

Delft University of Technology

Early warning model for anticipating extreme electricity prices

A comparison of early warning models for extreme prices in the Dutch day-ahead electricity market, using open data, cross-border information and weather conditions

S. Weites

Complex Systems Engineering & Management (CoSEM)
Student number 4914279
Date June 18, 2026

TU Delft

Early warning model for anticipating extreme electricity prices

A comparison of early warning models for extreme prices in the Dutch day-ahead electricity market, using open data, cross-border information and weather conditions

by

S. Weites

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday July 2, 2026 at 3:30 PM.

Student number:	4914279	
Project duration:	February 2, 2026 – July 2, 2026	
Thesis committee:	Dr. ir. P. W. G. Bots	TU Delft, Chair & second supervisor
	Dr. ir. M. Kroesen	TU Delft, first supervisor
	ing. Y. Z. Hu	Accenture B.V., client
	ir. R. S. Dijkstra	Accenture B.V., supervisor

This thesis is confidential and cannot be made public until December 31, 2026.

An electronic version of this thesis is available at http://repository.tudelft.nl/.
Cover image: AI-generated visual created with ChatGPT for this thesis.

TU Delft

Preface

Writing this thesis has been a challenging but very valuable academic journey. I would like to thank the people who supported me throughout this process.

First, I would like to thank my chair and second supervisor, Dr. ir. P. W. G. Bots, for the guidance, critical feedback, and support during the thesis process. I would also like to thank my first supervisor, Dr. ir. M. Kroesen, for the methodological insight and useful discussions during the modelling part of the research. These discussions helped me to reflect on modelling choices and identify aspects that I might have overlooked otherwise.

I would also like to thank my client and supervisor from Accenture B.V., ing. Y. Z. Hu, and ir. R. S. Dijkstra, for their guidance during the project. The weekly meetings were especially helpful for discussing progress, exploring next steps, and keeping the research focused during the different phases of the thesis.

This thesis marks the end of my master's programme in Complex Systems Engineering & Management. It has further strengthened my interest in the energy system and in the use of machine learning methods to support decision-making in complex electricity markets.

Sam Weites
Delft, June 18, 2026

Executive summary

The Dutch electricity system is becoming increasingly renewable, weather-dependent, and interconnected. This development is essential for the energy transition, but it also changes the behaviour of the day-ahead electricity market. High wind and solar generation can contribute to low or negative prices when supply exceeds demand, while scarcity and limited flexibility can contribute to price spikes. At the same time, the Netherlands is strongly connected to neighbouring bidding zones, which means that Dutch day-ahead prices are influenced not only by domestic conditions, but also by cross-border exchange, neighbouring residual load, renewable generation, and market coupling. Extreme prices should therefore be understood as events in an interconnected regional electricity system.

This creates a challenge for market participants. Traders, retailers, producers, flexibility providers, and large consumers make decisions before delivery, while the actual system conditions are not yet fully known. Anticipating hours with increased downward or upward price risk is therefore valuable. However, negative prices and price spikes are rare, irregular, and difficult to predict with standard price-forecasting approaches. This thesis therefore focuses on probabilistic early warning rather than exact price forecasting. The main research question is:

> *How can ex-ante probabilistic machine-learning models be developed and evaluated to estimate hourly probabilities of negative prices, normal prices, and price spikes in the Dutch day-ahead electricity market, using domestic, weather, and cross-border indicators available before market closure, and how can these probabilities support market decision-making under asymmetric error costs?*

The research addresses this question by developing a reproducible ex-ante early-warning framework. Dutch day-ahead prices are translated into a three-class target variable consisting of negative prices, normal prices, and price spikes. Negative prices are defined as prices below 0 €/MWh, while price spikes are defined as prices above the 95th percentile of the Dutch day-ahead price distribution within the corresponding month and year. The feature set combines Dutch price-history variables, domestic system forecasts, weather variables, cross-border indicators, neighbouring-system variables, and calendar controls. Only variables that can be retrieved before the day-ahead market closes are included. Multiple probabilistic classification models are compared using a chronological train-validation-test split, in which earlier years are used for model development and 2025 is kept as the final unseen test year. The candidate models are logistic regression, a PNN, an MLP, GBDT, and XGBoost. Because negative prices and price spikes occur much less often than normal prices, model performance is not evaluated through overall accuracy alone. Instead, the evaluation combines class-specific recall and precision, the macro F_2 score, balanced accuracy, warning behaviour, probability-quality metrics, calibration diagnostics, target-definition sensitivity, and feature-removal analysis.

The final model comparison shows that tree-based models provide the strongest early-warning performance. XGBoost is selected as the preferred final model because it gives the best overall balance between detecting negative-price and price-spike events, limiting unnecessary warnings, and producing useful probability estimates. On the final 2025 test set, XGBoost achieves a macro F_2 score of 0.812 and a balanced accuracy of 0.895, with high recall for both negative prices (0.969) and price spikes (0.845). It also shows the strongest probability-quality results among the compared models. GBDT achieves comparable results and therefore serves as an important benchmark. In line with the reviewed literature, logistic regression and PNN show that models can detect many rare event hours, while still producing lower precision, more false warnings, and weaker probability quality. This confirms that high recall alone is not sufficient for an early-warning model. The predicted probabilities must also translate into useful warning behaviour for decision support.

The results also show that cross-border information improves the prediction of Dutch extreme prices. For XGBoost, the macro F_2 score increases from 0.702 with the domestic baseline to 0.813 with the core cross-border feature set. This indicates that Dutch extreme-price prediction should not be treated as a purely domestic problem. The strongest contribution comes from neighbouring-system features,

such as load, generation, renewable generation, and residual-load forecasts. Germany/Luxembourg contributes most consistently to the predictive value of the cross-border feature set. The explainability results support this interpretation. Negative-price risk is mainly associated with residual-load conditions, renewable and neighbouring-system indicators, and scheduled exchange, whereas price-spike risk is more strongly linked to residual load, lagged Dutch prices, solar forecasts, and calendar timing. This suggests that Dutch extreme prices are shaped by a combination of domestic market stress, weather-dependent system conditions, and cross-border market coupling.

The estimated probabilities should be interpreted as decision-support signals rather than automatic trading, hedging, or dispatch rules. The model can identify hours with increased negative-price or price-spike risk before day-ahead market outcomes are known, but the appropriate actions depend on the position, contracts, flexibility options, and risk preferences of the user. A lower warning threshold may be suitable when missed events are more costly than false warnings, while a higher threshold may be more appropriate when false warnings trigger costly or unnecessary actions. The validation-selected thresholds are therefore useful for model comparison and warning-behaviour analysis, but actor-specific operational thresholds require additional information about costs and decision context.

The main contribution of this thesis is a Dutch ex-ante probabilistic early-warning framework that combines extreme-price classification, open data, cross-border information, model comparison, explainability, and cost-sensitive interpretation. The practical deliverable is a reproducible modelling pipeline that forms the basis for an early-warning tool. Such a tool could provide hourly probabilities of negative prices and price spikes, together with interpretable drivers that help market participants identify hours requiring closer attention or preparation. At the same time, operational use should also account for possible feedback effects. If many market participants respond to similar warning signals, their bidding, hedging, scheduling, or flexibility decisions may influence the price outcome that the model tries to predict. Several limitations should be taken into account. The final evaluation is based on one out-of-sample test year, and the analysis is limited to hourly day-ahead prices. Intraday and balancing-market outcomes, as well as actor-specific trading and flexibility strategies, are therefore outside the scope of the model. These limitations do not undermine the value of the framework, but they show that operational use requires further development and alignment with actor-specific decision-making processes.

Contents

Preface iii

Executive summary v

1 Introduction 1
- 1.1 Context . . . 1
- 1.2 Background on the subject and the day-ahead market . . . 1
- 1.3 Problem statement . . . 2
- 1.4 Societal relevance . . . 2
- 1.5 Research objective, deliverables, and scope . . . 3
- 1.6 Research questions . . . 3
  - 1.6.1 Main research question . . . 4
  - 1.6.2 Sub-question 1 – Extreme event definition . . . 4
  - 1.6.3 Sub-question 2 – Ex-ante feature design . . . 4
  - 1.6.4 Sub-question 3 – Probabilistic model development and comparison . . . 4
  - 1.6.5 Sub-question 4 – Added value of cross-border indicators . . . 4
  - 1.6.6 Sub-question 5 – Evaluation, error costs, and decision-support interpretation . . . 5
- 1.7 Contributions . . . 5
- 1.8 Thesis structure . . . 5

2 Literature review and knowledge gap 7
- 2.1 Literature review and search strategy . . . 7
  - 2.1.1 Review objective . . . 7
  - 2.1.2 Literature search and selection . . . 7
  - 2.1.3 Comparison of the selected literature . . . 8
- 2.2 Conceptual synthesis of selected literature . . . 8
  - 2.2.1 Day-ahead price formation and the ex-ante constraint . . . 9
  - 2.2.2 Extreme price events and threshold definitions . . . 9
  - 2.2.3 Probabilistic forecasting and early warning . . . 10
  - 2.2.4 Cross-border market coupling and neighbouring market conditions . . . 11
  - 2.2.5 Implications for feature selection and modelling design . . . 12
- 2.3 Conceptual model for this thesis . . . 13
- 2.4 Academic knowledge gap and thesis positioning . . . 15

3 Research design and methods 17
- 3.1 Research approach . . . 17
- 3.2 Workflow and phases . . . 18
  - 3.2.1 SQ1 – Phase 1: extreme event definition . . . 18
  - 3.2.2 SQ2 – Phase 2: ex-ante feature design . . . 18
  - 3.2.3 SQ3 – Phase 3: probabilistic model development and comparison . . . 18
  - 3.2.4 SQ4 – Phase 4: added value of cross-border indicators . . . 19
  - 3.2.5 SQ5 – Phase 5: evaluation, error costs, and decision-support interpretation . . . 19
- 3.3 Operational definitions . . . 19
  - 3.3.1 Definition of extreme events (labels) . . . 19
  - 3.3.2 Ex-ante constraint . . . 20
  - 3.3.3 Model-comparison KPIs . . . 21
- 3.4 Validation strategy . . . 21
- 3.5 Ethics, privacy, and reproducibility . . . 22

4 Data and feature engineering **23**
4.1 Data overview . . . 23
4.2 Target variable construction . . . 25
4.3 Dutch price-history and domestic system features . . . 25
4.4 Weather features . . . 26
4.5 Cross-border coupling and constraint features . . . 27
4.6 Calendar controls . . . 28
4.7 Exploratory data analysis and preprocessing . . . 28
4.7.1 Target-label distribution . . . 28
4.7.2 Time patterns and class imbalance . . . 29
4.7.3 Missing values and data coverage . . . 30
4.7.4 Feature-group differences across target classes . . . 30
4.7.5 Correlation and redundancy checks . . . 32
4.7.6 Final preprocessing choices . . . 33
4.8 Conclusion: ex-ante feature set . . . 33

5 Modelling and evaluation framework **35**
5.1 Modelling problem and prediction . . . 35
5.2 Train-validation-test design . . . 36
5.3 Candidate models and model pipelines . . . 36
5.4 Hyperparameter tuning . . . 37
5.5 Probability outputs and warning-threshold design . . . 38
5.6 Evaluation metrics . . . 39
5.6.1 Class-specific prediction performance . . . 39
5.6.2 Probability rankings . . . 40
5.6.3 Probability quality . . . 40
5.6.4 Warning behaviour and error structure . . . 41
5.7 Probability-quality assessment . . . 41
5.8 Explainability approach . . . 42
5.9 Feature-group removal design . . . 42
5.10 Target-definition sensitivity design . . . 44
5.11 Cost-sensitive interpretation framework . . . 45

6 Results **47**
6.1 Target-definition and test-set characteristics . . . 47
6.2 Final 2025 model comparison . . . 48
6.3 Warning behaviour and error structure . . . 49
6.4 Target-definition sensitivity results . . . 51
6.5 Added value of cross-border indicators . . . 52
6.6 Calibration and reliability . . . 54
6.7 Explainability insights . . . 55
6.8 Cost-sensitive decision interpretation . . . 55
6.9 Conclusion of the results chapter . . . 56

7 Discussion **59**
7.1 Interpretation of model performance . . . 59
7.2 Implications for Dutch market participants . . . 60
7.3 Comparison to literature . . . 61
7.4 Limitations . . . 62
7.5 Potential feedback effects . . . 63
7.6 Recommendations and future work . . . 63

8 Conclusion **65**
8.1 Answer to the sub-questions . . . 65
8.1.1 Sub-question 1 – Extreme event definition . . . 65
8.1.2 Sub-question 2 – Ex-ante feature design . . . 66
8.1.3 Sub-question 3 – Probabilistic model development and comparison . . . 66
8.1.4 Sub-question 4 – Added value of cross-border indicators . . . 66

8.1.5 Sub-question 5 – Evaluation, error costs, and decision-support interpretation . . . 67
8.2 Answer to the main research question . . . 67
8.3 Key contributions . . . 68
8.4 Practical deliverable . . . 69
8.5 Final reflection . . . 69

Appendices **75**

A Literature search strategy **75**
A.1 Search strings . . . 75
A.2 Database filtering criteria . . . 77

B PRISMA flow diagram and literature overview **79**
B.1 PRISMA flow diagram . . . 79
B.2 Extended literature overview . . . 81

C Feature dictionary and ex-ante availability **83**
C.1 Target variable . . . 83
C.2 Dutch price-history variables . . . 83
C.3 Domestic system variables . . . 84
C.4 Weather variables . . . 85
C.5 Cross-border variables . . . 85
C.6 Calendar controls . . . 87

D Model specifications and tuning details **89**
D.1 Candidate models . . . 89
D.2 Initial model configuration . . . 90
D.3 Hyperparameter grid . . . 91
D.4 Final selected model settings . . . 91

E Extended feature-removal results **93**

F Explainability outputs **95**

G AI statement **99**

List of Figures

2.1 Conceptual model linking literature mechanisms, ex-ante observable indicators, probabilistic early-warning output, and later evaluation of usefulness. 14

4.1 Dutch day-ahead prices and labelled extreme-price events across the development training, validation, and final test periods. 29
4.2 Share of negative prices and price spikes across hour of the day and month. 30
4.3 Comparison of the target-class distribution for missing values and the complete dataset 31

6.1 Final 2025 overall model performance 49
6.2 Final 2025 confusion matrices by model 50
6.3 Final 2025 warning counts by model. The vertical line indicates the number of realised event hours. 50
6.4 Target-definition sensitivity of macro F_2 score 51
6.5 Target-definition sensitivity of event-warning volume. The grey part indicates the number of realised extreme hours in each scenario, while the total bar height indicates the number of model warnings 52
6.6 Change in macro F_2 after removing one neighbouring bidding zone from the core cross-border feature set 53
6.7 Reliability curves for negative-price and price-spike probabilities on the final 2025 test set 54

B.1 PRISMA 2020 flow diagram for the scoping review. 80

F.1 SHAP summary plot for negative-price predictions of the final XGBoost model 96
F.2 SHAP summary plot for price-spike predictions of the final XGBoost model 97

List of Tables

4.1 Overview of feature groups used in the modelling framework. 24
4.2 Zero and low-price observations in the final modelling dataset. 25
4.3 Target-label distribution in the final modelling dataset. 28
4.4 Feature-group differences across target classes. 31
4.5 Strongest highly correlated feature pairs in the final dataset. 32
4.6 PCA summary by feature group. 32

5.1 Time-based data split used for model development and final evaluation 36
5.2 Candidate models and their role in the modelling framework 37
5.3 Feature-group scenarios used in the removal analysis 43
5.4 Target-definition scenarios used in the sensitivity analysis 44

6.1 Class distribution across modelling splits 47
6.2 Final 2025 model comparison 48
6.3 Reverse-year macro F_2 performance 49
6.4 Target-definition sensitivity results on the 2025 test set 51
6.5 Feature-group removal results on the 2025 test set 53
6.6 Top SHAP features for the final XGBoost model 55
6.7 Qualitative synthesis of model performance across the four model-comparison KPIs . . 56

A.1 Search strings used in Scopus and Web of Science. 75
A.2 Filtering criteria used for database searches. 77

B.1 Selected literature overview. 81

C.1 Target variable dictionary. 83
C.2 Dutch price-history variable dictionary. 84
C.3 Domestic system variable dictionary. 84
C.4 Weather variable dictionary. 85
C.5 Cross-border variable dictionary. 86
C.6 Calendar control variable dictionary. 87

D.1 Initial model definitions and fixed implementation settings 90
D.2 Hyperparameter grids used during model tuning 91
D.3 Final selected model settings used for the 2025 evaluation 92

E.1 Extended feature-removal results on the 2025 test set 94

List of Abbreviations

ACER Agency for the Cooperation of Energy Regulators. 1

aFRR automatic Frequency Restoration Reserve. 84

AI Artificial Intelligence. 22

API Application Programming Interface. 22, 26, 64

BE Belgian bidding zone. 24, 31, 43, 54, 55, 84–86

CET Central European Time. 11, 20

DE/LU Germany/Luxembourg bidding zone. 24, 31, 43, 54, 55, 57, 67, 68, 84–86

DK1 Denmark bidding zone 1. 24, 31, 43, 54, 84–86

DOI Digital Object Identifier. 8

ELM Extreme Learning Machine. 11, 81

ENTSO-E European Network of Transmission System Operators for Electricity. 23

FBMC Flow-Based Market Coupling. 81

GB Great Britain bidding zone. 24, 27, 62

GBDT Gradient Boosting Decision Trees. v, 36, 43, 44, 48–54, 56, 59, 60, 66–68, 89

KPI key performance indicator. 17–19, 21, 56

LSTM Long Short-Term Memory. 11, 89, 90

mFRR manual Frequency Restoration Reserve. 84

ML machine learning. 3, 4, 17, 18, 67

MLP Multilayer Perceptron. v, 37, 48, 49, 56, 66, 90

NL Dutch bidding zone. 23, 31, 84

NO2 Norway bidding zone 2. 24, 31, 43, 54, 55, 84–86

NTC Net Transfer Capacity. 24, 27, 84

PCA Principal Component Analysis. 32, 33, 37, 43, 53, 55

PNN Probabilistic Neural Network. v, 11, 37, 48–50, 56, 59, 61, 66, 68, 81, 89, 90

PR-AUC precision-recall area under the curve. 54, 93

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses. 8, 79

SDAC Single Day-Ahead Coupling. 1, 9, 18, 62, 81

SHAP SHapley Additive exPlanations. 42, 55

SQ Sub-question. 4, 5, 18–20, 27, 33, 42

SVM Support Vector Machine. 11

TSO Transmission System Operator. 23

UTC Coordinated Universal Time. 28

XGBoost Extreme Gradient Boosting. v, 11, 36, 37, 43, 44, 48–57, 59–61, 64, 66–68, 89, 90

1 Introduction

1.1. Context

The Dutch electricity system is moving towards an energy-generation mix that is increasingly dominated by renewable sources. In 2024, electricity production from renewable sources in the Netherlands rose by 10% compared to 2023, reaching 61 million MWh. As a result, renewables accounted for approximately half of total electricity production, while in the particularly sunny and windy month of April 2024 this share even reached 63% (Statistics Netherlands, 2025). This shift is key to the Dutch and European energy transition, but it also makes electricity supply more weather-dependent. As wind and solar generation fluctuate with weather conditions, the probability of both temporary surplus and temporary scarcity increases, sometimes even within the same day.

