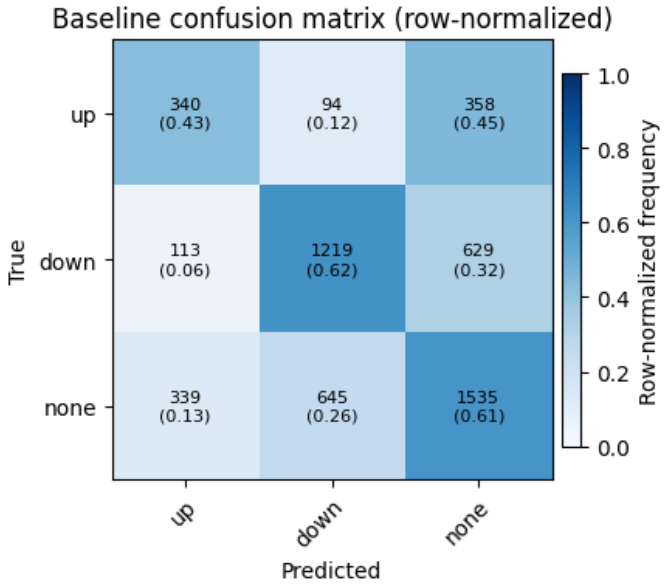


Figure 20: Confusion matrix (row-normalized) for naive model on post-March 4th 2025 dataset.

5.2 Machine Learning Model Results

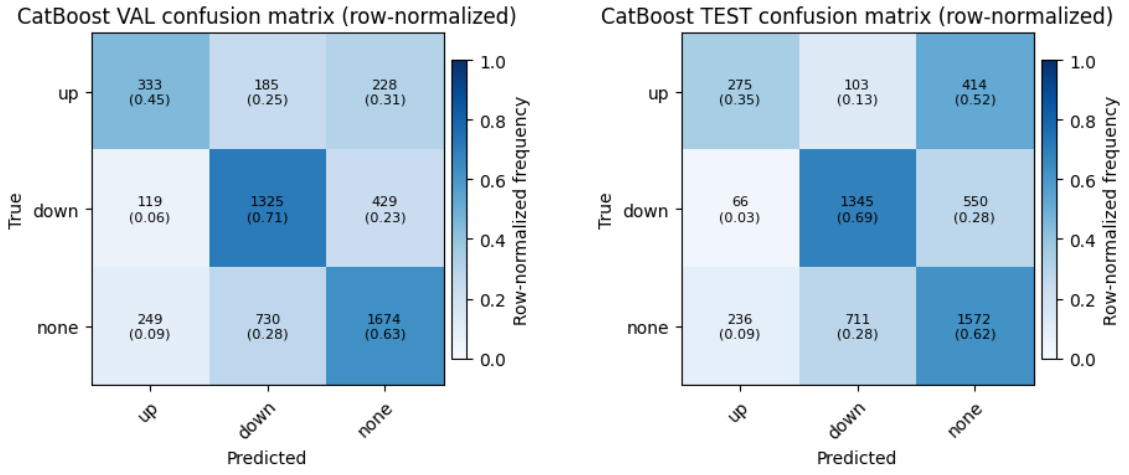
Models were trained using the open-source AutoML framework AutoGluon-Tabular [19]. Much of the appeal of AutoGluon lies in its method of ensembling multiple models and stacking them in multiple layers. Thus, individual models do not need to be considered in isolation, as the ensemble model often outperforms 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 6 and Figure 21 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 6: 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 21: 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 22 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.

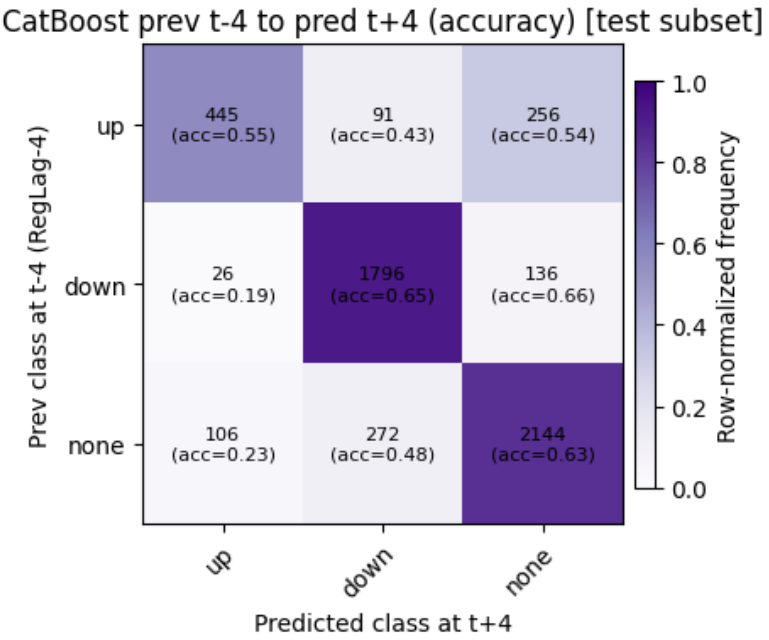


Figure 22: Confusion matrix (row-normalized) for CatBoost model on post-March 4th 2025 dataset.

I have previously made code which indicates sort of how accurate the model is at certain "internal probability thresholds". So, for instance, the model is quite accurate when probabilities for one class are >0.6 etc. This could be interesting to include here? Stand-in for more robust probabilistic modeling?

5.2.2 Classification Threshold Adjustment Impact

The classification threshold adjustment methodology described in Section 4.6.1 made among the most significant improvements to model performance during development. With default thresholds, most models avoided predicting up-activations to a major extent. By up-scaling up-activation probabilities post-training, the model is encouraged to put more weight on up-activation indicators. The thresholds were tuned on the validation set, configured to optimize F1-macro score. Figure 21 and Table 6 show results for the CatBoost model with adjusted thresholds. Pre-adjustment models consistently showed very low recall for up-activations (often <0.2), lowering F1-macro scores significantly. Threshold adjustments, though tuned on a particular validation set, generalized well to the test set, improving up-activation recall notably without sacrificing too much precision.

5.2.3 Feature Importance

Figure 23 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 is reasonable, 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.