These developments are reflected in the wholesale electricity prices. In the Dutch day-ahead market, hours with negative prices are no longer exceptional: 85 hours were negative in 2022, 316 in 2023, and 465 in 2024 (EPEX SPOT, n.d.-b). At the same time, high-price episodes continue to occur when supply is tight or flexibility is insufficient. Together, these developments indicate that the Dutch electricity system is entering a phase in which price volatility and extreme price events are becoming structurally more relevant features of market operation.

1.2. Background on the subject and the day-ahead market

Electricity prices in Europe are mainly formed in the day-ahead market. In this market, participants submit hourly supply and demand bids on the day before physical delivery. The market is then cleared through a pricing mechanism, in which the intersection of aggregated supply and demand determines the hourly market-clearing price (Chatzigiannis et al., 2016). Because the electricity network must be balanced continuously and large-scale storage remains limited, this market plays a large role in coordinating production, consumption, and cross-border exchanges.

The growing share of renewable electricity changes the functioning of this market. Wind and solar generation generally have low marginal costs and therefore tend to lower market prices when output is high. At the same time, their variability makes prices more sensitive to weather conditions, forecast errors, and system flexibility constraints. This increases the likelihood of extreme price outcomes, including both negative prices during oversupply and price spikes during moments of scarcity. Negative prices can emerge when available generation exceeds demand and producers are still willing to remain online. This can be caused, for example, by high shutdown costs or by market incentives that favour continued generation (EPEX SPOT, n.d.-b). In the Single Day-Ahead Coupling (SDAC) framework, the reference harmonised clearing price limits are set at -500 €/MWh and +4,000 €/MWh, with adjustment rules specified by Agency for the Cooperation of Energy Regulators (ACER) (Agency for the Cooperation of Energy Regulators, 2023).

It is also important to note that Dutch price formation is not purely domestic. The Netherlands is strongly interconnected to neighbouring electricity markets through several major interconnectors, including BritNed (1,000 MW) to Great Britain, NorNed (700 MW) to Norway, and COBRAcable (700 MW) to Denmark (TenneT, n.d.-a, n.d.-b, n.d.-e). In addition, the Dutch system is physically linked

to Germany and Belgium through high-voltage interconnections, including 380 kV cross-border infrastructure (TenneT, n.d.-c). Because of these interconnections, Dutch prices are influenced not only by domestic demand and renewable output, but also by weather conditions, generation patterns, demand developments, and congestion in electricity networks in directly connected neighbouring markets.

This cross-border dimension is strengthened by European market coupling. EPEX SPOT explicitly notes that, without coupled markets, negative prices would occur more often and price peaks would be more acute (EPEX SPOT, n.d.-b). Academic literature similarly shows that electricity spot markets in Central Western Europe increasingly share common drivers, especially renewable generation, load, and transmission constraints (De Blauwe et al., 2025). For the Dutch market, this means that extreme prices should not be analysed only from a national perspective, but as the outcome of an interconnected regional market.

1.3. Problem statement

Extreme day-ahead electricity prices are becoming more frequent as renewable generation increases and European market coupling allows surplus and scarcity to propagate across borders. However, it remains difficult to predict when negative prices or price spikes will occur in the Dutch day-ahead market, especially when cross-border coupling, interconnector capacity, and constraint-related signals are taken into account.

This uncertainty creates a practical problem for market participants. Traders, retailers, producers, flexibility platforms, and large consumers must make decisions before delivery and before full information on real-time system conditions is known. They need to determine bidding strategies, manage risk buffers, schedule flexible assets, and anticipate possible market stress.

The usefulness of an early-warning model therefore lies in supporting decisions before the day-ahead market closes, rather than in predicting the exact realised price. The model should identify delivery hours in which negative prices or price spikes are more likely, allowing market participants to focus their attention on periods with increased risk.

The challenge is particularly relevant in the Dutch context, where grid congestion is increasing the value of temporal flexibility. In 2025, TenneT announced that 9.1 GW of capacity could be made available through flexible off-peak contracts, illustrating both the severity of congestion and the growing importance of shifting demand and generation in time (TenneT, 2025). Despite this, there is still limited insight into which combination of Dutch system indicators, neighbouring market signals, and weather-related variables best predicts Dutch extreme day-ahead price events. As a result, decision-makers often lack a structured early-warning and decision-support mechanism for anticipating these events before market closure.

1.4. Societal relevance

The societal relevance of this research lies in the fact that extreme electricity prices affect not only market participants, but also the energy transition in general. Negative prices can undermine revenues for renewable electricity producers, trigger curtailment, and weaken investment incentives when they occur more frequently. Price spikes, in contrast, increase costs and financial risks for retailers, flexibility providers, and energy-intensive industries. This may increase consumer prices and reduce competitiveness (Davi-Arderius & Jamasb, 2026; EPEX SPOT, n.d.-b).

In addition, extreme price events are a signal of underlying system challenges. They indicate moments at which the electricity system struggles to balance and align variable renewable generation, electricity demand, cross-border exchange, and available flexibility. Better anticipation of these events can therefore contribute to more efficient use of batteries, demand response, and industrial flexibility. This is increasingly important in a system where decarbonisation not only depends on building more renewable generation, but also on operating the system in a smarter and more adaptive way.

For the Netherlands specifically, improved early warning of extreme day-ahead prices can support more efficient market decisions under conditions of congestion, weather dependency, and international market integration. By making hours of increased negative-price or price-spike risk more visible before market closure, it may help market participants and flexibility providers respond more effectively to moments of expected surplus or scarcity. This research therefore contributes not only to better forecasting, but also to the societal objective of building a sustainable, secure, and economically efficient electricity system.

1.5. Research objective, deliverables, and scope

This thesis aims to develop and evaluate an ex-ante probabilistic model for extreme prices on the Dutch day-ahead electricity market, using information available before market closure. The research compares multiple machine learning (ML) models and selects the model that performs best in estimating hourly probabilities of negative prices, normal prices, and price spikes for the next day.

The decision-support setting is characterized by three conditions. First, the model is used before day-ahead market closure, meaning that only information available before the market closes can be used. Second, the extreme events are rare compared with normal-price hours, so the model cannot only perform well on the normal class. Third, the error costs are asymmetric. Missing an extreme-price event and issuing a false warning do not have the same consequences, and these consequences differ across market participants.

The conditions are translated into four usefulness criteria for the early-warning model. The model must rely only on information available before market closure, detect rare negative-price and price-spike events, produce probabilities that can be interpreted as risk signals for upcoming delivery hours, and make the trade-off between missed events and false warnings visible.

The analysis is based on a final variable set that combines Dutch price variables, Dutch system variables, weather forecasts, cross-border market variables, scheduled exchange and capacity variables, and calendar controls. The objective is therefore not only to model extreme prices, but also to investigate which combination of ex-ante variables and models provides the most reliable early-warning signals in the Dutch day-ahead market context.

The main deliverables of this thesis are:

- A reproducible dataset with ex-ante features for Dutch day-ahead price classification.
- A comparison of trained ML models for the prediction of negative prices, normal prices, and price spikes.
- A validated final probabilistic early-warning model that outputs, before market closure, hourly class probabilities for the next day as well as the most important drivers for each hour.
- An interpretation of how these probabilities can support decision-making for different market participants, including warning behaviour and a cost-sensitive interpretation of false positives and false negatives.

The scope is limited to the Dutch day-ahead electricity market and directly connected cross-border influences in the feature set. Intraday and balancing markets are outside the scope of this research. Only information available before day-ahead market closure is included in the modelling framework. The focus of this research lies on model development, comparison, validation, and interpretation before market closure, including the implications for decision-making based on the estimated probabilities.

The resulting probabilities are treated not only as predictive outputs, but also as decision-support signals, depending on the relative costs of missed events and false warnings. They are intended to indicate which hours require closer attention, not to trigger automatic trading or dispatch actions.

This research fits within the Complex Systems Engineering & Management MSc, and specifically within the Energy & Industry track. Anticipating extreme prices in the Dutch day-ahead electricity market is not only a technical model-comparison problem. It concerns a socio-technical energy system in which renewable generation, weather conditions, cross-border market coupling, available flexibility, and the day-ahead market setting shape price outcomes together. The thesis therefore combines electricity-market understanding with a data-driven ex-ante modelling and evaluation framework. This reflects the CoSEM focus on complex systems in which technical, economic, and institutional components influence one another. Within the Energy & Industry track, the thesis contributes by developing and evaluating a decision-support framework for an increasingly renewable and interconnected electricity system.

1.6. Research questions

This section presents the main research question and sub-questions that guide the thesis. Together, they translate the research objective into a structured set of phases, covering target definition, feature engineering, model development, model comparison, and the interpretation of model outputs for decision-making.

1.6.1. Main research question

How can ex-ante probabilistic machine-learning models be developed and evaluated to estimate hourly probabilities of negative prices, normal prices, and price spikes in the Dutch day-ahead electricity market, using domestic, weather, and cross-border indicators available before market closure, and how can these probabilities support market decision-making under asymmetric error costs?

1.6.2. Sub-question 1 – Extreme event definition

Which threshold definition and labelling strategy is most appropriate for negative prices, normal prices, and price spikes in the Dutch day-ahead electricity market, taking price distributions, seasonality, and sensitivity to alternative definitions into account?
SQ 1 defines the target variable for the modelling framework. Dutch day-ahead price distributions are analysed across hours and seasons, after which various threshold definitions are derived from the literature. A sensitivity analysis is performed to compare the definitions. The output is a transparent labelling strategy and a labelled dataset in which each hour is classified as negative, normal, or spike.

1.6.3. Sub-question 2 – Ex-ante feature design

Which domestic, weather, cross-border, and calendar indicators can be operationalised as ex-ante predictors of Dutch day-ahead extreme price events, given the constraint that all features must be available before market closure?
SQ 2 translates the mechanisms identified in the literature and the available data sources into an ex-ante feature set. This includes Dutch price-history variables, domestic system indicators, weather forecasts, cross-border market and exchange variables, and calendar controls. Each variable is assessed in terms of publication timing and practical ex-ante availability. The output is a documented feature overview and the final modelling dataset aligned to day-ahead market closure.

1.6.4. Sub-question 3 – Probabilistic model development and comparison

How can candidate probabilistic machine-learning models be evaluated and compared for estimating hourly probabilities of negative prices, normal prices, and price spikes in an ex-ante setting, and which model provides the most useful early-warning performance?
SQ 3 develops the predictive modelling component of the thesis. Multiple probabilistic ML models are trained and compared in an ex-ante setting, with explicit attention to the imbalanced three-class structure of the problem. Model performance is therefore not measured by overall accuracy alone, but through a combination of class-specific performance, probability-quality metrics, calibration, and warning behaviour. This makes it possible to assess whether a model can detect rare negative-price and spike events, whether its predicted probabilities are meaningful as risk signals, and whether the number of false warnings remains practically interpretable.

The output is a structured comparison of the candidate models and their suitability as early-warning models for Dutch extreme price events. This comparison is used to select the model setup that is carried forward to the next steps of the thesis, including the cross-border feature removal, explainability analysis, target-definition sensitivity analysis, and the interpretation of model outputs for decision-making.

1.6.5. Sub-question 4 – Added value of cross-border indicators

To what extent do cross-border coupling, directional scheduled exchange, day-ahead net transfer capacity, and neighbouring market features improve the ex-ante prediction of Dutch negative prices and price spikes compared with models that only use domestic features, and which neighbouring feature groups and bidding zones contribute most?
SQ 4 examines whether the Dutch day-ahead extreme-price warning problem should be treated as a domestic prediction problem or as a regional, interconnected-market problem. The question focuses on whether cross-border capacity, scheduled exchange, and neighbouring market conditions add information beyond Dutch price-history, domestic system, weather, and calendar variables. The output of this sub-question is a conclusion on whether cross-border indicators should be included in the final early-warning framework and which types of neighbouring information are most relevant for the Dutch case.

1.6.6. Sub-question 5 – Evaluation, error costs, and decision-support interpretation

How should the estimated class probabilities be interpreted as decision-support signals for market participants, considering asymmetric false-positive and false-negative costs?
SQ 5 connects the probabilistic forecasts to their intended use as early-warning signals. Because missed negative-price or spike events and false warnings do not have the same consequences for all market participants, the probability outputs are interpreted through the trade-off between false positives and false negatives. The output of this sub-question is a decision-oriented interpretation of how model probabilities can support attention, monitoring, and preparation before day-ahead market closure.

1.7. Contributions

This thesis contributes scientifically in four ways. First, it develops a reproducible ex-ante modelling framework for extreme price events in the Dutch day-ahead electricity market. This framework combines Dutch price variables, domestic system indicators, weather forecasts, cross-border market and exchange variables, and calendar controls into one dataset.

Second, the thesis contributes methodologically by comparing multiple probabilistic machine-learning models for the classification of negative prices, normal prices, and price spikes. In doing so, it evaluates which modelling approach performs best under realistic ex-ante conditions.

Third, the thesis provides empirical evidence on the added value of cross-border indicators for Dutch extreme price events, thereby contributing to the literature on interconnected electricity markets and market coupling.

Fourth, the thesis interprets the estimated probabilities in terms of market decision-making under asymmetric error costs. This includes a cost-sensitive interpretation of false positives, false negatives, warning volumes, and probability levels. This way, the thesis not only contributes to the academic literature on extreme electricity price forecasting, but also to the practical understanding of how ex-ante probability estimates can support risk monitoring and decision preparation in electricity markets.

1.8. Thesis structure

This thesis is structured as follows. Chapter 2 presents the literature review and academic knowledge gap. It defines the core concepts of day-ahead price formation, extreme price events, probabilistic forecasting, and cross-border coupling. These are used to position the thesis in the existing literature and identify the academic knowledge gap.

Chapter 3 presents the research design and methods. It translates the main research question into the five phases described earlier: the definition of extreme price events, the design of an ex-ante feature set, the development and comparison of different probabilistic machine-learning models, the assessment of the added value of cross-border indicators, and the interpretation of model outputs as decision-support signals under asymmetric error costs.

Chapter 4 describes the data and feature engineering process. It introduces the data sources, target variable construction, and the final feature groups used in the modelling framework, consisting of Dutch price-history variables, domestic system indicators, weather features, cross-border variables, and calendar controls. It also discusses ex-ante data availability and preprocessing choices.

Chapter 5 describes the modelling and evaluation approach. It specifies the candidate models, the validation strategy, probability calibration, class imbalance treatment, explainability methods, and the framework used to evaluate predictive performance, robustness, and decision relevance.

Chapter 6 presents the modelling and evaluation results. It reports descriptive statistics, compares the forecasting performance of the different models, evaluates calibration and reliability, analyses the added value of cross-border indicators, and summarises the most important explainability insights.

Chapter 7 discusses the findings in relation to the literature and the Dutch electricity market context. It reflects on the interpretation of the modelling results, the practical implications of the estimated probabilities, the asymmetric costs of prediction errors, and the main limitations of the study. Potential feedback effects from model-based warnings are also discussed.

Finally, Chapter 8 concludes the thesis by answering the sub-questions and main research question, summarising the key contributions, describing the practical deliverable, and reflecting on the value and limitations of the early-warning framework.

2

Literature review and knowledge gap

2.1. Literature review and search strategy

This review focuses on the academic literature needed to position this thesis within existing research on extreme day-ahead electricity prices and forecasting, as well as cross-border market coupling. The purpose of the review is to identify the concepts, mechanisms, variables, and methodological choices that are relevant for developing an ex-ante early-warning model for extreme prices in the Dutch day-ahead market.

The literature review is guided by the main objective of the thesis: developing an ex-ante probabilistic model that estimates the hourly probability that the Dutch day-ahead price of the next day belongs to one of three classes: negative price, normal price, or price spike. The reviewed literature is therefore assessed in terms of how it contributes to four aspects of the thesis design. First, the literature is used to define and conceptualise extreme price events. Second, it helps identify the market mechanisms that can explain negative prices and price spikes. Third, it supports the use of probabilistic forecasting with classification. Finally, it guides the selection of domestic and cross-border indicators for the ex-ante feature set.

2.1.1. Review objective

The review is guided by four analytical questions. First, how does the literature conceptualise day-ahead price formation and the occurrence of extreme prices? Second, how are negative prices and price spikes defined and distinguished in existing studies? Third, which forecasting approaches are used to estimate the probability or occurrence of extreme electricity prices? Fourth, how does the literature take cross-border market coupling and neighbouring market conditions into account?

These questions reflect the objective of this thesis, which is to create a model that estimates the hourly probability that the Dutch day-ahead price of the next day belongs to one of three classes: negative price, normal price, or price spike. The literature review therefore focuses on studies that relate to the conceptualisation of these classes and the selection of explanatory variables, as well as the design of the probabilistic modelling approach. The findings of the review are synthesised in the following sections and are later translated into the feature-selection strategy in Chapter 4.

2.1.2. Literature search and selection

The literature search was conducted in Scopus and Web of Science and focused on academic studies in English published between 2006 and 2026. This range was chosen to capture relevant developments in coupled electricity markets, since the Dutch, French and Belgian power exchanges have been coupled since 2006 (TenneT, n.d.-d).

The search strategy was designed around the four analytical questions introduced in Section 2.1.1. Therefore, the search strings combined terms related to day-ahead electricity markets, extreme price events, probabilistic forecasting, and cross-border market coupling. The search terms included, among others, *day-ahead market*, *spot market*, *electricity price*, *negative prices*, *price spikes*, *extreme prices*, *probabilistic forecasting*, *classification*, *early warning*, *market coupling*, *cross-border flows*, and *interconnection*. Six search strings were developed in total and these are provided in Appendix A.

An example of one of the search strings is:

"*("day-ahead" OR "day ahead" OR dayahead OR "day-ahead market" OR "spot market" OR "DA price") AND (electricity OR power) AND price* AND ("negative price*" OR (price* W/3 spike*) OR ("extreme" W/2 price*) OR ("tail" W/2 event*) OR "scarcity pricing" OR "price volatility" OR "negative electricity price*") AND ("probabilistic forecast*" OR "probability forecast*" OR "quantile forecast*" OR "interval forecast*" OR "early warning" OR "risk forecast*" OR "event prediction" OR classification))*"

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) method (Page et al., 2021) was used for the screening process to support reproducibility. First, duplicate records were removed by comparing titles, author information, and Digital Object Identifiers (DOIs). Afterwards, titles and abstracts were screened for relevance to the scope of the thesis, and finally the remaining full texts were assessed using inclusion and exclusion criteria. The full screening process, including the PRISMA flowchart, search terms and search criteria, is shown in Appendix A and Appendix B. The final selection resulted in 12 journal articles, which are shown in Table B.1.

2.1.3. Comparison of the selected literature

After the final selection, the articles were compared through a structured literature overview. This step was used to identify how the selected studies contribute to the conceptualisation, variable selection, and modelling approach of this thesis.

For each article, attention was paid to the price phenomenon being studied and the market context in which the study was conducted, as well as the way in which extreme prices were defined or operationalised. The comparison also considered which explanatory variables were used, how these variables were linked to the underlying market mechanisms, and which methodological approach was applied. Finally, each article was assessed in terms of its relevance for the Dutch day-ahead market and its limitations for developing an ex-ante probabilistic early-warning model.