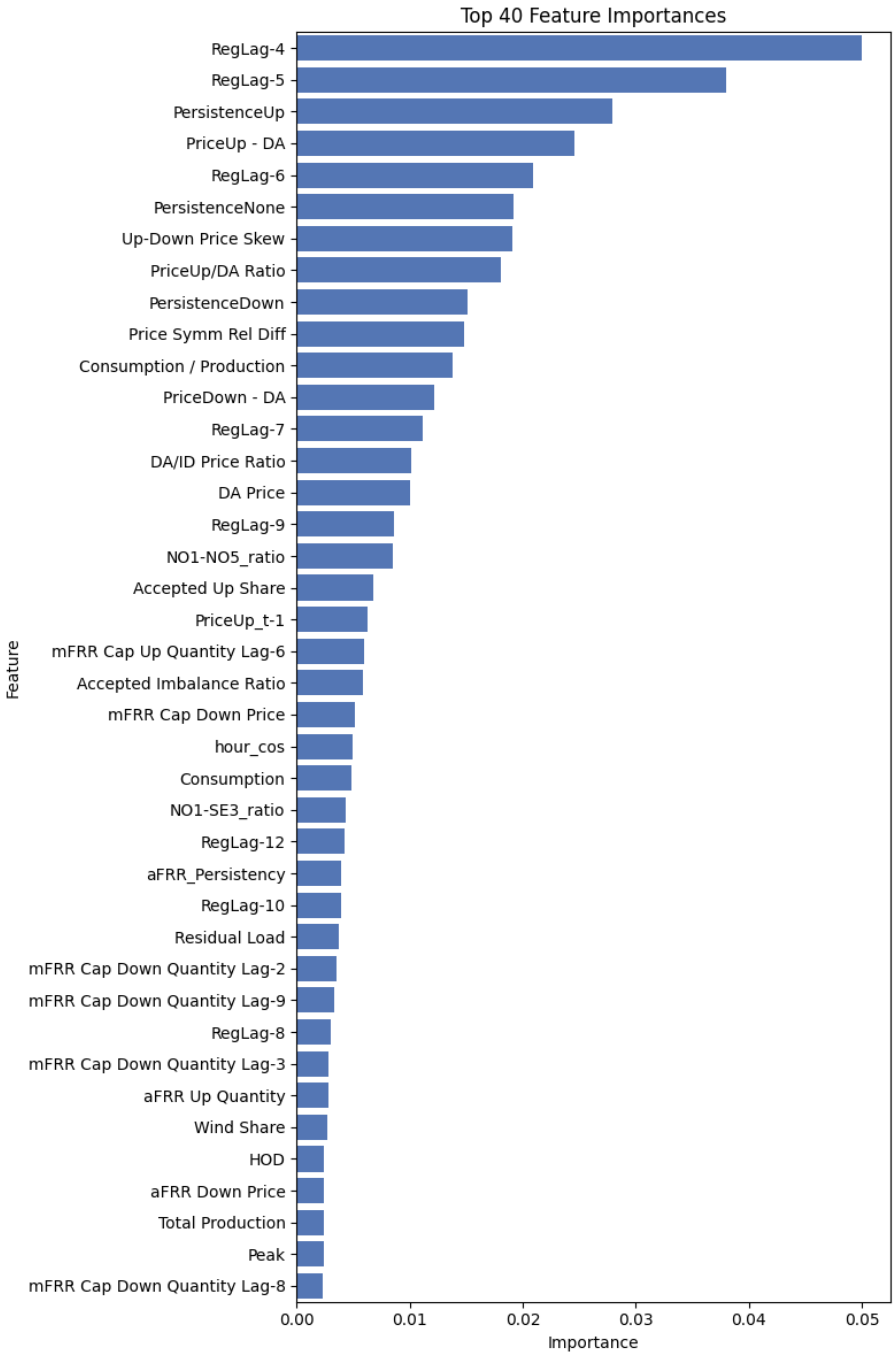


Figure 23: Feature importance for the CatBoost model trained on the post-March 4th 2025 dataset.

5.2.4 Price Difference Distribution

Market participants often consider the relationship between day-ahead prices and regulation prices when making bidding decisions. Figure 24 visualizes and Table 7 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.

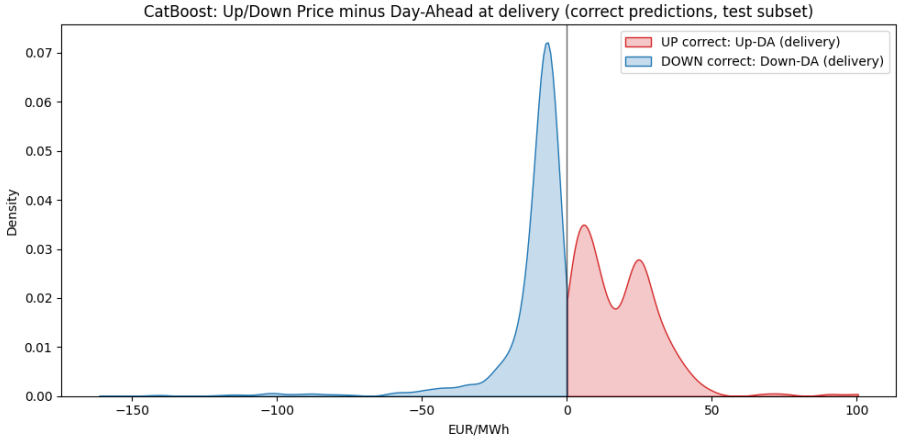


Figure 24: Distribution of Day-Ahead to Regulation Price Differences (PriceUp − DA and PriceDown − DA) across activation classes in the post-March 4th 2025 dataset.

Metric	n	mean	std	p05	p25	p50	p75	p95
Up Price – DA (correct UP)	275	17.98	14.33	1.81	6.32	15.82	25.82	40.05
Down Price – DA (correct DOWN)	1343	-12.34	16.78	-39.67	-11.84	-7.66	-5.54	-1.92

Table 7: Summary statistics for delivery-time spreads (CatBoost), restricted to correctly predicted UP/DOWN cases on test set.

Metric	n	mean	std	p05	p25	p50	p75	p95
Up Price – DA (correct UP)	340	15.83	13.50	1.21	6.50	13.82	22.80	37.4
Down Price – DA (correct DOWN)	1219	-11.34	15.50	-38.03	-12.50	-7.21	-4.50	-0.60

Table 8: Summary statistics for delivery-time spreads (Naive Model), restricted to correctly predicted UP/DOWN cases on test set.

Comparison with Naive Model. To assess whether the machine learning model captures more profitable activation opportunities than the naive model, the same price difference analysis was performed for the naive model's correct predictions. Table 8, when compared with Table 7 on the test set, shows that the machine learning model captures more profitable up-activations (mean 17.98 EUR/MWh vs 15.83 EUR/MWh) and down-activations (mean -12.34 EUR/MWh vs -11.34 EUR/MWh) than the naive model. This indicates that the machine learning model is able to identify more lucrative activation opportunities beyond what is captured by simple persistence patterns.

5.2.5 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 25 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. This relationship is quite obvious for activation lag and persistence features, as they are directly derived from past activation states. However, correlation may also be prominent among the remaining features. The information carried by wind production- and forecast-related features, for instance, may already be captured by persistence patterns. The same applies to cross-zonal flows, consumption, and price features.

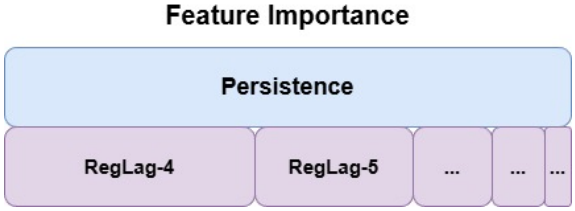


Figure 25: Visualization of feature importance correlation with persistence fatures.

Pairwise Feature Correlation. To quantify feature redundancy, pairwise Pearson correlation coefficients [54] were computed for all features in the final dataset. Table 9 lists feature pairs with absolute correlation coefficients greater than or equal to 0.60 on an already pruned feature set. Some feature pairs, such as consumption and residual load, and raw wind production and wind share, exhibit extremely high correlation (≥ 0.98), indicating near-redundancy. Other pairs, such as various lagged regulation features, also show high correlation (≥ 0.66), but not so high as to be redundant. Some less obvious correlations are also present, such as between day-ahead price and import features, confirming correlations across different system state indicators.

Principal Component Analysis (PCA). PCA is used in machine learning to reduce the dimensionality of datasets while preserving as much variance as possible (cite) by transforming the original features into a new set of uncorrelated variables called principal components. In this project, PCA is not applied as such, but rather as an analytic tool to assess feature redundancy. Figure 26 shows a PCA projection of the post-March 4th 2025 dataset. According to PCA, 90% of the variance in the dataset can be explained by 23 principal components. Given that the dataset contains 93 features, this indicates a high degree of redundancy among the features.

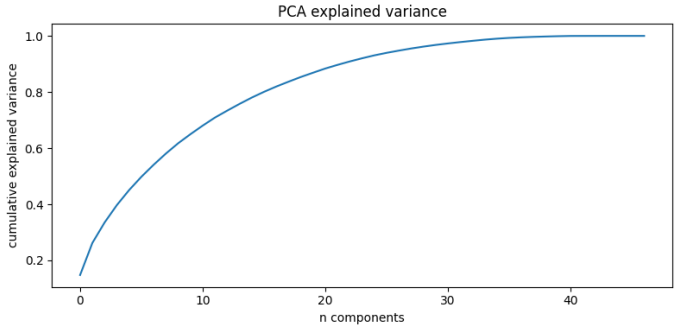


Figure 26: PCA projection of the post-March 4th 2025 dataset.

Table 9: Top correlated feature pairs (absolute Pearson correlation ≥ 0.60).