This structured comparison makes it possible to synthesise the selected literature across definitions, mechanisms, methods, and feature groups. The detailed comparison table is provided in Appendix B. The literature is synthesised around the main themes of this thesis: day-ahead price formation, negative prices, price spikes, probabilistic early warning, and cross-border market coupling. This synthesis provides the basis for the feature-selection process developed in Chapter 4.

2.2. Conceptual synthesis of selected literature

The selected literature is synthesised in this section to show how existing research supports the conceptualisation, feature selection, and modelling approach of this thesis. The synthesis is structured around four core concepts that correspond to the analytical questions introduced in Section 2.1.1: day-ahead electricity price formation, extreme price events, probabilistic forecasting and early warning, and cross-border market coupling. Together, these concepts define the scope of the thesis and provide the basis for the empirical design in following chapters.

A more descriptive overview of the selected articles, including their broader focus, market context, methods, and key takeaway, is provided in Appendix B. Taken together, the selected studies indicate that extreme Dutch day-ahead prices should be treated as tail events that occur under conditions of market stress and flexibility limitations, but can also occur as a result of cross-border influences. However, studies so far rarely integrate these insights into one ex-ante modelling framework for the Dutch day-ahead market.

For this thesis, the system is conceptualised as a day-ahead clearing process in which expected system conditions influence the probability that the market-clearing price falls into the lower or upper tail of the price distribution (Chatzigiannis et al., 2016). Negative prices are interpreted as lower-tail outcomes that can arise under surplus conditions and limited downward flexibility (De Vos, 2015). Price spikes are interpreted as upper-tail outcomes that can arise under tight system conditions, high demand, or limited reserve capacity (Liu et al., 2022). Cross-border coupling can affect both tails, since neighbouring markets and available transfer capacity may either absorb domestic surplus and scarcity or transmit external market conditions into the Dutch bidding zone (De Blauwe et al., 2025; Ovaere et al., 2023).

2.2.1. Day-ahead price formation and the ex-ante constraint

The first core concept is day-ahead electricity price formation. This refers to price formation in day-ahead markets, where electricity is traded before physical delivery and prices are cleared through auctions. Participants offer electricity and submit bids for delivery during specific hours of the next day. The resulting day-ahead price is an hourly market-clearing outcome that reflects expected supply and demand conditions before delivery (Chatzigiannis et al., 2016; De Blauwe et al., 2025).

In this mechanism, price formation is determined by the submitted supply and demand bids, the marginal bid needed to clear the market, and the physical and institutional constraints included in the auction (Chatzigiannis et al., 2016). Renewable generation with low marginal costs can shift the supply side downward when output is high, while scarce dispatchable capacity, high demand, or limited reserve capacity can move the clearing price upward. Transmission constraints and cross-border exchange possibilities can affect this outcome by determining how far surplus or scarcity conditions can be shared across bidding zones (Ovaere et al., 2023).

The auction-based view is important for this thesis, because the Dutch day-ahead price is not treated only as a historical time series. Instead, extreme prices are interpreted as outcomes of a market process in which relatively small shifts in expected system conditions can result in large hourly price changes (Chatzigiannis et al., 2016). This means that negative prices and price spikes should not only be interpreted as statistical outliers, but also as tail outcomes of a market process with both physical and institutional constraints.

The objective is therefore not to reproduce the full SDAC mechanism, but to identify conditions under which tail outcomes become more likely. Since the aim is to develop an early-warning model, these conditions must be observable before the day-ahead market is cleared. Features available before market clearing can then be used as indicators of situations in which the market-clearing process results in negative prices or price spikes. Variables that are only known after market clearing may be useful for explanation or robustness checks, but cannot be part of the main prediction model.

The literature on electricity price forecasting often uses historical prices, load, generation, renewable production, and interconnector information as explanatory variables. Liu et al. (2022), for example, include historical prices, reserve capacity, load demand, variable renewable energy proportion, and interconnector flow when forecasting the occurrence of extremely high and low prices. This is relevant because it shows that extreme-price forecasting should not only rely on past prices, but also on variables that represent expected system conditions. Historical prices can capture persistence, volatility, and recurring price patterns, while system and market variables capture the conditions under which the next day-ahead auction is cleared.

The main takeaway is that extreme price forecasting in the Dutch market should be treated as tail-risk estimation. The model should not only learn historical price patterns, but also use variables that represent expected system stress before the day-ahead auction closes. This provides the first link between the literature and the modelling design: the feature set must reflect expected market stress before the price is known.

2.2.2. Extreme price events and threshold definitions

The second core concept is extreme price events. Extreme prices are understood as observations in the lower and upper tails of the day-ahead price distribution, including both negative prices and price spikes. Negative prices refer to hours in which the day-ahead price falls below zero. Price spikes refer to unusually high prices compared to the rest of the price distribution (De Vos, 2015; Liu et al., 2022).

This distinction is important because negative prices and price spikes represent different directions of market stress within the day-ahead clearing process. They are therefore not necessarily caused by the same market conditions. Negative prices are often associated with surplus conditions, while price spikes are associated with scarcity conditions. Negative prices can occur when supply is difficult to reduce, demand is low, renewable production is high, or export and storage possibilities are insufficient (De Vos, 2015). Recent work also indicates that zero-negative prices may include additional system costs, such as balancing and ancillary-service costs (Davi-Arderius & Jamasb, 2026). Price spikes, on the other hand, are generally linked to tight system conditions, high demand, limited available generation, limited reserve capacity, outages, interconnector constraints, or sudden changes in system conditions (Amjady & Keynia, 2011; Liu et al., 2022).

For that reason, extreme prices are not treated as one single category. Instead, negative prices and price spikes are separated into two event classes. This avoids combining two different market situations

into one label. It also allows the model to learn different patterns for surplus-driven and scarcity-driven events.

Negative prices have a clear definition, because they occur when the day-ahead price is below zero. Price spikes, however, require an explicit threshold definition. There is no fixed boundary above which a price automatically becomes a spike. Existing studies therefore use fixed thresholds, percentile-based thresholds, or data-dependent threshold approaches to identify upper-tail events (Amjady & Keynia, 2011; Liu et al., 2022; Lu & Suthaharan, 2023).

These threshold choices are not only technical details. They influence how many hours are labelled as extreme, how imbalanced the modelling problem becomes, and what type of price behaviour the model is trained to detect. Fixed thresholds are easy to interpret, because they define price spikes above one absolute price level. However, they may be too rigid when market price levels change over time. A fixed threshold can classify many hours as extreme in a high-price period and almost none in a low-price period.

Percentile-based thresholds are more adaptive, because they define spikes relative to the observed price distribution. Liu et al. (2022) use threshold-based definitions to analyse extremely low and high prices and show that the forecasting performance can differ across threshold choices. This supports the use of a sensitivity analysis, because the definition of an extreme price affects the resulting event set. At the same time, percentile-based thresholds also depend on the selected sample period. If the price distribution changes strongly over time, the same percentile may correspond to different absolute price levels.

Data-dependent threshold approaches go one step further by trying to identify spikes from the behaviour of the price series itself. Lu and Suthaharan (2023) use a two-regime threshold autoregression to determine price spikes before modelling the clustering of spike events. This is useful because it recognises that spikes may be part of a separate price regime. However, it also makes the definition more complex and less transparent. The advantage of these approaches is that they are closer to the statistical structure of the data, while the disadvantage is that the threshold becomes harder to explain and reproduce.

The literature does not provide one universally accepted spike definition. Instead, it shows a trade-off between interpretability, robustness, and sensitivity to changing market conditions. This is important for this thesis because the spike threshold must be suitable for the Dutch day-ahead market and for a classification model with rare events. A fixed threshold would be easy to explain, but may not account for seasonal and yearly differences in the price distribution. A percentile-based threshold is more consistent with the aim of identifying the upper tail of the Dutch price distribution. The exact definition and operationalisation of price spikes, as well as the labelling strategy, are discussed in Chapter 3.

This leads to a three-class target variable: negative prices, normal prices, and price spikes. The exact operationalisation of these labels is discussed in Chapter 3, while the sensitivity of the results to alternative target definitions is evaluated later in Chapter 5 and Chapter 6.

2.2.3. Probabilistic forecasting and early warning

The third core concept is probabilistic forecasting and early warning. Since this thesis focuses on early warning, estimating probabilities is more important than predicting the exact height of the electricity price in each hour. The intended output is an hourly probability that the Dutch day-ahead price of the next day belongs to one of the three price classes: negative, normal, or spike. This follows the probabilistic forecasting literature, which argues that probabilistic forecasts are useful when electricity prices are volatile and when uncertainty around future outcomes is relevant for decision-making (Haben et al., 2021; Liu et al., 2022).

This changes the modelling problem from a standard price-level forecasting problem into an event-probability problem. Instead of asking what the exact day-ahead price will be, the model estimates the probability of a relevant price event. This distinction is important because exact price-level forecasting and event prediction serve different purposes. Price-level forecasts aim to minimise the numerical distance between predicted and realised prices, which can work well for normal prices, but extreme prices are rare and often behave as outliers. Liu et al. (2022) argue that existing point-forecasting models often perform well for normal electricity prices, but have lower accuracy for extreme prices because these events occur suddenly and with large magnitude. For an early-warning model, the main question becomes not whether the exact price level can be predicted, but whether risky hours can be identified early enough. Market participants can use probabilities to identify hours that require closer

monitoring or possible actions. The model output should therefore be interpreted as a risk signal, and not as an exact prediction of the hourly price.

The relevance of uncertainty representation is also shown by Loizidis et al. (2024). Their results show that forecasts based on intervals can represent normal prices and extremely high prices. Negative prices, however, are harder to predict because they fall outside the lower bounds of these intervals. This supports the use of probabilistic outputs and shows that negative prices require explicit attention, rather than being treated as forecast errors (Loizidis et al., 2024).

Literature also shows that probabilistic and classification models can be used to estimate the occurrence of extreme prices. Logistic regression provides an interpretable probability baseline and can be used to relate extreme-price probabilities to market variables (Liu et al., 2022). Spike-oriented studies also show that separating event occurrence from price level can improve the forecasting task. This is because predicting whether an event occurs is not the same as predicting the exact value of the price during that event (Amjady & Keynia, 2011).

Loizidis et al. (2025) are especially relevant because they compare several classification models for negative day-ahead price forecasting. In their study of the German market, a Probabilistic Neural Network (PNN) classifier is benchmarked against Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM) models. The PNN performs best overall for the negative-price classification task, while XGBoost provides a useful benchmark in the comparison as a common tree-based classifier with relatively limited computational demands. This comparison shows why classification models are relevant for forecasting rare electricity price events and provides a basis for selecting similar model types in this thesis.

At the same time, their modelling setting differs from the aim of this research. Loizidis et al. (2025) distinguish between negative and positive prices and integrate the classifier into an Extreme Learning Machine (ELM)-Bootstrap forecasting framework. This thesis instead focuses on the Dutch market, estimates hourly class probabilities, and separates prices into three classes: negative prices, normal prices, and price spikes. The models compared by Loizidis et al. (2025) are therefore not adopted directly, but are used to support the use of comparable classification approaches in the modelling phase.

A central constraint of this early-warning approach is that the model can only use information available ex-ante. This means that explanatory variables must be available before the day-ahead market is cleared, which is at 12:00 Central European Time (CET) before delivery on the next day (De Vos, 2015; Haben et al., 2021). Variables that are only known after market clearing may be useful for explanation or robustness checks, but cannot be part of the main prediction model. This constraint is used in the feature-selection strategy in Chapter 4.

The probabilistic forecasting aspect also affects the way model performance is evaluated. Since the model produces probabilities, performance should not only be assessed by whether the most likely class is correct. It is also important to look at the quality of the predictions. This makes probabilistic and class-specific performance relevant for assessing whether the model can function as an early-warning tool. This performance should be interpreted in terms of class-specific errors, probability quality, calibration, and decision relevance. Recent negative-price forecasting literature shows that model evaluation should account for asymmetric error costs, since missed events and false warnings do not have the same consequences (Loizidis et al., 2025).

The literature therefore does not identify one model type as clearly superior for the specific forecasting task addressed in this thesis. Existing studies use and compare different approaches, including statistical models, neural networks, tree-based methods, hybrid models, and probabilistic classifiers (Haben et al., 2021; Liu et al., 2022; Loizidis et al., 2024). The implication for this thesis is that the model choice should not be fixed in advance. Instead, several candidate models are compared in Chapter 5 to assess which approach is most suitable for estimating early-warning probabilities for Dutch negative prices and price spikes.

2.2.4. Cross-border market coupling and neighbouring market conditions

The fourth core concept is cross-border market coupling. Dutch electricity prices are influenced by cross-border flows and coupled market dynamics. The Netherlands is connected to neighbouring bidding zones, and these interconnections allow electricity to be exchanged across borders when transmission capacity is available. As a result, Dutch day-ahead prices may be affected not only by domestic influences, but also by neighbouring market conditions (De Blauwe et al., 2025; Ovaere et al., 2023).

The literature on cross-border effects indicates that coupled electricity markets are shaped by a combination of domestic and neighbouring-market conditions. Relevant factors include neighbouring prices, demand, renewable generation, interconnector flows, available transmission capacity, and scheduled exchanges (De Blauwe et al., 2025; Ovaere et al., 2023). Market coupling can reduce price differences between bidding zones by enabling cross-border trade and allowing surplus or scarcity to be absorbed. However, it can also transmit scarcity or surplus conditions across markets when neighbouring systems experience similar stress or when network constraints limit the available exchange. For the Dutch day-ahead market, this implies that neighbouring-system and transmission variables may contain useful information for anticipating extreme-price events. Studies on flow-based market coupling further show that cross-border exchanges and price convergence depend on the commercial capacity made available to the market and on how network constraints are represented in the market-clearing process (Ovaere et al., 2023).

This means that cross-border capacity and flow features can contain information about whether surplus or scarcity conditions can be absorbed through trade. They may also indicate whether these conditions are more likely to remain within the Dutch bidding zone or be transmitted into it. Evidence from the Danish market shows that wind production and planned cross-border flows can affect the probability of price differences between connected bidding zones (Unger et al., 2018). These findings suggest that cross-border indicators should not only be treated as background information, but as potential predictors of Dutch price risk.

At the same time, the literature does not imply that every cross-border variable is automatically useful for prediction. Cross-border coupling can smooth local scarcity or surplus when transmission capacity is available, but it can also be limited by congestion or by insufficient capacity. This means that neighbouring conditions are only predictive when they can affect the Dutch market through actual or expected coupling mechanisms. The question then becomes not whether neighbouring markets matter in general, but whether cross-border indicators add predictive value beyond domestic Dutch features in an ex-ante forecasting setting.

For this reason, cross-border indicators are included as part of the research. It is not assumed in advance that all neighbouring-market variables improve prediction. Their added value is tested by comparing models and feature groups. Chapter 4 later translates this concept into usable variables such as lagged neighbouring prices, cross-border flows, available transmission capacity, and neighbouring demand and renewable generation indicators.

2.2.5. Implications for feature selection and modelling design

The synthesis of the selected literature leads to several implications for the design of this thesis. First, the target variable should separate negative prices, normal prices, and price spikes. This follows from the fact that the lower and upper ends of the price distribution are related, but are associated with different market conditions. Negative prices and price spikes occur under different system conditions, and are therefore likely to require different predictive signals and different interpretations of prediction errors.

Second, the feature set should include variables that represent domestic system stress, such as load, renewable generation, residual load, price history, and recent event indicators. This follows from the literature on day-ahead price formation and extreme-price forecasting, where both historical prices and market fundamentals are used to explain or predict extreme events. Price history is relevant because extreme prices can show persistence and clustering over time, while system variables are relevant because they represent the expected supply-demand balance before market clearing.

Third, the feature set should include cross-border variables, because neighbouring market conditions may cause scarcity or surplus in the Dutch bidding zone. Neighbouring prices, neighbouring load and renewable forecasts, scheduled exchanges, and available transfer-capacity variables may all provide early signals of Dutch price risk. These variables reflect neighbouring market stress and cross-border constraints. Their added predictive value is not assumed in advance and is therefore tested empirically.

Fourth, the model should be restricted to ex-ante available information. The aim of this research is to create an early-warning model, which means that variables known after market clearing cannot be used. The ex-ante constraint is therefore not only a practical data choice, but a conceptual constraint for an early-warning model.

Fifth, the modelling approach should produce probabilities rather than only class labels. This allows the model output to be evaluated as a risk signal and later linked to decision thresholds. Since extreme prices are rare and uncertain, the model should not only predict the most likely class, but also provide probability estimates that can be evaluated and interpreted as warning signals.

Together, these implications define the position of the thesis. It translates explanatory literature on extreme prices into an ex-ante modelling framework that estimates hourly event probabilities before the day-ahead market is cleared. Day-ahead electricity prices and auctions define the market context, extreme price events define the target variable, probabilistic forecasting and early warning define the modelling objective and data availability constraint, and cross-border market coupling defines the neighbouring-market aspect of the thesis. Chapter 3 defines the research design and operational choices, and Chapter 4 operationalises the feature groups into the final modelling dataset.

2.3. Conceptual model for this thesis

The literature synthesis in Section 2.2 is translated into a conceptual model that links the reviewed mechanisms to the empirical design of this thesis. The model uses ex-ante observable indicators to represent conditions under which negative prices or price spikes become more likely.

The conceptual model is structured around three market-stress mechanisms: domestic market stress, flexibility stress, and cross-border transmission or absorption of surplus and scarcity. These mechanisms explain why the selected feature groups are expected to contain information about Dutch price extremes. Domestic system conditions, weather variables, calendar effects, and cross-border indicators are therefore not treated as direct causes of price extremes, but as observable signals of market conditions that may increase the likelihood of one of the three price classes. This conceptual translation is summarised in Figure 2.1, which shows how the literature mechanisms are operationalised through ex-ante indicators and then translated into probabilistic warning signals.

The figure should be read from top to bottom. The day-ahead market context defines the ex-ante constraint, while the literature mechanisms explain why different groups of indicators may contain information about future price outcomes. The mechanisms are operationalised through multiple indicators, because system conditions, flexibility limitations, and cross-border coupling interact during the market-clearing process. Together, the indicator groups form the input for the ex-ante model, which is used to estimate hourly class probabilities that can be interpreted as warning signals.

The first mechanism is domestic market stress. Since day-ahead prices are formed through expected supply and demand conditions before delivery, domestic indicators can provide information about the likelihood of negative prices or price spikes (Chatzigiannis et al., 2016; Liu et al., 2022). In this mechanism, surplus conditions increase the likelihood of the lower tail, while scarcity conditions increase the likelihood of the upper tail. Domestic market stress is mainly reflected in domestic system indicators, Dutch and neighbouring weather indicators, and calendar controls, but can also be captured partly by price-history or recent-event indicators.

The second mechanism is flexibility stress. Negative prices can occur when the system cannot absorb surplus electricity or reduce production, while price spikes can occur when the system cannot respond to scarcity (De Vos, 2015; Lu & Suthaharan, 2023). Flexibility stress therefore explains why surplus or scarcity conditions may move the price into one of the tails of the distribution. In the feature set, this mechanism is operationalised through domestic system indicators and recent-event indicators, as well as neighbouring-market and cross-border indicators. Neighbouring-market and cross-border indicators are included because they can signal whether surplus or scarcity conditions are likely to be absorbed through connected bidding zones.