Feature A	Feature B	$\lvert\rho\rvert$
Consumption	Residual Load	0.997733
Accepted Imbalance Ratio	Accepted Up Share	0.994552
Wind Production	Wind Share	0.989141
Consumption / Production	Residual Load	0.957463
Consumption	Consumption / Production	0.956545
wind_error_t+2	wind_error_t+4	0.947148
PriceDown_t-1	PriceUp_t-1	0.888687
month_cos	month_sin	0.884542
DA Price	PriceDown_t-1	0.850790
DA/ID Price Ratio	Price Symm Rel Diff	0.841696
DA Price	PriceUp_t-1	0.837728
Consumption	Import/Consumption	0.786014
Import/Consumption	Residual Load	0.782683
Consumption / Production	Import/Consumption	0.782593
HOD	hour_sin	0.775928
Net Import	PriceDown_t-1	0.699544
PriceDown_t-1	Total Imports	0.672832
Net Import	PriceUp_t-1	0.661550
RegLag-4	RegLag-6	0.660286
RegLag-6	RegLag-8	0.660251
RegLag-10	RegLag-8	0.660216
RegLag-10	RegLag-12	0.660180
Net Import	Total Imports	0.650611
DA Price	Net Import	0.643849
ID Price 3	ID3 Mom_1h	0.643338
PriceDown - DA	PriceUp - DA	0.623193
Net Import	Residual Load	0.608814
DA Price	Total Imports	0.606887
Consumption	Net Import	0.605301
PriceUp_t-1	Total Imports	0.600970

Code Availability

The entire codebase for data processing, feature engineering, model training, and evaluation is available in a public GitHub repository: https://github.com/haakonnh/mfrr_activation_classify. The repository is segmented into various modules, each corresponding to different stages of the project workflow. Comprehensive documentation is provided within the repository, including a README file that outlines the project structure, dependencies, and instructions for reproducing the results presented in this report.

6 Discussion

6.1 Summary of Findings

This study demonstrates that short-term mFRR activation direction at a 15-minute resolution is predictable to a limited but meaningful extent using only participant-feasible information available at gate closure. Across all evaluated models, short-term predictability is dominated by temporal persistence in recent activation states, with a simple persistence-based baseline proving difficult to outperform in macro-averaged classification performance.

Clear differences in predictability are observed across activation classes. Down-activations are the most consistently predictable class: on the test set, 69% of all down-activations, which constitute 37.2% of all activations, are correctly identified by the best-performing model, with 62% precision (Table 21). Up-activations are comparatively less predictable, reflecting their more sporadic and less persistent nature.

Other at-gate-closure features, including prices, cross-zonal flows, and production and load forecasts, provide only marginal incremental value once persistence is accounted for. Overall, the results indicate that activation direction forecasting under strict information constraints is most effective when treated as a short-horizon, persistence-aware classification problem, and that class-conditional performance differences are critical for assessing practical usefulness.

Finally, when comparing economic relevance via the price difference analysis, the machine learning model captures more profitable activations than the naive model among its correct predictions. On the test set, Table 8 compared with Table 7 shows higher mean price spreads for both up-activations (17.98 EUR/MWh vs 15.83 EUR/MWh) and down-activations (-12.34 EUR/MWh vs -11.34 EUR/MWh), suggesting that the model identifies more lucrative opportunities beyond what is explained by simple persistence patterns.

6.2 Answers to Research Questions

RQ1: To what extent can mFRR energy activation direction be predicted at a 15-minute resolution using only information available to market participants at gate closure?

The results indicate that mFRR activation direction is predictable to a limited but non-negligible extent at gate closure when evaluated at a 15-minute resolution. Across all evaluated models, short-term predictability is primarily driven by temporal persistence in recent activation states. Models are able to correctly classify a substantial share of down-activation events, with the best-performing model correctly predicting 69% of all down-activations in the test set with 62% precision. Up-activations are more difficult to predict, with lower recall and precision.

RQ2: Which categories of participant-feasible features contribute most to the short-term predictability of mFRR activation direction?

Feature importance analyses show that short-term temporal persistence features dominate the predictive signal. Recent activation states consistently emerge as the most influential predictors across the models, reflecting strong auto-correlation in mFRR activation behavior. Other feature categories, including prices, cross-zonal flows, and production and load forecasts, provide only marginal incremental information once persistence is accounted for. Feature correlation analyses further suggest that many of these additional features correlate with recent activation states, limiting their unique predictive value. PCA results corroborate these findings, indicating that most variance in the feature set can be captured by a significantly reduced number of components.

RQ3: Does explicit forecasting of mFRR activation direction provide value over simple persistence-based baselines, and can it serve as an intermediate step toward conditional forecasting of volumes and prices?

Explicit activation direction forecasting provides limited but potentially meaningful incremental value over a pure persistence baseline, particularly in class-specific performance and decision-relevant contexts. While overall macro performance gains are modest, the models demonstrate improved discrimination for certain activation states, most notably down-activations, relative to the naive baseline. Combined with the observed price differences conditional on correctly predicted activation direction, this suggests that direction forecasting can support participant decision-making in settings where correct directional signals entail economic value. Furthermore, treating activation direction as a separate classification problem provides a conceptually and methodologically useful intermediate layer for future conditional forecasting of activation volumes and prices, where direction uncertainty otherwise complicates direct regression approaches.

6.3 Implications

A central implication of this study is that **mFRR down-activation direction exhibits relatively strong short-term predictability under participant-feasible information constraints in NO1**. The evaluated models achieve materially higher recall and precision for down-activations than for up-activations, indicating that down-regulation events tend to persist over short horizons and are therefore more amenable to anticipation at bid close.

These findings are particularly relevant for demand-side aggregators of flexible resources, for whom short-term activation direction constitutes a key operational signal. For such participants, the results suggest that forecast-driven decision support is most reliable when used to **prioritize down-regulation-oriented mFRR strategies**. Directional forecasts can serve as a filtering mechanism that helps identify intervals with elevated likelihood of down-activation, supporting decisions on whether flexibility should be reserved for mFRR participation or allocated to alternative uses.

While correct up-activation predictions are associated with larger deviations from day-ahead prices and therefore potentially higher value per event, the lower predictability of up-activations makes such opportunities more difficult to exploit consistently under gate-closure information constraints. As a result, down-regulation strategies are likely to offer more robust and repeatable decision support for demand-side aggregators in practice.

More broadly, the strong role of persistence implies that simple heuristics remain competitive, and any activation forecasting model should be benchmarked against a persistence-based baseline. The limited incremental value of other at-gate-closure features further suggests that performance improvements are more likely to arise from richer temporal modelling of activation regimes or access to better-aligned information closer to real time, rather than from incremental expansion of coarse gate-closure feature sets.

6.4 Limitations

Several factors can explain why performance gains over a persistence baseline are limited: (i) class imbalance; (ii) mFRR activations exhibit strong auto-correlation; (iii) information constraints at gate closure, which remove the most informative recent observations; (iv) varying data availability and granularity; and (v) limited data span and representativeness.

Class imbalance was addressed through decision threshold tuning, but more sophisticated rebalancing techniques could have been explored. It seems, however, that the "up" class is inherently more difficult to predict due to it being more sporadic and less persistent than the other classes.

The strong auto-correlation in activation patterns, especially for down- and no-activations, means that recent activation states are highly predictive of future states. Other more nuanced features may provide only marginal, and potentially noisy, signals that a model may not value highly. Persistence features may therefore have overshadowed other potentially useful features. Attempts were made to exclude persistence features to force models to learn from other signals, but this led to substantially worse performance. This suggests that other features do not contain sufficient predictive information beyond what is already captured by persistence.

The information constraints at gate closure are a significant limitation. This study deliberately restricts inputs to publicly available information, while actual market participants may have access to additional and more recent proprietary signals, which could improve short-horizon predictability beyond what is achievable here. Activation directions are predicted an hour in advance, thus no real-time signals are guaranteed to represent system conditions at the time of activation. This creates an information gap that models cannot bridge. Recent activations are the largest exceptions as they often persist across the gap. In addition, data publication delays can change over time: during early stages of this project, Nord Pool data appeared to be released with a delay of four 15-minute time slots, whereas it now seems to be published with a delay of two time slots. Such changes affect what is realistically "participant-feasible" at gate closure and can influence both model performance and the comparability of results across periods. Varying data granularity also poses challenges, as some features are only available at hourly resolution and must be resampled. This removes important intra-hour dynamics that could aid prediction. Some data streams have the additional limitation of being released at specific times during the day, limiting availability dependent on time of day (see Table 3).