The third mechanism is cross-border transmission or absorption of surplus and scarcity. The Dutch day-ahead market is connected to neighbouring bidding zones, meaning that domestic price formation may be affected by external neighbouring market conditions. Cross-border coupling can reduce price differences when sufficient capacity is available, but it can also transmit market stress from one bidding zone to another (De Blauwe et al., 2025; Ovaere et al., 2023). It is therefore operationalised through neighbouring-market and cross-border indicators, with neighbouring weather indicators included to represent conditions in connected bidding zones.

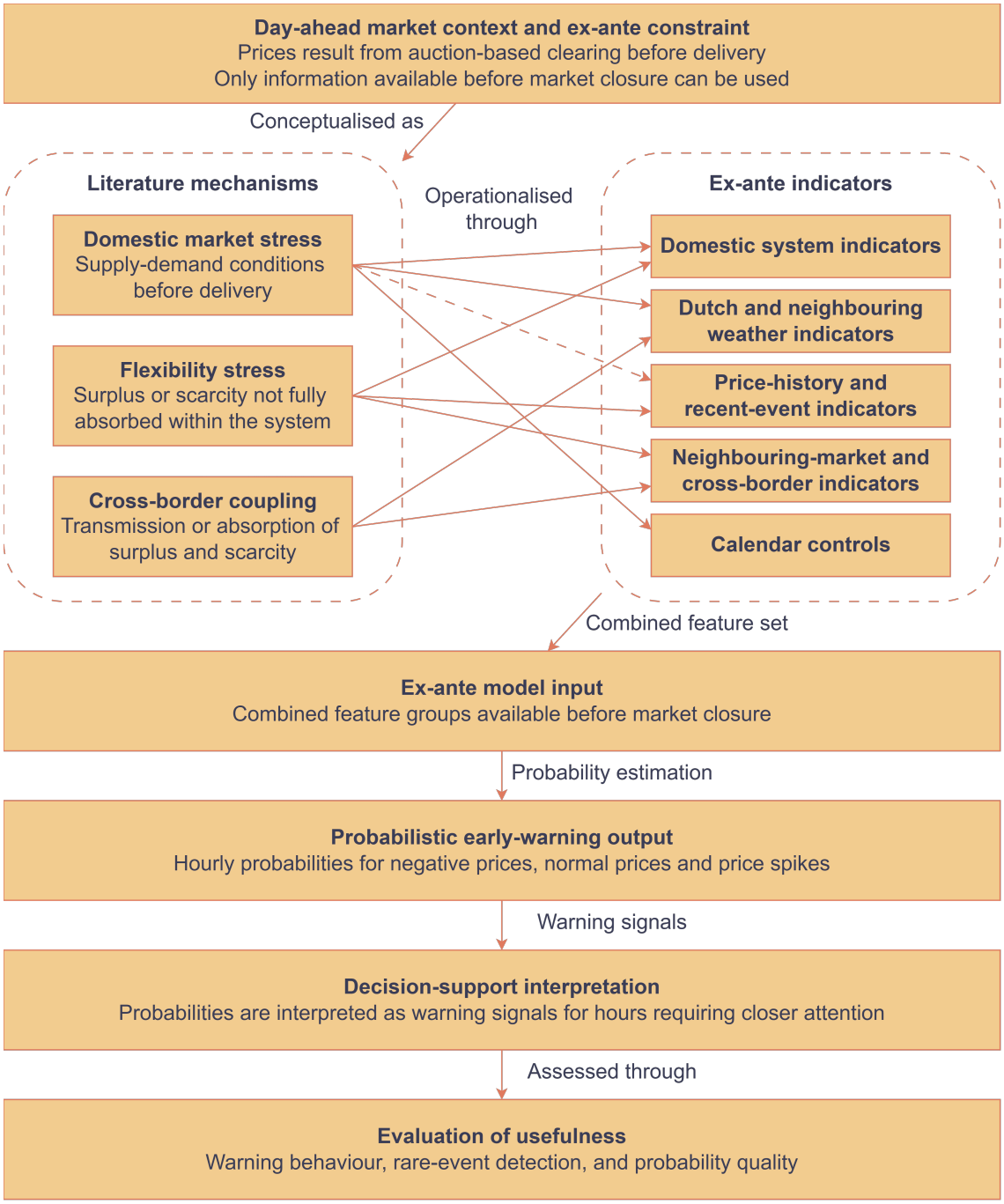

Figure 2.1: Conceptual model linking literature mechanisms, ex-ante observable indicators, probabilistic early-warning output, and later evaluation of usefulness.

Together, these mechanisms define the conceptual model used in this thesis. Ex-ante observable variables are used as indicators for domestic market stress, flexibility stress, and cross-border transmission or absorption. These indicators do not reproduce the full market-clearing process, but help identify conditions under which the realised day-ahead price is more likely to fall into one of three classes: negative price, normal price, or price spike. The empirical modelling task is then to estimate the probability of each of these classes before market closure.

The conceptual model forms the bridge between the literature review and the empirical design. Chapter 3 defines the target classes and research design, Chapter 4 translates the mechanisms into feature groups, and Chapter 5 develops the probabilistic modelling and evaluation approach. In Chapter 5, the decision-support interpretation is operationalised through warning behaviour, rare-event detection and probability quality.

2.4. Academic knowledge gap and thesis positioning

The literature reviewed in this chapter provides useful insights into separate parts of the research problem. Market-clearing literature explains how day-ahead prices are formed through auction-based clearing under technical and transmission constraints. Literature on negative prices explains the role of surplus conditions, renewable generation, low demand, and limited downward flexibility. At the other end of the price distribution, literature on price spikes shows that upper-tail events are rare, can cluster, and are linked to scarcity conditions. Probabilistic forecasting literature shows that event probabilities can be useful for decision-making under uncertainty. Finally, cross-border literature shows that neighbouring market conditions can influence prices in coupled electricity markets.

However, these insights are not yet fully combined in one ex-ante early-warning framework for the Dutch day-ahead market. Existing studies often focus on either negative prices or price spikes, rather than modelling both tails of the price distribution together. Other studies focus on exact price-level or interval forecasting, rather than on multi-class event probabilities. Cross-border studies show that neighbouring market conditions matter, but the added predictive value of cross-border features for Dutch extreme price events has not yet been tested explicitly in an ex-ante probabilistic model. The academic knowledge gap addressed in this thesis is therefore the lack of a Dutch ex-ante early-warning framework that combines both price tails, domestic and cross-border indicators, and probabilistic multi-class prediction.

This thesis addresses that gap by treating negative prices and price spikes as separate but related tail events, testing whether cross-border indicators improve the prediction of Dutch extreme price events, and evaluating models based on probabilistic early-warning performance rather than only overall classification accuracy.

The next chapter translates this thesis positioning into the research design, operational definitions, and methodological choices used in the empirical part of the thesis.

3

Research design and methods

Building on Chapters 1 and 2, this research treats extreme prices in the Dutch day-ahead market as tail events that emerge under conditions of domestic market stress, flexibility stress, and cross-border market coupling. The objective is to develop and evaluate an ex-ante probabilistic modelling framework that estimates, for each hour of the next day, the probability of negative prices, normal prices, and price spikes using only information available before day-ahead market closure. This chapter translates the knowledge gap and conceptual model from Chapter 2 into the research design. It defines the phased workflow, operational definitions, validation strategy, and reproducibility choices used in Chapter 4 and Chapter 5.

3.1. Research approach

A design-oriented modelling approach is used to answer the main research question (Hevner et al., 2004):

How can ex-ante probabilistic machine-learning models be developed and evaluated to estimate hourly probabilities of negative prices, normal prices, and price spikes in the Dutch day-ahead electricity market, using domestic, weather, and cross-border indicators available before market closure, and how can these probabilities support market decision-making under asymmetric error costs?

This approach fits the intended contribution of the research by developing and evaluating a reproducible modelling framework for a complex socio-technical market setting. In this way, technical signals from the Dutch and neighbouring electricity systems can be translated into probabilistic risk estimates that are relevant for decisions before market closure.

The research design follows directly from the conceptual model discussed in Chapter 2. The mechanisms of domestic market stress, flexibility stress, and cross-border transmission of scarcity and surplus are translated into different feature groups. The modelling part is then framed as an ex-ante multi-class classification problem, where each hour is assigned a probability of belonging to one of three classes: negative price, normal price, or price spike.

In this chapter, the conceptual model is translated into an early-warning framework. The approach moves from extreme-event definition and ex-ante feature design to probabilistic model development, model comparison and evaluation, cross-border feature assessment, explainability, and cost-sensitive decision interpretation. These steps link the conceptual mechanisms from Chapter 2 to the empirical design, the model-comparison key performance indicators (KPIs), and the final decision-support interpretation.

Within this design-oriented approach, the research combines two connected modelling components:

1. **Probabilistic risk estimation**, the predictive component. Multiple probabilistic ML models are used to estimate, for each hour of the next day, the probabilities of three price classes: negative price, normal price, and price spike. Special attention is given to the rarity of tail events, the clustering of extreme prices, and the constraint that only ex-ante features may be used.

Because the literature does not identify one modelling approach as clearly most appropriate for the Dutch case, model choice is evaluated through explicit comparison.

2. **Decision-oriented interpretation**, the evaluative component. The estimated class probabilities are interpreted in terms of market risk, warning behaviour, and asymmetric error costs before market closure. This component is included because the model output should be useful not only as a prediction, but also as a risk signal. Model evaluation is therefore not based on predictive performance alone, but also on the extent to which the outputs can be interpreted for decision-making.

One limitation of this approach is that the modelling framework cannot reproduce the full SDAC market-clearing process. Instead, it approximates tail-risk conditions using ex-ante variables that act as indicators of domestic market stress, flexibility stress, and cross-border pressure. This limitation is addressed through time-based validation, explicit documentation of ex-ante feature availability, and robustness checks.

3.2. Workflow and phases

To answer the main research question, the research is structured into five phases that correspond to the sub-questions introduced in Chapter 1. Each phase produces an output that is used in later phases. Together, these phases operationalise the early-warning framework by defining the target variable, constructing the ex-ante feature set, developing and comparing possible probabilistic models, testing the added value of cross-border information, and evaluating the final probability outputs in terms of explainability, warning behaviour, asymmetric error costs, and decision-support interpretation.

3.2.1. SQ1 – Phase 1: extreme event definition

Which threshold definition and labelling strategy is most appropriate for negative prices, normal prices, and price spikes in the Dutch day-ahead electricity market, taking into account price distributions, seasonality, and sensitivity to alternative definitions?

SQ 1 defines the target variable for the modelling framework. Dutch day-ahead price distributions are analysed across hours and seasons, after which alternative threshold definitions are compared through sensitivity analysis. This phase determines how each hour is classified as negative, normal, or spike, and therefore directly affects both class imbalance and the practical interpretation of model outputs. The output of this phase is a transparent labelling strategy and a labelled hourly dataset for the subsequent modelling phases.

3.2.2. SQ2 – Phase 2: ex-ante feature design

Which domestic, weather, cross-border, and calendar indicators can be operationalised as ex-ante predictors of Dutch day-ahead extreme price events, given the constraint that all features must be available before market closure?

SQ 2 translates the three mechanisms identified in Chapter 2 and the available data sources into an ex-ante feature set. This includes Dutch price-history variables, domestic system indicators, weather forecasts, cross-border market and exchange variables, and calendar controls. Each variable is assessed in terms of publication timing and practical ex-ante availability. The output of this phase is a documented feature overview and the final modelling dataset aligned to day-ahead market closure.

3.2.3. SQ3 – Phase 3: probabilistic model development and comparison

How can candidate probabilistic machine-learning models be evaluated and compared for estimating hourly probabilities of negative prices, normal prices, and price spikes in an ex-ante setting, and which model provides the most useful early-warning performance?

SQ 3 develops the predictive modelling component of the research. Multiple probabilistic ML models are trained and compared in an ex-ante setting. The focus here lies on predictive performance, class imbalance, calibration, and the quality of the predicted probabilities. The comparison is guided by the model-comparison KPIs defined in Section 3.3.3. In line with the implications from Chapter 2, this phase

also considers explainability as part of the model evaluation. This is because the probability outputs should be interpretable enough to support decision-making. The output of this phase is a comparison of candidate models and their performance as probabilistic early-warning models for Dutch extreme price events.

3.2.4. SQ4 – Phase 4: added value of cross-border indicators

To what extent do cross-border coupling, directional scheduled exchange, day-ahead net transfer capacity, and neighbouring market features improve the ex-ante prediction of Dutch negative prices and price spikes compared with models that only use domestic features, and which neighbouring feature groups and bidding zones contribute most?

SQ 4 is operationalised through the feature-group removal analysis. The final modelling framework is compared across a domestic baseline, the core cross-border feature set, and additional cross-border feature groups, including scheduled exchange, transfer capacity, neighbouring-system forecasts, neighbouring lagged prices, and neighbouring weather and calendar variables. The analysis also includes leave-one-neighbour scenarios to assess whether the predictive value is different for specific bidding zones. The resulting performance metrics are used to determine how much cross-border information improves the ex-ante prediction task and which cross-border components contribute most to model performance.

3.2.5. SQ5 – Phase 5: evaluation, error costs, and decision-support interpretation

How should the estimated class probabilities be interpreted as decision-support signals for market participants, considering asymmetric false-positive and false-negative costs?

SQ 5 is operationalised through the warning-behaviour, calibration, probability-quality, explainability, and error-structure analyses of the selected model. Validated warning thresholds are used to examine correct warnings, missed events, and false warnings, while the cost-sensitive framework interprets these outcomes in terms of asymmetric false-positive and false-negative costs. This phase assesses how the predicted probabilities can be used as decision-support signals and what additional actor-specific information would be needed for operational deployment.

3.3. Operational definitions

This section defines the main operational choices used in the modelling part of the research. These definitions translate the conceptual model from Chapter 2 into labels, data constraints, and model-comparison KPIs. Three operational choices are central to the modelling framework: the classification of each hour into one of three price classes, the constraint that all explanatory variables must be available before day-ahead market closure, and the KPIs used to compare the candidate models. These choices are mainly linked to SQ 1, SQ 2, and SQ 3. SQ 1 defines the event labelling strategy, SQ 2 defines which features are available in an ex-ante setting, and SQ 3 defines how candidate models are compared.

3.3.1. Definition of extreme events (labels)

SQ 1 is operationalised by defining how the hourly Dutch day-ahead price is translated into usable labels. The dependent variable is the hourly Dutch day-ahead market-clearing price. Each hour is assigned to one of three classes: negative price, normal price, or price spike. This label is used as the target variable in the probabilistic classification models.

A negative price is defined as an hour in which the Dutch day-ahead price is below 0 €/MWh. This threshold follows directly from the market definition of negative prices and does not require an additional statistical threshold. However, there is no fixed price level above which an observation automatically becomes a spike. This means that an explicit threshold is required.

In this research, a price spike is defined as an upper-tail event relative to the price distribution of that period. For each separate month in each year, the 95th percentile of the Dutch day-ahead price is calculated. An hour is labelled as a price spike when its price is above this month-year-specific threshold. For example, January 2022 and January 2023 receive separate spike thresholds, and February

2022 receives a different threshold from January 2022. This means that the spike definition adapts to changing market conditions and seasonal price patterns.

The use of a percentile-based threshold is in line with variable thresholds used in literature. Liu et al. (2022) distinguish between fixed thresholds and variable thresholds, and use the 95th quantile to identify extremely high prices (Liu et al., 2022). Lu and Suthaharan (2023) add to this by arguing that a single threshold across changing market periods is not appropriate, and use data-dependent thresholds that vary by market and year (Lu & Suthaharan, 2023). The definition of the spike threshold used here combines both of these ideas: it keeps the 95th-percentile interpretation, while allowing the threshold to vary over time.

In the current dataset covering 2022–2025, negative prices account for approximately 4% of all hourly observations. Using the upper 5% of prices as the spike definition creates a similar lower-tail and upper-tail event frequency. An initial full-sample 95th-percentile threshold was considered, but this produced an uneven spike distribution across the years because market price levels differ over time. The month-year threshold was therefore selected to define spikes relative to the price conditions of each period. This does not imply that negative prices and price spikes are symmetric events. However, it supports the use of a 95th-percentile threshold for defining upper-tail price extremes while keeping both extreme classes similar in size.

The labelling strategy is therefore defined as follows:

- **Negative price:** day-ahead price < 0 €/MWh;
- **Price spike:** day-ahead price $>$ 95th percentile of prices in the same month and year;
- **Normal price:** all remaining hours.

All observations are initially labelled as normal prices. Afterwards, the month-year-specific spike threshold is calculated for each separate month and year. Finally, observations are relabelled as negative prices when the day-ahead price is below 0 €/MWh, and as price spikes when the day-ahead price is above the relevant month-year 95th-percentile threshold. Observations that do not meet one of these conditions remain labelled as normal prices. The spike threshold is calculated using the full price distribution within each month-year period, including negative prices.

This labelling strategy has two advantages. First, it ensures that price spikes represent the highest prices within the relevant seasonal and yearly market context, rather than only the highest prices in the full dataset. Second, it prevents periods with structurally high or low prices from dominating the spike definition. As a result, approximately the upper 5% of prices within each month-year period are treated as spike observations, making the target variable more sensitive to changing market regimes.

Because the threshold choice has an impact on class balance and model performance, alternative definitions are included in the sensitivity analysis. The baseline month-year 95th-percentile threshold is compared with alternative threshold choices, such as a fixed full-sample 95th percentile or alternative percentile levels. This tests whether the main conclusions remain the same when the spike definition is changed.

3.3.2. Ex-ante constraint

SQ 2 is operationalised by defining which features are allowed in the model. The model is intended as an early-warning model before day-ahead market closure. Therefore, features may only be included in the main modelling framework if they are available before market closure. Ex-ante availability means that a feature must be known before the day-ahead auction closes at 12:00 CET on the day before delivery.

This constraint influences the entire feature-selection process. Forecast variables are preferred over realised variables, because realised values are generally only known after delivery. Historical variables may be included if they refer to information that would already be available at the prediction moment, such as lagged prices or previous-day system conditions. Variables that are only published after market closure are excluded from this research.

Publication times are verified for each feature where possible. This is done by checking source documentation and the retrieval process itself. For each candidate variable, the feature overview includes the data source and the observed publication time. If the publication time cannot be verified reliably, the variable is excluded from the model and analysis. The ex-ante status of each feature is documented

in the feature overview in Chapter 4. This prevents the model from relying on information that would not have been available in a real early-warning setting and ensures that the empirical design remains consistent with the objective of estimating price-event risks before day-ahead market closure. Manual spot checks were performed during the data retrieval and feature-selection process by retrieving data for upcoming delivery periods before market closure. These checks were used to verify whether the selected data sources could provide the required forecast, capacity, exchange, weather, and calendar variables in time for an ex-ante prediction setting.

3.3.3. Model-comparison KPIs

The model comparison is based on four KPIs that follow from the early-warning objective introduced in Chapters 1 and 2. These KPIs translate the usefulness of the early-warning model into criteria for comparing candidate models:

- **Detection of relevant extreme events.** The model should identify negative prices and price spikes, since these extreme events are the central focus of the early-warning problem. This criterion is evaluated through classification performance, with attention to the rare-event classes rather than only overall accuracy.
- **Limitation of unnecessary warnings.** The model should not achieve high event detection only by producing many false alarms, because this would reduce its usefulness. This criterion focuses on warning behaviour and the trade-off between event detection and false warnings.
- **Useful probability estimates.** The model outputs should provide meaningful hourly risk estimates rather than predict the most likely class label. This criterion concerns probability quality and calibration, which assess whether the predicted probabilities are reliable enough to support interpretation.
- **Support for actor-specific decision-making.** The model should provide outputs that can be interpreted in relation to different market positions, flexibility options, and risk preferences. This criterion is addressed through error-structure and cost-sensitive interpretation, where false positives and false negatives are related to possible actor-specific consequences.