It is difficult to know whether or not the dataset used in this study is representative enough to capture a sufficient variety of patterns. The final sample period covers 10 months at 15-minute resolution, amounting to 28 800 data points. This is not a small dataset per se, but as the dataset consists of under a year of data, seasonal patterns may not be fully captured. Features that indicate seasonality or patterns that may repeat annually are not possible to learn from this dataset. Datasets on such short timeframes also risk being poorly generalizable to future periods. A similarly featured dataset trained over multiple years may yield better results. It is also important to note that this study only considers the price area NO1. Other areas may have different characteristics that could affect predictability, such as less imbalance, less persistence, or differing system dynamics.

6.5 Future Work

Section 5.2.2 showed that adjusting classification thresholds post-training can improve minority-class and macro performance. Future work could explore other class imbalance techniques, such as **integrating cost-sensitive learning** directly into model training to better handle class imbalance [55]. There is also room for improved upsampling or downsampling techniques to rebalance the training data.

Although this study incorporated a diverse set of features available at gate closure and machine learning-based models, future research could explore: (a) more nuanced weather-related features to better capture wind power production deviations and intermittency effects; (b) a more thorough investigation of the predictability of non-persistence-driven features for mFRR activations; and (c) alternative model architectures, such as recurrent neural networks or transformers, that may better capture temporal dependencies.

This project focused on predicting discrete mFRR activation direction at a 15-minute resolution. Future work could extend this to a **probabilistic classification framework**, providing likelihood estimates for each activation direction. Such frameworks are perhaps more difficult to evaluate, but could provide more useful information for market participants. Additionally, more comprehensive bidding strategy architectures, such as reinforcement learning-based agents, could utilize probabilistic distributions to inform optimal decisions under uncertainty.

7 Conclusion

This study has examined the feasibility and relevance of short-term forecasting of manual Frequency Restoration Reserve (mFRR) energy activation direction at a 15-minute resolution following the Nordic power system's transition to automated mFRR activation on 4 March 2025. Motivated by the structural change in balancing market operations and the resulting increase in temporal granularity, the analysis adopts a participant-centric perspective, focusing on the information constraints faced by market participants at gate closure.

Using data from the Norwegian NO1 bidding zone, a supervised multi-class classification framework was developed to predict mFRR activation direction—up-regulation, down-regulation, or no activation—based solely on participant-feasible information. Extensive feature engineering was performed across market prices, cross-zonal flows, production and load forecasts, and temporal activation indicators. Tree-based machine-learning models were trained and evaluated using temporally consistent training, validation, and test splits, and benchmarked against a persistence-based naïve baseline to ensure realistic performance assessment.

The results demonstrate that short-term mFRR activation direction is predictable to a meaningful extent at a 15-minute horizon under realistic participant information constraints. A key finding is that the majority of predictive power arises from short-term temporal persistence in activation patterns, indicating that recent activation states provide strong signals about near-term balancing behavior. While additional system and market features contribute incremental information, their influence is secondary relative to persistence-based indicators.

The findings further indicate that explicit forecasting of activation direction is both practically and methodologically valuable. From a practical standpoint, correctly anticipating activation direction constitutes a critical short-term signal for market participants—particularly demand-side aggregators of flexible resources—where correct directional predictions can yield meaningful economic value independently of precise volume or price forecasts. From a methodological perspective, activation direction forecasting serves as a useful intermediate step toward conditional forecasting of activation volumes and prices, mitigating challenges associated with zero-inflation and mixed activation regimes.

By providing one of the first empirical studies of participant-feasible mFRR activation direction forecasting at 15-minute resolution under the new Nordic market design, this work contributes evidence on short-term predictability in automated balancing markets. The results establish a foundation for future work on conditional, probabilistic, and decision-integrated forecasting frameworks aimed at supporting data-driven participation in evolving balancing markets.

Appendix

AI Declaration

In this project, AI tools were utilized in three primary ways. Firstly, AI-assisted coding tools, more specifcially GitHub Copilot, were used to aid in the programming process. These tools provided suggestions through autocomplete features, which were then reviewed and modified. All code generated by AI tools was thoroughly evaluated and adjusted to ensure correctness.

Secondly, AI language models, specifically ChatGPT, were employed to help refine the language and structure of certain sections of the text. In an increasingly expanding report, AI tools were used to evaluate and suggest improvements to the structure and clarity of the writing. Spelling and grammar checks were also performed using AI tools to enhance the overall quality of the text.

Finally, AI tools were used to assist in the search of relevant literature. Google Scholar Labs, an AI-powered literature search tool, was used to explore relevant sources for the literature review and other sections of the report. This tool was simply used for the search process, as subsequent evaluation of the relevance and quality of the sources was made without AI assistance.

An important consideration when using AI tools in project work is the potential of over-reliance on these tools, which can lead to a lack of critical thinking and also possible derailment from the original intent of the work. To mitigate this, it is important to avoid viewing AI tools as factual sources, and instead treat them as suggestive tools that require oversight. This was the approach taken in this project.

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, *Nordic Balancing Philosophy ENTSOE*, 2024. Accessed: 13th Nov. 2025.

[3] Statnett, *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] 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.

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

[8] 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.

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

[10] 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.

[11] A. Soares and C. H. Antunes, 'The role of aggregators in energy transition', in *Research Handbook on Energy Management*, Edward Elgar Publishing, Oct. 2025, ch. Research Handbook on Energy Management, pp. 370–400, ISBN: 978-1-80037-650-2. Accessed: 16th Dec. 2025.

[12] *Nordicbalancingmodel*, https://nordicbalancingmodel.net/. Accessed: 16th Dec. 2025.

[13] *Wholesale market: Timeframes - NVE*, https://www.nve.no/norwegian-energy-regulatory-authority/wholesale-market/wholesale-market-timeframes/. Accessed: 16th Dec. 2025.

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

[15] *Balancing Service Provider (BSP)*, https://www.svk.se/en/stakeholders-portal/electricity-market/balancing-service-provider-bsp/, May 2024. Accessed: 16th Dec. 2025.

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

[17] 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.

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

[19] 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.

[20] eSETT, 'Nordic Imbalance Settlement Handbook', Sep. 2024.

[21] 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.

[22] Svenska Kraftnät, *Balancing market outlook 2030*, https://www.svk.se/en/stakeholders-portal/electricity-market/provision-of-ancillary-services/balancing-market-outlook-2030/, Dec. 2024. Accessed: 18th Dec. 2025.

[23] *Manually Activated Reserves Initiative*, https://www.entsoe.eu/network_codes/eb/mari/. Accessed: 17th Dec. 2025.

[24] *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.

[25] 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.

[26] C. Edling and D. Azarang, *Machine Learning-Based Prediction and Key Drivers of mFRR Activations : A Swedish Balancing Market Study*. 2025. Accessed: 18th Dec. 2025.

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

[28] J. Bankefors, *Day-Ahead Modelling of the Electricity Balancing Market : A Study of Linear Machine Learning Models Used for Modelling Predictions of mFRR Volumes*. 2024. Accessed: 18th Dec. 2025.

[29] 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.

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

[31] 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.

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

[33] 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.

[34] *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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

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

[42] I. Pavić, H. Pandžić and T. Capuder, 'Tight Robust Formulation for Uncertain Reserve Activation of an Electric Vehicle Aggregator', in *2021 IEEE Madrid PowerTech*, Jun. 2021, pp. 1–6. DOI: 10.1109/PowerTech46648.2021.9495038 Accessed: 9th Dec. 2025.

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

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

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

[46] *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.

[47] 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.

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

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

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

[51] *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.

[52] 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.

[53] *3.3. Tuning the decision threshold for class prediction*, https://scikit-learn/stable/modules/classification_threshold.html. Accessed: 14th Dec. 2025.

[54] *Correlation Coefficient: Simple Definition, Formula, Easy Steps*. Accessed: 15th Dec. 2025.

[55] *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.