The detailed metric and implementation choices for these KPIs are provided in Chapter 5.

3.4. Validation strategy

The validation strategy is designed to reflect the intended use of the model as closely as possible. Since the goal is to estimate future day-ahead extreme price risks using information available before market closure, random train-test splits are not appropriate. Random splits would mix past and future observations, which could lead to overly optimistic performance estimates. Additionally, temporal patterns from the test period may indirectly appear in the training data.

For this reason, the research uses a time-based validation strategy. The model is trained on earlier years and evaluated on a later period outside this training sample. Observations from 2022–2023 are used as the development training period, 2024 as the validation period, and 2025 as the final test period. After model selection and threshold tuning, the selected models are retrained on the combined 2022–2024 period before final evaluation on the 2025 test period. Differences between validation and final test performance are used to assess temporal robustness under changing market conditions. This setup also reflects the intended operational setting, in which a model trained on historical data is used to estimate price-event probabilities for future hours.

Model comparison is performed using the same split for all models. Differences in performance can therefore be attributed to the model specification and feature set, rather than to different train-test divisions. The final model performance is then reported on the held-out test year.

The evaluation follows the model-comparison criteria defined in Section 3.3.3. It considers both classification performance and probability quality. Classification performance is used to assess whether the model correctly identifies negative prices, normal prices, and price spikes, while probability quality is evaluated because the model outputs are intended as risk signals rather than only class labels. The detailed metrics used for these assessments are defined in Chapter 5.

The validation strategy also includes robustness checks. The sensitivity of the results to the spike definition is tested by comparing the baseline month-year 95th-percentile threshold with alternative threshold definitions. Afterwards, the added value of cross-border features is evaluated by comparing models with and without cross-border variables. Third, the influence of class imbalance is assessed by using class-specific evaluation metrics and model weighting strategies. Performance is inspected across relevant time periods and feature groups as a final check, because negative prices and price spikes may occur under different market conditions. A reversed year-order evaluation is included as an additional robustness check to test how sensitive the conclusions are to the chronological ordering of the available years.

Together, these validation steps evaluate model performance beyond average predictive accuracy. The validation strategy also tests whether the model produces reliable and interpretable early-warning signals for extreme price events.

3.5. Ethics, privacy, and reproducibility

This research uses public and secondary data sources. The analysis does not involve personal data, individual-level behavioural data, or confidential customer information. The ethical risks related to privacy are therefore limited. The main ethical and methodological concerns relate to transparency, reproducibility, and the interpretation of model outputs.

The modelling framework is designed to be reproducible. Data retrieval is performed through publicly accessible Application Programming Interfaces (APIs). The preprocessing, feature engineering, exploratory data analysis, model training, and evaluation steps are implemented in Python and organised into separate scripts or notebooks. This structure makes it straightforward to rerun the pipeline when data sources are updated or when alternative modelling choices are tested.

Reproducibility also includes documenting the ex-ante availability of features. It is important to show that the model does not rely on information that would only have been available after market closure to meet the ex-ante constraint. Publication times and data availability are therefore checked where possible, or documented manually during the data retrieval process. The use of Artificial Intelligence (AI)-based tools during the research and writing process is documented in Appendix G.

The model outputs should only be used as supporting information for decision-making. Predicted probabilities translate into estimated risks under the available feature set and model assumptions. They do not remove uncertainty and should not be interpreted as guaranteed predictions. The decision-oriented part of the research therefore explicitly considers false positives, false negatives, and the asymmetric costs of both errors.

Together, these choices define how the conceptual model from Chapter 2 is translated into a modelling framework. The next chapter applies this design to the data collection and feature-selection process by constructing the target variable and operationalising the domestic, weather, cross-border, and calendar features.

4

Data and feature engineering

This chapter describes the construction of the dataset used for the ex-ante prediction of Dutch day-ahead price classes. Building on the research design in Chapter 3, it explains how the target variable is constructed, which feature groups are included, and how the ex-ante availability constraint is respected during data collection and feature engineering.

The dataset is structured at an hourly resolution. Each observation represents one hour in the Dutch day-ahead market and contains one target label: negative price, normal price, or price spike. The variables are grouped into Dutch price-history variables, domestic system variables, weather variables, cross-border coupling and constraint variables, and calendar controls. Together, these feature groups operationalise the mechanisms identified in Chapter 2: domestic market stress, flexibility stress, and cross-border transmission of scarcity and surplus.

The full variable dictionary, including operationalisation, literature support, expected importance, source, and publication timing, is provided in Appendix C. This chapter summarises the feature groups and explains how they are transformed into the final modelling dataset.

4.1. Data overview

The dataset is developed around the Dutch hourly day-ahead market-clearing price for the Dutch bidding zone (NL). Each observation represents one delivery hour and is linked to one target label: negative price, normal price, or price spike. The target labels are constructed according to the operational definitions introduced in Chapter 3. Negative prices are defined as prices below 0 €/MWh, and price spikes are defined using a month-year-specific 95th-percentile threshold.

The variables are selected to represent the mechanisms identified in the literature review and conceptual model. The final dataset is structured around five feature groups: Dutch price-history variables, domestic system variables, weather variables, cross-border coupling and constraint variables, and calendar controls. Together, these groups cover the main ex-ante information sources before day-ahead market closure: recent market outcomes, expected Dutch supply-demand conditions, weather-dependent system conditions, cross-border exchange and capacity conditions, and recurring calendar patterns. They provide observable indicators of domestic market stress, flexibility stress, and cross-border influences that can contribute to Dutch prices moving into the lower or upper tail of the distribution.

The electricity-market and system data are retrieved from the European Network of Transmission System Operators for Electricity (ENTSO-E) Transparency Platform (ENTSO-E, n.d.). This platform provides open access to European electricity-market data published by Transmission System Operators (TSOs) and other data providers, including prices, load, generation, forecasts, transmission, and cross-border exchange information. It supports electricity-market transparency by making fundamental market information available. The platform is used in this thesis because it provides reproducible access to relevant variables for the Dutch day-ahead market. Forecast series from the platform are treated as external inputs and are used as ex-ante indicators of expected system conditions before day-ahead market closure.

The first feature group consists of Dutch price-history variables. These variables capture volatility and the clustering of extreme price events. Examples include lagged Dutch day-ahead prices and recent counts of negative prices and price spikes.

The second feature group consists of domestic Dutch system variables. These variables capture expected supply-demand conditions in the Dutch bidding zone. Examples include load forecasts, renewable generation forecasts, and other system indicators that are available before market closure.

The third feature group consists of weather variables. These variables are included because electricity demand and renewable generation are weather-dependent, and can therefore affect both lower-tail and upper-tail price outcomes (Loizidis et al., 2025). Because the model is intended to operate before day-ahead market closure, only archived forecast variables are included.

The fourth feature group consists of cross-border coupling and constraint variables. These variables capture neighbouring market conditions and the ability of the Dutch system to import or export electricity. The cross-border scope is limited to directly connected bidding zones: Germany/Luxembourg bidding zone (DE/LU), Belgian bidding zone (BE), Great Britain bidding zone (GB), Denmark bidding zone 1 (DK1), and Norway bidding zone 2 (NO2). For each connected bidding zone, the dataset includes scheduled exchange variables and day-ahead Net Transfer Capacity (NTC) when these data are available. Import variables are defined as scheduled exchange or transfer capacity from the neighbouring bidding zone to the Netherlands. Export variables are defined as scheduled exchange or transfer capacity from the Netherlands to the neighbouring bidding zone. This is applied consistently across all cross-border variables.

The fifth feature group consists of calendar controls. These variables capture recurring temporal patterns in electricity demand, renewable production, and market behaviour. Examples include hour-of-day, day-of-week, weekend, month, season, and public-holiday indicators.

All variables are aligned to an hourly resolution and matched to the corresponding Dutch day-ahead delivery hour. The dataset covers the years 2022–2025. Data from 2022–2024 are used for training and validation, while observations from 2025 are used as the final out-of-sample test period, as defined in the validation strategy in Chapter 3.

Table 4.1 summarises the feature groups used in the modelling framework. The full variable dictionary, including operationalisation, unit, literature support, expected importance, data source, and ex-ante availability, is provided in Appendix C.

Table 4.1: Overview of feature groups used in the modelling framework.

Feature group	Role in the model	Example variables	Main source type
Target variable	Defines the hourly price class to be predicted.	Dutch day-ahead price, price category.	ENTSO-E
Dutch price-history variables	Capture price persistence, volatility, and recent extreme-event behaviour.	Lagged Dutch prices, recent negative-price and spike counts, indicators of negative prices and spikes the same hour one day before.	ENTSO-E
Domestic system variables	Represent Dutch supply-demand balance and flexibility stress.	Load forecasts, wind and solar forecasts, residual load, system indicators.	ENTSO-E
Weather variables	Capture weather-dependent demand and renewable-generation conditions.	Historical forecast of temperature, irradiance, sunshine duration, wind speed.	Open-Meteo
Cross-border variables	Represent neighbouring market pressure and cross-border transmission of scarcity or surplus.	Neighbouring lagged prices, directional scheduled imports and exports, directional day-ahead net transfer capacity, total Dutch import and export capacity, neighbouring load and renewable forecasts.	ENTSO-E
Calendar controls	Capture recurring temporal and seasonal patterns.	Hour, weekday, weekend, month, season, public holidays.	Python library / Nager public holiday API

4.2. Target variable construction

The target variable is constructed from the hourly Dutch day-ahead market-clearing price. Each delivery hour is assigned one of three labels: negative price, normal price, or price spike. This label is used as the dependent variable in the probabilistic classification models.

The labelling procedure is in line with the operational definitions discussed in Chapter 3. First, all observations are initially labelled as normal prices. Second, for each separate month in each year, the 95th percentile of the Dutch day-ahead price is calculated. This creates a specific spike threshold for each month and year. Third, observations are relabelled as negative prices when the day-ahead price is below 0 €/MWh, and as price spikes when the price is above the varying 95th-percentile threshold. Observations that do not meet one of these conditions remain labelled as normal prices.

Exactly zero prices require separate attention, since they lie on the boundary between normal and negative prices. In the main target definition, exactly zero prices are retained in the normal class because the operational definition of negative prices is restricted to prices below 0 €/MWh. However, zero prices can still reflect surplus-system conditions that are closer to negative-price events than to regular market prices. Zero and low positive prices are therefore reported separately in Table 4.2. Alternative lower-tail definitions, as well as upper-tail definitions, are reserved for the sensitivity analysis in Chapter 5. The sensitivity analysis tests whether the main modelling conclusions are robust to changes in the definition of both price spikes and low-price events.

Table 4.2: Zero and low-price observations in the final modelling dataset.

Price range	Count	Share (%)
Price < 0	1,429	4.15
Price $= 0$	228	0.66
$0 <$ price ≤ 5	676	1.96
$5 <$ price ≤ 10	348	1.01
Price > 10	31,772	92.22

Table 4.2 shows that zero prices account for only 0.66% of the final modelling dataset. This low share supports keeping the strict negative-price definition in the main model, while still motivating its inclusion in the later sensitivity analysis together with alternative low-price thresholds.

This creates a varying spike threshold in which each price is compared with the upper tail of its own month-year period. The target label therefore accounts for seasonal and yearly differences in the Dutch price distribution, rather than applying one fixed threshold to the full dataset.

The resulting target variable is used throughout the modelling framework. Negative prices represent the lower end of the price distribution, price spikes represent the upper end, and normal prices represent all remaining observations. The distribution of these labels is reported in the exploratory data analysis in Section 4.7. The number of observations in each class determines the degree of class imbalance and affects model evaluation.

4.3. Dutch price-history and domestic system features

The Dutch feature group captures domestic market stress and flexibility stress within the Dutch bidding zone. It consists of two types of variables: price-history variables and domestic system variables. Price-history variables capture persistence, volatility, and the clustering of extreme price events. Domestic system variables capture expected supply-demand conditions before day-ahead market closure. This feature group represents the Dutch part of the electricity system in the conceptual model, where recent market conditions and expected domestic balance conditions signal whether the Dutch bidding zone is moving towards surplus or scarcity.

Price-history variables are included because recent price behaviour can contain information about future extreme-price risk. This includes lagged Dutch day-ahead prices, recent counts of negative prices and price spikes, and time-since-event indicators. These variables are especially relevant because extreme price events may be clustered. They also provide a baseline to assess the added value of domestic system and cross-border variables.

Domestic system variables describe the expected Dutch electricity balance before delivery. This includes load forecasts, renewable generation forecasts, and other system indicators that are available

before market closure. The forecast variables are not produced by the modelling pipeline itself, but retrieved as archived forecast series. They are matched to the relevant delivery hour, so that the feature value represents information that would have been available before day-ahead market closure rather than realised system outcomes. These variables operationalise the domestic supply-demand mechanism identified in Chapter 2.

Residual load is treated as an important derived indicator because it combines demand forecast and variable renewable generation forecast into one measure of expected system pressure. Residual load is calculated as expected load minus expected wind and solar generation. A high residual load indicates that more demand must be met by non-renewable generation, imports, storage, or demand response. A low or negative residual load indicates potential surplus conditions and may increase the probability of negative prices. Residual load overlaps with the load, wind, and solar forecast variables from which it is calculated. This overlap is accepted because residual load represents the combined domestic balance condition, while the separate forecast variables retain information about which component drives that balance.

$$RL_t = L_t^{forecast} - W_{t,onshore}^{forecast} - W_{t,offshore}^{forecast} - S_t^{forecast}$$

where RL_t is residual load in hour t, $L_t^{forecast}$ is the Dutch load forecast, $W_{t,onshore}^{forecast}$ is the Dutch onshore wind generation forecast, $W_{t,offshore}^{forecast}$ is the Dutch offshore wind generation forecast, and $S_t^{forecast}$ is the Dutch solar generation forecast (Ruggles & Caldeira, 2022).

All Dutch price-history and domestic system variables are checked against the ex-ante constraint. Historical price variables are only used for prices that are already known before prediction. Forecast variables are used instead of realised variables, because realised load and renewable generation are only known after delivery. Candidate variables for which the publication time cannot be verified are excluded from the dataset. The full list of Dutch price-history and domestic system variables, including operationalisation, source, and publication timing, is provided in Appendix C.

4.4. Weather features

Weather features are included because weather conditions can affect both electricity demand and renewable electricity generation. They fit the conceptual model because weather conditions are an ex-ante observable source of variation in both demand-side pressure and renewable supply conditions. Temperature can influence load demand, which in turn affects electricity prices. Weather also affects the supply side through wind and solar generation, because high renewable production can contribute to negative prices and low renewable production can contribute to price spikes (Amjady & Keynia, 2011; Loizidis et al., 2025).

In this thesis, weather variables are included as supporting predictors of both demand and supply conditions. Temperature is used to capture possible demand effects, while wind-speed and solar-related variables are used to capture renewable-generation conditions. Load forecasts, wind generation forecasts, and solar generation forecasts already reflect part of the expected weather effect. The separate weather variables are therefore partly overlapping with these system forecasts. This overlap is accepted because the system forecasts describe expected electricity-system outcomes, while the weather variables capture the underlying weather conditions that may contain additional information for extreme-price prediction.

Weather variables are only used when they are available before day-ahead market closure. Archived weather-forecast variables are retrieved from the Open-Meteo API (Open-Meteo, n.d.). These variables are used instead of realised weather observations. Realised weather data would only be known after the delivery hour and is therefore not available in a real early-warning setting. Each archived weather-forecast value is matched to the delivery hour for which the forecast is valid, and only forecast values that satisfy the publication-timing constraint are used.

All weather variables are aligned to the hourly delivery structure of the Dutch day-ahead market. The weather data are available at a different resolution, so they are interpolated to match the hourly modelling dataset. The final weather variables, including their operationalisation, source, and ex-ante availability status, are documented in Appendix C.

4.5. Cross-border coupling and constraint features

Cross-border coupling and constraint features are included to capture the influence of connected electricity markets on the Dutch day-ahead price. As discussed in Chapter 2, the Dutch bidding zone is connected to neighbouring markets through interconnectors and coupled market-clearing procedures. Cross-border variables therefore operationalise the mechanism of cross-border transmission of scarcity and surplus. Within the electricity system, these variables indicate whether Dutch surplus or scarcity conditions can be absorbed through connected bidding zones, or whether stress in neighbouring zones may affect the Dutch market.

The scope is limited to directly connected bidding zones. These are Germany/Luxembourg, Belgium, Great Britain, Denmark (DK1), and Norway (NO2). This selection keeps the feature set focused on direct cross-border connections (TenneT, n.d.-c).

The cross-border feature group contains three types of variables. The first type consists of neighbouring market variables, such as lagged day-ahead prices, load forecasts, renewable generation forecasts, and residual-load forecasts. Neighbouring load, renewable-generation, and residual-load forecast variables are constructed in the same way as the Dutch domestic system variables, but for the connected bidding zones. These variables capture whether connected systems are likely to experience scarcity or surplus conditions. The second type consists of scheduled exchange variables, such as scheduled imports and exports between the Netherlands and connected bidding zones. These variables indicate whether electricity is planned to flow in or out of the Dutch market during a specific delivery hour. The third type consists of day-ahead NTC variables. These variables represent the cross-border capacity made available for day-ahead exchange between the Netherlands and the directly connected bidding zones for which these data are available.

All cross-border exchange and capacity variables are defined from the Dutch perspective. Scheduled imports represent planned exchanges from a neighbouring bidding zone into the Netherlands. Scheduled exports, on the other hand, represent planned exchanges from the Netherlands to a neighbouring bidding zone. In addition to border-specific capacity variables, aggregate capacity indicators are calculated by summing the directional day-ahead NTC across all directly connected bidding zones. Total import capacity represents the total available cross-border import capacity into the Dutch bidding zone, while total export capacity represents the total available cross-border export capacity from the Dutch bidding zone. These variables indicate whether the Netherlands has cross-border room to absorb scarcity through imports or surplus through exports.

$$C_{NL,t}^{import} = \sum_{z \in Z} C_{z \to NL,t} \tag{4.1}$$

$$C_{NL,t}^{export} = \sum_{z \in Z} C_{NL \to z,t} \tag{4.2}$$

where $C_{NL,t}^{import}$ is the total day-ahead import capacity into the Netherlands in hour t, $C_{NL,t}^{export}$ is the total day-ahead export capacity from the Netherlands in hour t, and $C_{z \to NL,t}$ and $C_{NL \to z,t}$ are the day-ahead NTCs between the Netherlands and neighbouring bidding zone z. GB is included in the broader cross-border feature set where data are available, but not in the day-ahead transfer-capacity variables. The set Z contains the directly connected bidding zones included in this thesis:

$$Z = \{\text{Germany/Luxembourg}, \text{Belgium}, \text{Denmark 1}, \text{Norway 2}\}. \tag{4.3}$$

These variables are included because cross-border exchange is linked to market stress. Surplus electricity can be exported and scarcity can be reduced through imports if sufficient transfer capacity is available. When capacity is limited or neighbouring markets experience similar stress, cross-border coupling may instead transmit scarcity or surplus conditions into the Dutch bidding zone. This makes cross-border variables relevant for predicting both negative prices and price spikes (De Blauwe et al., 2025; Ovaere et al., 2023; Unger et al., 2018).

The inclusion of cross-border features is also used to answer SQ 4. The modelling framework compares models with and without cross-border variables. This tests whether neighbouring-market information improves the prediction of Dutch negative prices and price spikes. Therefore, cross-border variables are not assumed to improve prediction in advance. Their added value is evaluated through the model-comparison and removal strategy.

All cross-border features follow the ex-ante availability constraint. The detailed list of cross-border variables, including operationalisation, source, and ex-ante availability status, is provided in Appendix C.

4.6. Calendar controls

Calendar controls are included to account for recurring patterns in electricity demand and market behaviour. The feature group includes hour of the day, day of the week, weekend, month, season, and public-holiday indicators. These variables help control for variation between hours, weekdays, weekends, seasons, and holidays. They fit the conceptual model because part of electricity demand, renewable production, and market behaviour follows temporal patterns that are known before delivery. Calendar controls are not treated as a main explanatory feature group, but as supporting variables that help separate recurring time patterns from domestic, weather, and cross-border market-stress signals. The full list of calendar controls is provided in Appendix C.

4.7. Exploratory data analysis and preprocessing

Before modelling, an exploratory analysis is performed to assess data quality and verify the structure, target-label construction, missing-value pattern, and overall suitability of the dataset. This section first describes the retrieval and cleaning checks, after which the target distribution, labelled price series, possible time patterns, and feature diagnostics are discussed.

The dataset is merged based on hourly Coordinated Universal Time (UTC) timestamps. Dutch local time was retained separately for calendar-feature construction and then aligned with the UTC-based dataset. This avoids ambiguity during daylight-saving transitions, where local Dutch days can contain 23 or 25 hours. Timestamp checks showed that all retained observations were aligned to whole hourly timestamps and that no duplicate UTC timestamps were present.

The full 2022–2025 period contains 35,064 expected hourly timestamps. The Dutch day-ahead price series contained 35,061 observed timestamps, with three missing target-price hours. These three missing target-price hours are gaps in the source data. Since the Dutch day-ahead price is required to construct the target label, these hours are not included in the analysis. They correspond to less than 0.01% of the expected hourly timestamps, so removing them is not expected to affect the results.

The target labels are constructed from the Dutch day-ahead price. Negative prices are defined as prices below 0 €/MWh, while price spikes are defined using the month-year-specific 95th percentile. This threshold is calculated after retrieving the Dutch day-ahead prices and used to construct the target label. It is not included as a feature in the final dataset.

4.7.1. Target-label distribution

The final complete-case modelling dataset contains 34,453 hourly observations. Table 4.3 shows the distribution of the three target classes after missing-value removal. Normal prices dominate the dataset, with 31,286 observations, corresponding to 90.81% of all hours. Negative prices occur in 1,429 hours, corresponding to 4.15% of the dataset, while price spikes occur in 1,738 hours, corresponding to 5.04%.

Table 4.3: Target-label distribution in the final modelling dataset.

Class	**Label**	**Count**	Share (%)
0	Negative price	1,429	4.15
1	Normal price	31,286	90.81
2	Price spike	1,738	5.04
Total		34,453	100.00

The distribution confirms that the modelling task is strongly imbalanced. The normal class is much larger than both extreme-price classes, while negative prices and price spikes each represent only a small share of the observations. This has direct consequences for model evaluation. Overall accuracy is not sufficient, because a model could achieve high accuracy while failing to detect extreme-price events. Therefore, the modelling chapter evaluates class-specific performance, including precision, recall, F_2-score, balanced accuracy, and confusion matrices.

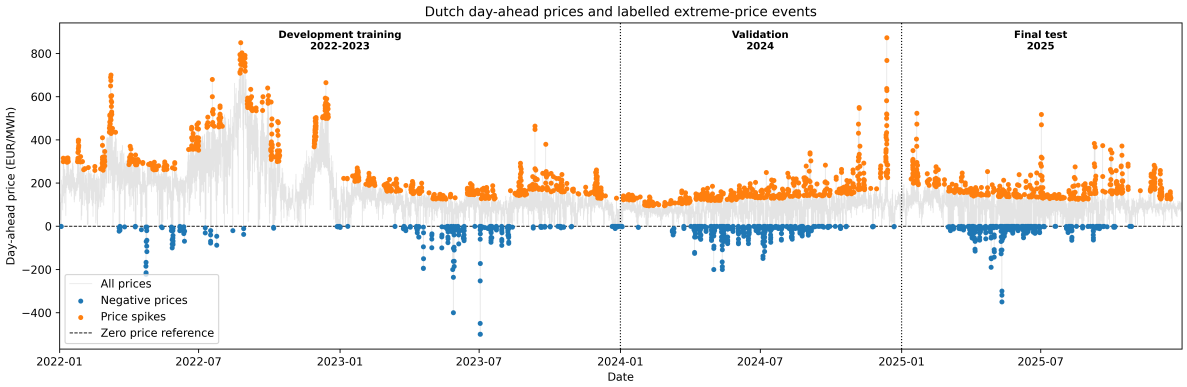

Figure 4.1: Dutch day-ahead prices and labelled extreme-price events across the development training, validation, and final test periods.

Figure 4.1 shows the labelled Dutch day-ahead price series across the development training, validation, and final test periods. The figure confirms that both negative prices and price spikes occur throughout the dataset, but not evenly across time. The development period contains the strongest price spikes, especially during the high-price market conditions in 2022. Negative prices become more frequent after 2022 and remain visible in both the validation and final test periods.

The figure also suggests that extreme events are temporally clustered rather than randomly distributed. Some seasonal structure is already visible, especially for negative-price periods. Detailed hourly and monthly patterns are assessed separately in Figure 4.2. This supports the use of a chronological validation strategy, because the model must be evaluated on later market conditions that differ from the development period.

4.7.2. Time patterns and class imbalance

Figure 4.2 is used to inspect whether the extreme target classes are evenly distributed across hours and months. This is relevant because negative prices and price spikes may occur under different temporal market conditions, and because the spike class is defined using a month-year-specific threshold. The figure reports shares within each hour and month rather than absolute counts. The normal class is left out to make the two extreme-price classes visible.

The hourly distribution shows a clear intraday pattern. Negative prices are concentrated around the middle of the day, especially between late morning and early afternoon. These are the hours in which solar generation is typically high, because solar irradiance is highest around midday. This supports the interpretation that many negative-price hours are related to surplus conditions during periods with relatively high renewable generation. Price spikes follow a different pattern and occur more often during morning and evening hours, when demand is higher or system flexibility may be more limited.

The monthly distribution shows that negative prices also vary across the year. They are most visible in the spring and summer months, when solar output is relatively high and midday surplus conditions occur more frequently. The monthly spike shares are more stable because the spike label is defined using a month-year-specific 95th-percentile threshold. This assigns the upper 5% of prices within each month-year period to the spike class. Seasonal conditions, such as renewable generation patterns and demand levels, still affect the probability of observing negative prices or price spikes. The observed patterns therefore support the inclusion of hour-of-day and month variables as calendar controls in the modelling dataset.

Overall, the figure shows that class imbalance is not only a general frequency problem, but also varies over time. This supports the evaluation of model performance with class-specific metrics rather than relying on overall accuracy alone, which will be further discussed in Chapter 5.

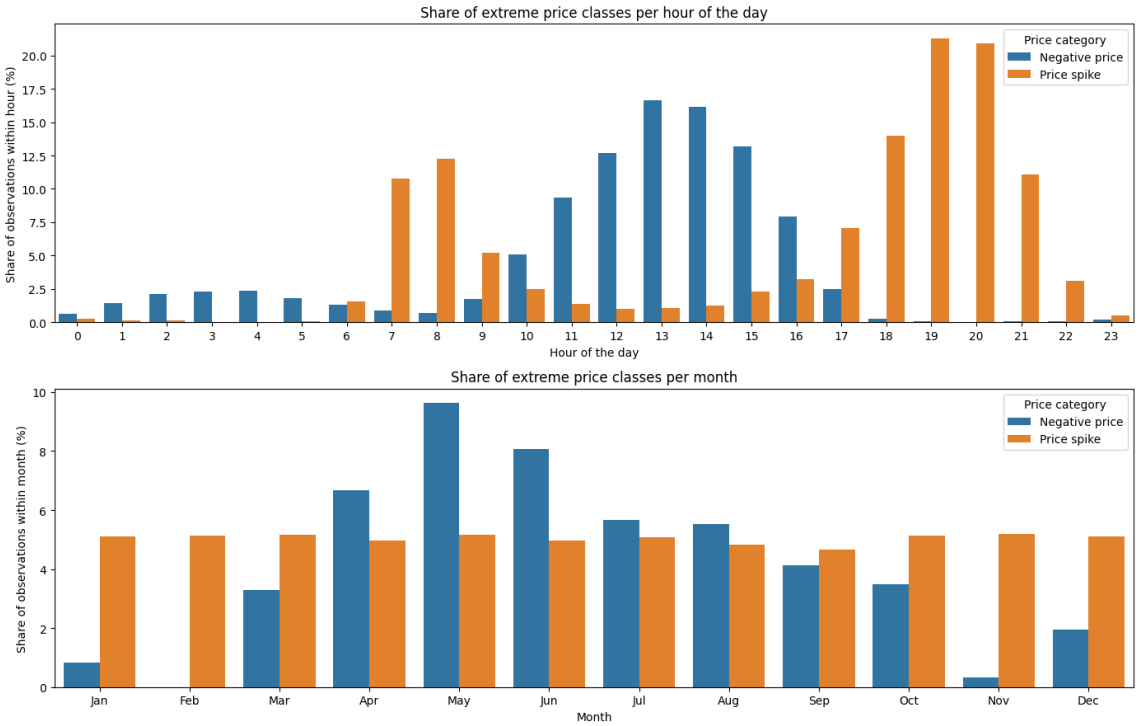

Figure 4.2: Share of negative prices and price spikes across hour of the day and month.

4.7.3. Missing values and data coverage

After merging the target and feature data, the dataset contained 35,061 hourly observations. Of these observations, 608 rows contained at least one missing feature value. This corresponds to approximately 1.73% of the merged dataset. The strategy for missing values is not based on the percentage alone, since the proportion of missing data should not be the only basis for choosing a missing-data strategy (Madley-Dowd et al., 2019). Instead, missing data is evaluated in relation to the target variable and the temporal pattern of the missing rows. This is in line with the recommendation to consider missing-data patterns when choosing a strategy for missing data (Hughes et al., 2019).

A closer inspection of the rows with missing values indicated that these were concentrated in a limited number of forecast and exchange variables, rather than being distributed evenly across all variables. The largest missing blocks occurred in a small number of neighbouring forecast series, while many other variables contained only a few missing observations. Missing values in some residual-load variables were caused by missing observations in one or more of the load, wind, or solar forecasts used in the calculations.

Figure 4.3 compares the price category distribution of the missing rows with the overall target distribution. The distributions are broadly similar: normal prices dominate both groups, while price spikes are slightly overrepresented and negative prices slightly under-represented among the missing values. Missing data was also checked by year. The highest missing share occurred in 2022, but the share remained limited in all years, ranging from 0.85% to 3.57%.

These checks do not prove that the missing values are completely random. However, they do suggest that removing rows with missing feature values is unlikely to result in strong bias in the target-class balance, because the missing rows have a broadly similar target distribution. Rows with remaining missing values were therefore removed before modelling. This results in a final modelling dataset with 34,453 hourly observations, 102 explanatory features, no missing feature values, and no missing target values.

4.7.4. Feature-group differences across target classes

The explanatory feature groups also differ across negative-price, normal-price, and spike hours. This diagnostic links the feature set to the mechanisms that may differ between the lower and upper price tails.



Figure 4.3: Comparison of the target-class distribution for missing values and the complete dataset

The analysis is summarised at feature-group level. Table 4.4 reports the average absolute standardised median difference for each feature group. Negative and spike hours are compared separately to normal hours. Standardising the median differences makes variables with different units comparable. The table also reports the variable with the largest class difference within each feature group. If the reported largest-variable differences are positive, the median value is higher in the extreme-price class than in the normal class. Negative values, however, indicate that it is lower.

Table 4.4: Feature-group differences across target classes.

Feature group	Variables	Mean $\lvert\Delta_{neg}\rvert$	Mean $\lvert\Delta_{spike}\rvert$	Largest neg.–normal variable	Δ_{neg}	Largest spike–normal variable	Δ_{spike}
NL price variables	10	0.56	0.31	Negative count last week	1.33	Spike count day before	0.91
NL domestic variables	15	0.82	0.39	NL solar forecast	2.18	NL residual load forecast	0.82
DE/LU variables	11	0.75	0.42	DE/LU solar forecast	2.71	DE/LU residual load forecast	1.36
BE variables	11	0.69	0.37	BE solar forecast	2.64	BE residual load forecast	1.19
NO2 variables	10	0.80	0.26	NL scheduled export to NO2	2.60	NO2 generation forecast	0.75
DK1 variables	10	0.73	0.29	NL scheduled export to DK1	2.32	DK1 residual load forecast	0.94
Weather variables	24	1.29	0.12	GB sunshine duration	2.50	NL wind speed at 100m	-0.61
Calendar controls	11	0.16	0.18	Day of the week	1.00	Hour of the day	1.01

Note: Mean differences are average absolute standardised median differences relative to the normal class. Largest-variable differences are signed.

The table shows that several feature groups differ more strongly between negative-price and normal-price hours than between spike and normal-price hours. This is especially visible for the weather, Dutch domestic, and neighbouring-market feature groups. For negative prices, the largest differences are often linked to solar generation, sunshine duration, and scheduled exports. This is consistent with surplus-related conditions. For price spikes, the largest differences are more often linked to residual-load indicators, recent spike occurrence, and calendar timing, which is consistent with scarcity or tight-system conditions.

The table is interpreted as an exploratory check rather than as evidence of predictive value. It shows that the selected feature groups contain variables that differ between normal and extreme-price hours, while predictive value is evaluated later through model performance and explainability methods.

Extreme values are retained in the modelling dataset. Extreme observations can contain information about the market conditions that lead to negative prices or price spikes. Removing them automatically would risk removing the conditions that the early-warning model is intended to detect.

4.7.5. Correlation and redundancy checks

The final dataset contains 102 explanatory variables, several of which describe related market, weather, or cross-border conditions. Correlation and Principal Component Analysis (PCA) checks are used to inspect redundancy before modelling.

Table 4.5 reports the strongest correlations in the final dataset. The strongest correlations are expected given the electricity-market context. Lagged Dutch and neighbouring prices reflect market coupling, correlated temperatures reflect spatially related weather conditions, and some capacity and wind variables describe related system conditions.

Table 4.5: Strongest highly correlated feature pairs in the final dataset.

Variable 1	Variable 2	$\lvert r \rvert$
1-day lagged NL price	BE lagged day-ahead price	0.97
DE/LU lagged day-ahead price	DK_1 lagged day-ahead price	0.97
NL temperature	BE temperature	0.97
DE/LU lagged day-ahead price	BE lagged day-ahead price	0.96
1-day lagged NL price	DE/LU lagged day-ahead price	0.96
NL import capacity from NO_2	NL export capacity to NO_2	0.95
NL wind offshore forecast	NL wind onshore forecast	0.95
1-day lagged NL price	DK_1 lagged day-ahead price	0.93
NO temperature	DK temperature	0.93
BE lagged day-ahead price	DK_1 lagged day-ahead price	0.93

These correlations indicate redundancy, but they do not justify removing variables before modelling. Correlated variables may still contain useful information for non-linear models. Therefore, no variables are removed solely because of high correlation.

PCA is used as an additional diagnostic within feature groups. PCA transforms correlated variables into uncorrelated components that explain as much variance in the original variables as possible (Jolliffe & Cadima, 2016). This makes PCA useful for inspecting whether a feature group can be summarised by a smaller number of dimensions, especially when several variables describe related market, weather, or cross-border conditions. PCA is not used as a modelling transformation, because it is unsupervised. Additionally, it does not take predictive importance into account for negative prices or spikes.

Table 4.6: PCA summary by feature group.

Feature group	Vars.	PC1 var.	PC1–PC2 var.	PCs 80%	PCs 90%	PCs 95%
NL price variables	10	0.40	0.56	5	6	8
NL domestic variables	15	0.26	0.44	6	8	9
DE/LU variables	11	0.26	0.44	6	7	8
BE variables	11	0.23	0.42	6	7	8
NO_2 variables	10	0.37	0.57	4	5	6
DK_1 variables	10	0.30	0.50	5	6	7
Weather variables	24	0.42	0.55	7	11	14
Calendar controls	11	0.45	0.55	5	7	8

Table 4.6 shows that the first principal component explains between 23% and 45% of the variance across the feature groups. Most groups require multiple components to explain 90% of the variance, such as eight components for the domestic system variables and eleven for the weather variables. This indicates that the feature groups contain redundancy, but cannot be reduced to one or two dimensions without losing information.

The original variables are retained for the modelling stage. The correlation and PCA results are treated as evidence of redundancy, not as a reason to transform the feature set before modelling. As a result, the predictive value of the original variables can first be evaluated without dimensionality reduction. Whether feature reduction is useful is assessed later through model performance comparisons, feature-removal experiments, and explainability results.

4.7.6. Final preprocessing choices

The exploratory analysis results in a modelling-ready dataset. Observations without a Dutch day-ahead price are excluded because the price is required to construct the target label. After inspecting the missing-data pattern, rows with remaining missing feature values are also removed. The final dataset contains 34,453 hourly observations, 102 explanatory features, no missing feature values, and no missing target values.

The day-ahead price used for target construction is kept separate from the explanatory feature set. Historical price information enters the model only through lagged prices and lagged event indicators. Exactly zero prices remain in the normal class under the main target definition, while alternative lower-tail definitions are reserved for the sensitivity analysis in Chapter 5.

The feature diagnostics do not lead to automatic feature removal. Although the correlation and PCA checks show signs of redundancy, this is expected in market, weather, and cross-border data. The original 102 variables are therefore kept for the first modelling stage. Feature reduction is postponed until evidence is available through model performance comparisons, feature-removal experiments, and explainability results.

4.8. Conclusion: ex-ante feature set

This chapter constructed the dataset used for the ex-ante prediction of Dutch day-ahead price classes. The target variable was defined as a three-class label, while the explanatory variables were separated into Dutch price-history variables, domestic system variables, weather variables, cross-border variables, and calendar controls. The detailed overview of each variable, including its operationalisation, source, and ex-ante availability, is documented in Appendix C.

The final dataset contains 34,453 hourly observations and 102 explanatory features. The exploratory analysis assessed data quality through timestamp checks, missing-value inspection, target-label validation, feature-group comparison, correlation analysis, and PCA. These checks support the use of the dataset for the modelling stage.

Together, Chapter 4 answers the data-preparation part of SQ 2 by translating domestic, weather, cross-border, and calendar indicators into a validated ex-ante feature set. The modelling and results chapters evaluate which variables are most predictive and whether feature reduction is warranted. This is done through model performance comparisons and explainability results.

5

Modelling and evaluation framework

This chapter describes the modelling and evaluation framework used to develop an ex-ante early-warning model for extreme Dutch day-ahead electricity prices. Building on the labelled dataset and ex-ante feature set constructed in Chapter 4, the model is formulated as a multi-class probabilistic classification problem. The chapter first defines the prediction task, then explains the train-validation-test design, and finally introduces the candidate models and their modelling pipelines. The following sections explain how the models are tuned, how warning thresholds are selected, and how performance is evaluated across classification, probability-quality, calibration, and warning-behaviour metrics. The chapter then describes the explainability approach, the target-definition sensitivity analysis, the cross-border removal analysis, and the cost-sensitive interpretation of model errors from a decision-support perspective.

5.1. Modelling problem and prediction

The model in this thesis should predict whether a Dutch day-ahead electricity price for a given delivery hour will be negative, normal, or extremely high. Each hourly observation is therefore assigned to one of three target classes: *negative price*, *normal price*, or *price spike*. The target definition and construction from Chapter 4 are used throughout the modelling phase. Negative prices are defined as prices below 0 €/MWh, while price spikes are defined as prices above the month-year 95th percentile threshold. All remaining observations are classified as normal prices. Alternative definitions are addressed in the sensitivity analysis in Section 5.10.

The prediction task is formulated as a multi-class classification problem. For every delivery hour, each model estimates three class probabilities:

$$P(y_t = \text{negative}), \quad P(y_t = \text{normal}), \quad P(y_t = \text{spike})$$

where y_t is the price category for hour t. These probabilities are used instead of only hard class labels. The practical aim of the model is not just to classify past observations, but to provide early-warning information before market outcomes are known. A predicted probability can be interpreted as a risk signal: a higher probability of a negative price or price spike indicates that the hour is more likely to require attention from market participants.

The Dutch day-ahead price is used only to construct the target label. It is not used as an explanatory feature, because this would violate the ex-ante forecasting constraint. However, lagged Dutch price information is included to capture price persistence and spike clustering (Lu & Suthaharan, 2023).

This formulation has two implications for the model design. First, the dataset is imbalanced because negative prices and price spikes are less frequent than normal prices. A model that mainly predicts the normal class can therefore achieve high overall accuracy, while having little value as an early-warning model. Second, the probabilities themselves matter. The model should not only identify extreme events after applying a classification threshold, but also be able to rank hours according to their estimated risk. The evaluation therefore combines class-specific prediction performance, probability-quality metrics, warning-threshold behaviour, and explainability analysis to assess whether the models are useful as early-warning tools rather than only as classifiers.

5.2. Train-validation-test design

A time-based validation strategy is used to recreate the intended forecasting setting. The model is trained on historical observations and evaluated on later periods, rather than on randomly selected observations from the full dataset. Random train-test splits are avoided because they would allow observations from the same market regime, season, or neighbouring time period to appear in both training and test data. This would overstate the model's performance in a setting where the real objective is to forecast future market conditions from past information.

The final modelling design separates the data into three chronological periods. The years 2022 and 2023 are used as the development training period. The year 2024 is used as the validation period for model selection, hyperparameter tuning, and warning-threshold selection. The year 2025 is kept as the final out-of-sample test period. After model selection and threshold tuning have been completed, the selected models are retrained on the combined 2022–2024 period and evaluated once on the 2025 test period.

Table 5.1: Time-based data split used for model development and final evaluation

Period	Role	Purpose
2022–2023	Development training	Initial model fitting and hyperparameter search
2024	Validation	Model selection and warning-threshold tuning
2022–2024	Final training	Retraining selected models before final testing
2025	Test	Final out-of-sample evaluation

This design separates model development from final evaluation. The validation year is used to make modelling decisions, including the selection of hyperparameters and probability thresholds. These choices are described in Sections 5.4 and 5.5. The test year is not used for these choices. This is important because the final 2025 performance should represent an out-of-sample estimate of how the early-warning model behaves when applied to unseen future market conditions.

The split also supports the practical interpretation of the results. Electricity prices are affected by changing renewable generation patterns, interconnector conditions, demand developments, and policy or market-design changes. These conditions may differ across years. A model that performs well on the 2025 test year is therefore more useful than a model that performs well under a random split of historical observations.

A reversed year-order setup is used as a robustness check. The roles of the years are reversed, so that observations from 2024–2025 are used as the training period, 2023 is used for validation, and 2022 is used for final testing. After model selection and threshold tuning, the final reverse-year models are refitted on the combined 2023–2025 period before evaluation on 2022. This assesses whether the model ranking and main conclusions depend on the choice of years used for development and evaluation. The reversed year-order robustness check is implemented as a full rerun of the modelling pipeline. This means that hyperparameter selection, warning-threshold tuning, final refitting, and test-set evaluation are repeated within the reversed split.

5.3. Candidate models and model pipelines

The literature on electricity price forecasting does not identify one model class as clearly superior for predicting extreme day-ahead prices. Existing studies use different methodological approaches, depending on the market, target definition, available features, and evaluation objective (Haben et al., 2021; Loizidis et al., 2024). Model performance may depend on how well different model structures handle rare events, non-linear relations, and probability estimation. Several candidate models are therefore compared before selecting the most suitable model structure for the final evaluation.

The model set is chosen to compare model families that are relevant for probabilistic classification and differ in complexity and interpretability. A dummy classifier is included as a naive benchmark. Logistic regression is included as an interpretable probabilistic baseline and is supported by its use in extreme-price occurrence forecasting (Liu et al., 2022). This provides a linear and transparent reference point for evaluating more complex models. Gradient Boosting Decision Trees (GBDT) and XGBoost are included as tree-based models that can capture non-linear relations and feature inter-

actions between domestic, weather, and cross-border variables. XGBoost is also included in recent negative-price forecasting work as a benchmark model (Loizidis et al., 2025).

The Multilayer Perceptron (MLP) and PNN are included as neural-network-based alternatives. The MLP provides a neural-network benchmark. The PNN is included as a probabilistic reference model, since PNN approaches have been used in both price-spike and negative-price forecasting (Amjady & Keynia, 2011; Loizidis et al., 2025). Other model classes used in the literature, such as statistical time-series models, deep-learning models, and hybrid forecasting approaches, are not included in the main comparison (Haben et al., 2021; Loizidis et al., 2024, 2025). These model classes either focus primarily on price-level forecasting or add substantial complexity beyond the focused comparison of probabilistic classifiers.

Because the candidate models differ in their assumptions and preprocessing requirements, model-specific pipelines are used. Scaling is applied to logistic regression, the MLP, and the PNN, because these models are sensitive to the magnitude of input variables. Tree-based models are used without scaling, because they split variables according to threshold values. Other preprocessing choices are kept consistent across models where possible, so that performance differences mainly reflect model structure rather than data-preparation differences. PCA is therefore not included as a standard preprocessing step, because the main comparison first evaluates the original ex-ante feature set across all candidate models.

Table 5.2: Candidate models and their role in the modelling framework

Model	Model type	Role
Dummy classifier	Benchmark	Reference
Logistic regression	Linear classifier	Interpretable baseline
GBDT	Boosted trees	Tree-based candidate
XGBoost	Boosted trees	More efficient tree-based candidate
MLP	Neural network	Neural-network benchmark
PNN	Probabilistic neighbourhood model	PNN reference

The detailed model descriptions, fixed implementation settings, hyperparameter grids, and final selected settings are reported in Appendix D. The main text therefore focuses on the comparison logic, while the appendix provides the technical details needed for reproducibility.

5.4. Hyperparameter tuning

After defining the candidate models, hyperparameter tuning is used to select suitable model settings before the final test evaluation. Hyperparameters are model settings that are selected before fitting the model, rather than estimated directly from the training data. Examples include the number of trees in a boosting model, the regularisation strength of logistic regression, the hidden-layer structure of an MLP, and the bandwidth of the PNN model. These settings influence the trade-off between underfitting and overfitting, and are therefore selected through the tuning procedure described below.

The tuning procedure follows the time-based validation logic introduced in Section 5.2. Hyperparameter search is performed before the final 2025 evaluation. Earlier observations are used to train the model and later observations are used for validation within each fold. This is more appropriate than random cross-validation, because the model is intended to learn from past market conditions and estimate future price-event probabilities.

The scoring metric used during grid search is macro F_2. This choice follows from the early-warning objective of the thesis. Negative prices and price spikes are rare compared to normal prices, so the tuning metric should not be dominated by the majority class. Macro averaging gives equal weight to the three classes, while the F_2-score gives more weight to recall than precision. This is useful in an early-warning setting because missing an extreme-price event is generally more problematic than issuing an additional warning, although both types of error remain relevant. The full evaluation framework is defined in Section 5.6.

The dummy classifier is not tuned, because it is only included as a naive benchmark. For the other candidate models, predefined parameter grids are used. These grids are kept relatively limited to reduce computation time and to avoid overfitting on the development period. Parameters not included in

the grids are kept fixed at their predefined values, as defined in Appendix D.2. The full hyperparameter grids, fixed implementation settings, and final selected model settings are reported in Appendix D.

After grid search, the best estimator for each model is retained and used to generate class probabilities for the 2024 validation period. These validation probabilities are then used for warning-threshold tuning, which is described in the next section. The 2025 test set is not used during hyperparameter tuning or warning-threshold selection. After the model settings and warning thresholds have been selected, the tuned models are retrained on the combined 2022–2024 training period and evaluated once on the 2025 test period.

5.5. Probability outputs and warning-threshold design

All candidate models produce a probability distribution over the three target classes for each delivery hour:

$$P(y_t = \text{negative}), \quad P(y_t = \text{normal}), \quad P(y_t = \text{spike})$$

These probabilities are the main output of the early-warning model. Normally, the class with the highest probability would be predicted. This is not sufficient for the early-warning task in this thesis, because negative prices and price spikes are rare. The normal class can be the most likely class even when the probability of an extreme event is high enough to be relevant for decision-making. For classification with imbalanced data, fixed default thresholds can be unsuitable and affect model performance (Zou et al., 2016). Therefore, separate warning thresholds are applied to the two extreme events. These thresholds are selected separately for each model, because the distribution and calibration of predicted probabilities can differ between model families. As a result, the same numerical threshold can lead to different warning volumes and different precision–recall trade-offs across models. A single shared threshold would therefore not provide a fair comparison of early-warning performance.

Let τ_{neg} indicate the negative-price threshold and τ_{spike} indicate the price-spike threshold. Each observation is first assigned to the normal class by default. A negative-price warning is triggered when:

$$P(y_t = \text{negative}) \geq \tau_{\text{neg}}$$

A price-spike warning is triggered when:

$$P(y_t = \text{spike}) \geq \tau_{\text{spike}}$$

If only the negative-price threshold is exceeded, the observation is predicted as a negative-price warning. If only the price-spike threshold is exceeded, the observation is predicted as a price-spike warning. However, both thresholds can be exceeded for the same hour. In that situation, the model assigns the hour to the event class with the higher predicted probability. Ties are assigned to the negative-price class. This results in the following prediction rule:

$$\hat{y}_t = \begin{cases} \text{negative}, & \text{if } P(y_t = \text{negative}) \geq \tau_{\text{neg}} \text{ and } \big(P(y_t = \text{spike}) < \tau_{\text{spike}} \text{ or } P(y_t = \text{negative}) \geq P(y_t = \text{spike})\big) \\ \text{spike}, & \text{if } P(y_t = \text{spike}) \geq \tau_{\text{spike}} \text{ and } \big(P(y_t = \text{negative}) < \tau_{\text{neg}} \text{ or } P(y_t = \text{spike}) > P(y_t = \text{negative})\big) \\ \text{normal}, & \text{otherwise.} \end{cases}$$

The threshold design separates probability estimation from the decision rule used to issue warnings. This distinction is important because the model first estimates class probabilities, after which the warning rule determines when those probabilities are high enough to trigger an action. Using different thresholds is a common way to convert probability estimates into class labels or actions, especially when different types of errors are not equally important (Sheng & Ling, 2006).

The thresholds are selected using the 2024 validation period. For each model, a grid search is performed over possible values of τ_{neg} and τ_{spike}. The threshold grid is:

$$\{0.001, 0.002, 0.005, 0.01, 0.02, 0.03, 0.05, 0.075, 0.10, 0.15, 0.20, 0.25, 0.30, 0.40, 0.50\}$$

Each threshold pair is applied to the probabilities calculated on the validation set and evaluated using the macro F_2-score. The threshold pair with the highest validation macro F_2 is selected for that

model. Macro F_2 is used because it gives equal weight to the three target classes and gives more weight to recall than precision. This fits the early-warning objective and is explained further in Section 5.6.

The selected thresholds are used for the model comparison and warning-behaviour analysis. They should not be interpreted as actor-specific operational thresholds, because operational use would require additional information on the costs, constraints, and risk preferences of the market participant.

After the model-specific thresholds are selected on the validation period, they are fixed. The 2025 test set is not used to choose or adjust thresholds. This avoids using the final test year for model-selection decisions and reduces the risk of selection bias in the reported performance (Cawley & Talbot, 2010). In the final evaluation, each model is retrained on the combined 2022–2024 period and evaluated once on the 2025 test set using the thresholds selected during validation. The selected threshold values are reported in Appendix D.4 and interpreted together with the final model-comparison results in Chapter 6.

5.6. Evaluation metrics

The models are evaluated from the perspective of their intended use as early-warning tools. This means that performance is not only assessed by whether the final class prediction is correct. The model must also be able to detect negative-price and price-spike events, produce probability estimates that can be interpreted as risk signals, and avoid an excessive number of false warnings. The evaluation framework therefore links the usefulness criteria from Section 3.3.3 to four questions: are extreme events detected, do false warnings remain limited, is the probability ranking useful, and are the predicted probabilities reliable enough for decision-oriented interpretation.

5.6.1. Class-specific prediction performance

Since the target variable is imbalanced, class-specific metrics are used as the main basis for evaluating prediction performance. Overall accuracy is reported as a secondary metric. It is not used as the main model-selection criterion due to the imbalance of the target variable. Most observations are normal-price hours, while negative prices and price spikes are only a small part of the dataset. A model that predicts all hours as normal, such as the dummy classifier, can therefore obtain a high accuracy while failing to detect extreme events. This problem is well established in imbalanced classification, where accuracy can give misleadingly high performance when a model mainly predicts the majority class (Brodersen et al., 2010). In the context of extreme electricity prices, model evaluation should therefore focus on the detection of rare event classes rather than on average classification accuracy alone. Liu et al. (2022) treat extremely low and high electricity prices as occurrence events that can be predicted with classification methods, while Loizidis et al. (2025) evaluate negative-price classification with class-specific recall, precision, and F_1-score.

The first part of the evaluation framework therefore focuses on class-specific prediction performance, using precision, recall, macro F_2, balanced accuracy, and confusion matrices. The mathematical definitions follow the implementations used in the modelling pipeline, which are based on the corresponding functions in *scikit-learn* (scikit-learn developers, 2026a). For each class c, precision and recall are defined as (scikit-learn developers, 2026g, 2026h):

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}$$

where TP_c, FP_c, and FN_c indicate the number of true positives, false positives, and false negatives for class c. Precision and recall describe two different aspects of warning performance. Recall measures the share of realised events in class c that are detected by the model. Missed events correspond to false negatives, making recall especially important for the negative-price and price-spike classes. Precision measures the share of correctly predicted warnings. It is needed to assess whether high recall is achieved at the cost of too many false warnings. Both metrics are standard classification measures derived from the confusion matrix (Powers, 2011).

The F-score combines precision and recall into a single metric. This thesis uses the F_2-score rather than the F_1-score to give more emphasis to recall (scikit-learn developers, 2026f):

$$F_{\beta,c} = (1+\beta^2)\frac{\text{Precision}_c \cdot \text{Recall}_c}{\beta^2 \cdot \text{Precision}_c + \text{Recall}_c}$$

with $\beta = 2$. This choice reflects the early-warning purpose of the model. Missing an extreme-price event is given more weight than issuing an additional warning, although false warnings also remain relevant. The F_2-score is therefore more suitable than F_1 when recall is considered more important than precision. Since F-scores can hide the precision-recall trade-off, class-specific precision and recall are also reported separately (Christen et al., 2023).

The class-specific F_2-scores are reported for negative prices and price spikes. Macro F_2 is used as the main aggregate classification metric by averaging the class-specific F_2-scores across the three target classes:

$$\text{Macro F2} = \frac{1}{|C|}\sum_{c \in C} F_{2,c}$$

where C is the set of three target classes: negative price, normal price, and price spike. This corresponds to macro averaging, where each class contributes equally to the final score. Since the goal is to compare models on their ability to handle all three classes, this is especially important.

Balanced accuracy is included for the same reason. In the multiclass setting, balanced accuracy is calculated as the average recall across classes (scikit-learn developers, 2026c):

$$\text{Balanced accuracy} = \frac{1}{|C|}\sum_{c \in C} \text{Recall}_c$$

Balanced accuracy reduces the influence of class imbalance by treating each class as equal. It therefore provides a more informative measure than regular accuracy, when a classifier could perform well by mainly predicting the majority class (Brodersen et al., 2010).

5.6.2. Probability rankings

The second part of the evaluation framework assesses whether the models rank risky hours above less risky hours. The main ranking metric is average precision, reported separately for the negative-price class and the price-spike class in a one-versus-rest setting. Average precision summarises the precision-recall curve across thresholds and is used here as the PR-curve summary for the two event probabilities (scikit-learn developers, 2026b):

$$AP = \sum_n (R_n - R_{n-1})P_n$$

where P_n and R_n are the precision and recall at threshold n.

5.6.3. Probability quality

The third part of the evaluation framework assesses probability quality. Since the models produce probabilities rather than only hard class labels, the predicted probability distributions should be meaningful as risk estimates. Two scoring-rule metrics are used for this purpose: log loss and the multiclass Brier score. Scoring rules are useful for evaluating probabilistic forecasts because they assess the full predictive distribution (Gneiting & Raftery, 2007). This is consistent with probabilistic electricity-price forecasting literature, where forecast uncertainty is evaluated rather than only point predictions (Haben et al., 2021).

Log loss evaluates the probability assigned to the realised class. It is defined as the negative log-likelihood of the true labels given the predicted class probabilities. In the multiclass setting used in this thesis, log loss is calculated as:

$$L_{\log}(Y,P) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k}\log(p_{i,k})$$

where N is the number of observations, K is the number of classes, $y_{i,k}$ equals 1 if observation i belongs to class k and 0 otherwise, and $p_{i,k}$ is the predicted probability for class k. In this thesis,

$K = 3$, corresponding to negative prices, normal prices, and price spikes. Lower log loss values indicate better probability estimates. Because log loss penalises confident incorrect predictions, it is useful for evaluating whether the predicted probabilities are meaningful as early-warning risk signals.

The multiclass Brier score measures the squared distance between the predicted probability vector and the realised class vector (scikit-learn developers, 2026d):

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(y_{i,k} - p_{i,k})^2$$

where $y_{i,k}$ equals 1 if observation i belongs to class k and 0 otherwise, and $p_{i,k}$ is the predicted probability for class k. It is a multiclass version of the Brier score originally developed for probabilistic forecast verification (Brier, 1950; Gneiting & Raftery, 2007). Lower Brier scores indicate that the predicted probabilities are closer to the actual outcomes. The Brier score and log loss are interpreted as the probability-quality metrics. Calibration and reliability are discussed further in Section 5.7.

5.6.4. Warning behaviour and error structure

Finally, confusion matrices and warning counts are used to interpret the error structure of each model. A confusion matrix reports how observations from each realised class are distributed over the predicted classes (scikit-learn developers, 2026e). It shows whether a model mainly misses extreme events, confuses event types, or produces many false event warnings. Warning counts report how many hours are classified as negative-price warnings or price-spike warnings after applying the tuned probability thresholds from Section 5.5. They are included because high recall alone is not sufficient: a model that detects most events but warns during too many normal hours can be difficult to use in practice. Together, these metrics provide a consistent basis for comparing the candidate models in the final evaluation.

5.7. Probability-quality assessment

The previous section defined log loss and the Brier score as probability-quality metrics. This section explains how the probability outputs are further assessed in terms of calibration and reliability. The models in this thesis are used to estimate the probability of negative prices and price spikes. These probabilities should therefore be meaningful as risk signals.

Calibration refers to the relation between predicted probabilities and observed event frequencies. A model is well calibrated when predicted probabilities correspond to the frequency with which events actually occur. For example, if a model assigns a negative-price probability of around 10% to a group of hours, approximately 10% of those hours should result in a negative price. A model can still achieve useful classification performance after applying warning thresholds, while producing probabilities that are systematically too high or too low. In that case, the probabilities may still help with ranking risky hours, but they are less reliable as direct risk estimates.

Calibration is assessed using reliability curves. In a reliability curve, predicted probabilities are divided into bins. For each bin, the average predicted probability is compared with the observed event frequency. The diagonal line represents perfect calibration. Points above this line indicate that the model underestimates the event probability, because the event occurs more often than predicted. Points below the line indicate overestimation, because the event occurs less often than predicted. Reliability curves are commonly used to determine whether predicted probabilities can be interpreted as actual event frequencies (Dimitriadis et al., 2021).

Because the modelling problem has three target classes, calibration is assessed separately for the two extreme-event probabilities. The negative-price reliability curve compares $P(y_t = \text{negative})$ with the realised occurrence of negative prices. The spike reliability curve compares $P(y_t = \text{spike})$ with the realised occurrence of price spikes. The normal-price probability is excluded because the early-warning interpretation depends mainly on the two extreme-event probabilities. The full three-class probability distribution is still evaluated through the multiclass Brier score and log loss.

The reliability curves are created from the final 2025 test-set predictions. They are used as diagnostic evidence after model training, hyperparameter tuning, and warning-threshold selection have been completed. This keeps the calibration assessment separate from model selection and preserves the role of 2025 as the final evaluation period.

The calibration results are interpreted together with the Brier score and log loss. Together, these outputs show whether a model is not only able to detect extreme events, but also whether its probabilities can be interpreted as useful early-warning risk signals. The results of the calibration and reliability assessment are reported in Chapter 6.

5.8. Explainability approach

Explainability is used to examine how the final model outputs for negative prices and price spikes relate to the input features. The analysis is applied after model selection and is used to interpret the selected final model. The purpose is to assess which variables contribute most strongly to the predicted event risk and whether these contributions are consistent with the mechanisms identified in Chapter 2.

The main explainability method used in this thesis is SHapley Additive exPlanations (SHAP). SHAP values are based on Shapley values from cooperative game theory (Shapley, 1953) and were introduced as a unified additive feature-attribution framework by Lundberg and Lee (2017). The method decomposes a model output into a baseline value and feature-specific contributions:

$$f(x) = \phi_0 + \sum_{j=1}^{p} \phi_j$$

where ϕ_0 is the expected model output, ϕ_j is the contribution of feature j, and p is the number of input features. The SHAP value for feature j is calculated as the weighted average of all possible differences across feature subsets:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_x(S \cup \{j\}) - f_x(S)]$$

where F is the full feature set, S is a subset of features not containing feature j, and $f_x(S)$ is the model output when only the features in subset S are known. The difference $f_x(S \cup \{j\}) - f_x(S)$ represents the marginal contribution of feature j for that subset. This contribution is averaged over all possible feature orderings. A SHAP value explains a single prediction by showing how much a feature moves the model output for a specific class away from the expected model output. A positive SHAP value increases the model output for the class being predicted, while a negative SHAP value means that the feature decreases it. These contributions describe how the trained model uses the features for prediction, and can therefore not be directly interpreted as causal effects on market outcomes.

Since the target variable has three classes, the SHAP analysis is performed separately for the two extreme classes. For the negative-price class, SHAP values show which features increase or decrease the model output associated with negative-price predictions. For the spike class, SHAP values show which features increase or decrease the model output associated with spike predictions. This class-specific interpretation is necessary because the same feature can contribute differently to negative-price and spike predictions.

Two SHAP outputs are used. SHAP importance is calculated by taking the mean absolute SHAP value of each feature across the test observations. This gives a ranking of the features that contribute most strongly to the model output for each event class. SHAP summary plots are then used to assess the direction of feature effects, by showing whether high or low feature values generally increase or decrease the model output.

The SHAP analysis is performed on the final 2025 test-set predictions using an explainer that matches the structure of the selected final model. The SHAP results are used only after the evaluation has been completed. The values are interpreted as model-specific explanations of the predictions of the selected final model. They show which features contribute most strongly to the model output for negative-price and price-spike risk. The explainability results are used to support the interpretation of the selected early-warning model in Chapter 6.

5.9. Feature-group removal design

The feature-group removal analysis tests the added value of cross-border indicators. It directly addresses SQ 4 by comparing a domestic baseline feature set with alternative feature sets that add different types of cross-border information. The purpose is to assess whether cross-border variables

improve the ex-ante prediction of Dutch negative prices and price spikes beyond information from Dutch market and system conditions.

The domestic baseline consists of Dutch price-history variables, Dutch system variables, Dutch weather variables, and Dutch calendar controls. Cross-border variables are added in separate blocks: scheduled exchange, transfer capacity, neighbouring-system forecasts, neighbouring lagged prices, and neighbouring weather and calendar variables. The core cross-border scenario combines the cross-border variables most directly related to market coupling: scheduled exchange, transfer capacity, neighbouring-system forecasts, and neighbouring lagged prices. The full feature set includes all available variables. Table 5.3 summarises the feature-group scenarios used in the removal analysis.

Table 5.3: Feature-group scenarios used in the removal analysis

Scenario	Description	Features
Domestic baseline	Dutch price-history, Dutch system, Dutch weather, and Dutch calendar variables	32
Domestic + scheduled exchange	Domestic baseline plus directional scheduled import and export variables	40
Domestic + transfer capacity	Domestic baseline plus directional and aggregate day-ahead transfer-capacity variables	42
Domestic + neighbouring system	Domestic baseline plus neighbouring load, generation, renewable-generation, and residual-load forecasts	54
Domestic + neighbouring prices	Domestic baseline plus lagged neighbouring day-ahead prices	36
Domestic + neighbouring weather/calendar	Domestic baseline plus neighbouring weather and holiday controls	58
Domestic + core cross-border	Domestic baseline plus scheduled exchange, transfer capacity, neighbouring-system forecasts, and neighbouring lagged prices	76
Full feature set	All available domestic, weather, cross-border, and calendar variables	102

The removal analysis is performed for logistic regression, GBDT, and XGBoost. For each model and feature scenario, the same time-based evaluation design is used as in the main model comparison. Models are trained on 2022–2023, warning thresholds are selected on 2024, and final performance is evaluated on the 2025 test set after refitting on 2022–2024. The tuned model specifications from the main modelling pipeline are reused, so the removal focuses on the effect of changing the feature set rather than on repeated hyperparameter tuning. This keeps the original variables in the model and assesses overlap between feature groups by comparing their predictive contribution, rather than by replacing the original feature set with PCA components.

The added value is measured relative to the domestic baseline. For each model and feature scenario, the difference in a performance metric is calculated as:

$$\Delta M_{m,s} = M_{m,s} - M_{m,\text{domestic baseline}}$$

where $M_{m,s}$ is the value of metric M for model m under feature scenario s, and $M_{m,\text{domestic baseline}}$ is the value of the same metric for the domestic baseline. Positive values indicate that the additional feature block improves performance compared with domestic information only, while negative values indicate weaker performance.

An additional analysis assesses which neighbouring bidding zones contribute most to the core cross-border feature set. In this analysis, DE/LU, BE, NO2, and DK1 are removed one at a time. Total import and export capacity variables are excluded from this comparison because they combine information across neighbours.

The removal results are interpreted as predictive added value. A performance increase after adding or keeping a feature group indicates that the feature group contains useful information for prediction in the 2025 test period. The results of the removal analysis are reported in Chapter 6.

5.10. Target-definition sensitivity design

The target-definition sensitivity analysis tests whether the modelling conclusions depend on the chosen definition of negative prices and price spikes. The baseline target definition defines negative prices as prices below 0 €/MWh and price spikes as prices above the month-year-specific 95th percentile. Because this definition affects class balance, warning thresholds, and model performance, alternative target definitions are evaluated as well.

The sensitivity analysis varies both the upper-tail definition and the lower-tail definition. For the upper tail, the baseline is compared with a broader 90th percentile, a stricter 97.5th percentile, a fixed full-sample 95th percentile, and rolling 95th-percentile thresholds. The rolling thresholds are included because they are more directly aligned with a real-time early-warning setting. At the start of a new month, the full month-year price distribution is not yet known, while thresholds based on recent past prices can already be calculated. For the lower tail, two alternatives are tested: a zero-negative definition, which includes prices equal to 0 €/MWh in the lower-tail class, and a broader low-price definition in which prices up to 5 €/MWh are treated as low-price events. Table 5.4 summarises the target-definition scenarios used in the sensitivity analysis.

Table 5.4: Target-definition scenarios used in the sensitivity analysis

Scenario	Definition	Role
Month-year 95th	Price < 0; spike above month-year 95th percentile	Baseline
Month-year 90th	Price < 0; spike above month-year 90th percentile	Broader spike check
Month-year 97.5th	Price < 0; spike above month-year 97.5th percentile	Stricter spike check
Full-sample 95th	Price < 0; spike above full-sample 95th percentile	Fixed-threshold check
Rolling past 7-day 95th	Price < 0; spike above previous 7-day 95th percentile	Ex-ante rolling check
Rolling past 14-day 95th	Price < 0; spike above previous 14-day 95th percentile	Ex-ante rolling check
Rolling past 28-day 95th	Price < 0; spike above previous 28-day 95th percentile	Ex-ante rolling check
Centred 28-day 95th	Price < 0; spike above centred 28-day 95th percentile	Contextual check
Zero-negative	Price ≤ 0; spike above month-year 95th percentile	Lower-tail check
Low-price stress	Price ≤ 5 EUR/MWh; spike above month-year 95th percentile	Lower-tail check

For each scenario, a new three-class target variable is constructed. Observations are first assigned to the normal class. They are then relabelled as lower-tail events when they satisfy the scenario-specific lower-tail rule, and as spike events when they exceed the scenario-specific upper-tail threshold. The sensitivity dataset is restricted to the modelling period from 2022 to 2025. Additional observations outside this period are only used where needed to calculate rolling threshold values and are not included in model training, validation, testing, or class-balance reporting.

The sensitivity analysis is performed for logistic regression, GBDT, and XGBoost. The tuned model specifications from the main modelling pipeline are reused. For each target-definition scenario, the same time-based evaluation design is applied: models are trained on 2022–2023, warning thresholds are selected on 2024, and final performance is evaluated on 2025 after refitting on 2022–2024. This keeps the analysis focused on the effect of changing the target definition rather than on model-tuning.

The same warning-threshold rule and evaluation metrics are used in every scenario. The results are used to assess whether the main model ranking, event-detection performance, probability quality, and warning behaviour remain stable under alternative definitions of negative prices and price spikes. The results of the target-definition sensitivity analysis are reported in Chapter 6.

5.11. Cost-sensitive interpretation framework

The final part of the evaluation interprets the model outputs from a cost-sensitive perspective. A missed negative price or missed price spike may expose a market participant to unexpected price risk, while a false warning may lead to unnecessary bidding adjustments, hedging actions, or flexibility activation.

The cost-sensitive interpretation is applied separately to negative-price risk and price-spike risk. For each event class e, where e represents either a negative price or a price spike, the warning problem is defined as follows. A false negative occurs when event e is realised but no warning is issued. A false positive occurs when a warning for event e is issued but the event does not occur.

Let p_e be the predicted probability of event e, C_e^{FN} the cost of missing the event, and C_e^{FP} the cost of issuing a false warning. If correct decisions are assigned zero cost, the expected cost of not warning is:

$$EC_{\text{no warning},e} = p_e C_e^{FN}$$

The expected cost of issuing a warning is:

$$EC_{\text{warning},e} = (1 - p_e) C_e^{FP}$$

A warning is cost-sensitive when the expected cost of warning is lower than the expected cost of not warning. This gives the following decision condition:

$$p_e > \frac{C_e^{FP}}{C_e^{FP} + C_e^{FN}}$$

The appropriate warning threshold depends on the relative costs of false positives and false negatives. If missed events are more costly than false warnings, a lower threshold is justified. If false warnings are costly, a higher threshold is more appropriate. This follows the minimum expected-cost logic of cost-sensitive classification and the use of thresholds (Elkan, 2001; Sheng & Ling, 2006).

Monetary values are not assigned to false positives and false negatives, because these costs depend on the position, flexibility options, and risk preferences of the market participant. The cost-sensitive framework is used as a decision-support interpretation, rather than as a set of economic guidelines. Chapter 6 applies this framework to the final model outputs by interpreting false positives, false negatives, warning counts, and probability levels.

6

Results

This chapter presents the results of the modelling, target-definition sensitivity, feature-removal, calibration, and decision-interpretation analyses. The chapter is structured to answer the empirical parts of the research question while keeping the main text concise. Section 6.1 first describes the final target definition and the class distribution in the final 2025 test set. Section 6.2 compares the final models on predictive performance and probability quality, including a reversed year-order robustness check. Section 6.3 then examines warning behaviour and the structure of false positives and false negatives. Section 6.4 evaluates the robustness of the model conclusions to alternative target definitions. Section 6.5 reports the added value of cross-border indicators. Finally, Sections 6.6–6.8 interpret the reliability, explainability, and decision relevance of the final model.

6.1. Target-definition and test-set characteristics

The results in this chapter are evaluated using the baseline target definition and train-validation-test design introduced in Chapters 4 and 5. This section only reports the class composition needed to interpret the model comparison, warning behaviour, and calibration results.

Table 6.1 shows the class distribution across the modelling splits. After model settings and warning thresholds have been selected, the final models are refitted on the combined 2022–2024 period and evaluated once on the unseen 2025 test set.

Table 6.1: Class distribution across modelling splits

Split	Negative prices	Normal prices	Price spikes	Total
Development training (2022–2023)	400 (2.3%)	15,893 (92.6%)	862 (5.0%)	17,155
Validation (2024)	457 (5.2%)	7,808 (89.7%)	443 (5.1%)	8,708
Final training (2022–2024)	857 (3.3%)	23,701 (91.6%)	1,305 (5.0%)	25,863
Final test (2025)	572 (6.7%)	7,585 (88.3%)	433 (5.0%)	8,590

Note: Percentages are calculated within each split. The final training set combines the development training and validation periods after model settings and warning thresholds have been selected.

The table shows that all splits remain imbalanced. Normal-price hours account for most observations, while negative prices and price spikes form the two smaller event classes. This is important for the following sections, because high overall accuracy can be achieved even when a model fails to detect rare events. The model comparison therefore focuses on class-specific recall and precision, macro F_2 scores, balanced accuracy, probability quality, and warning behaviour.

The main difference between the splits is the increasing share of negative-price hours in the later years. Negative prices account for 2.3% of the development training period, 5.2% of the validation period, and 6.7% of the final test period. The final evaluation is therefore performed in a period in which negative prices are more frequent than in the original training period. This should be taken into account when interpreting the 2025 test results. However, before final testing, the models are refitted on the combined 2022–2024 final training set, so the higher frequency of negative prices in

2024 is included in the final training data. The price-spike share remains close to 5% in each split by construction. The final 2025 test set contains enough observations in both event classes to evaluate early-warning performance separately for negative prices and price spikes. At the same time, the class imbalance is still present, motivating the use of event-specific metrics.

6.2. Final 2025 model comparison

This section compares the final model performance on the unseen 2025 test set. The models are evaluated after hyperparameter selection and warning-threshold tuning on the validation period. The final models are then refitted on the combined 2022–2024 training period and evaluated once on the 2025 test set. The comparison therefore reflects the final out-of-sample performance of the candidate early-warning models.

Table 6.2 summarises the main performance indicators. The table focuses on balanced and event-specific metrics, because the class distribution in Section 6.1 showed that overall accuracy is not sufficient for evaluating rare-event warning performance. The colour scale is applied per metric, using blue–orange to indicate stronger and weaker values.

Table 6.2: Final 2025 model comparison

Model	Macro F_2	Bal. acc.	Rec. neg.	Prec. neg.	Rec. spike	Prec. spike	Log loss	Brier
XGBoost	0.812	0.895	0.969	0.459	0.845	0.521	0.161	0.095
GBDT	0.806	0.885	0.932	0.490	0.850	0.477	0.178	0.101
MLP	0.741	0.813	0.752	0.571	0.827	0.329	0.715	0.170
Logistic regression	0.734	0.892	0.984	0.392	0.935	0.294	0.533	0.314
PNN	0.693	0.871	0.948	0.426	0.961	0.217	0.566	0.366
Dummy baseline	0.325	0.333	0.000	0.000	0.000	0.000	4.041	0.234

Note: Blue indicates stronger values, while orange indicates weaker values. Higher values are better for macro F_2, balanced accuracy, recall, and precision. Lower values are better for log loss and the multiclass Brier score. Bal. acc. refers to balanced accuracy. Rec. and Prec. refer to recall and precision.

The strongest overall performance is obtained by XGBoost and GBDT. XGBoost has the highest macro F_2 score, with a value of 0.812, and the highest balanced accuracy, with a value of 0.895. It also has the lowest log loss and multiclass Brier score, which indicates that it provides the strongest probability quality among the evaluated models. GBDT performs very similarly, with a macro F_2 score of 0.806 and a balanced accuracy of 0.885. The difference between the two tree-based models is therefore small, but XGBoost provides the best overall combination of event detection and probability quality.

Figure 6.1 confirms that the tree-based models form the strongest model group in terms of macro F_2 and balanced accuracy. The dummy baseline obtains a relatively high normal-class accuracy, but its balanced accuracy and macro F_2 score are low because it does not detect any negative-price or price-spike events. This confirms that the final model selection cannot be based on overall accuracy.

The comparison also shows that high recall alone is not sufficient for selecting the final model. Logistic regression and PNN achieve high recall for both event classes, but this comes at the cost of substantially lower precision and weaker probability quality. In practical early-warning applications, this would lead to many additional false warnings. The MLP achieves the highest precision for negative prices, but its lower event recall and weaker probability metrics make it less suitable as the final model.

Based on the final 2025 model comparison, XGBoost is selected as the preferred model for the remaining results. GBDT is kept as the main benchmark because its performance is close to XGBoost and because it provides an additional check on whether the results are specific to one tree-based implementation. Logistic regression is kept in further analyses as an interpretable baseline, allowing the performance and warning behaviour of the tree-based models to be compared with a less flexible probabilistic classifier.

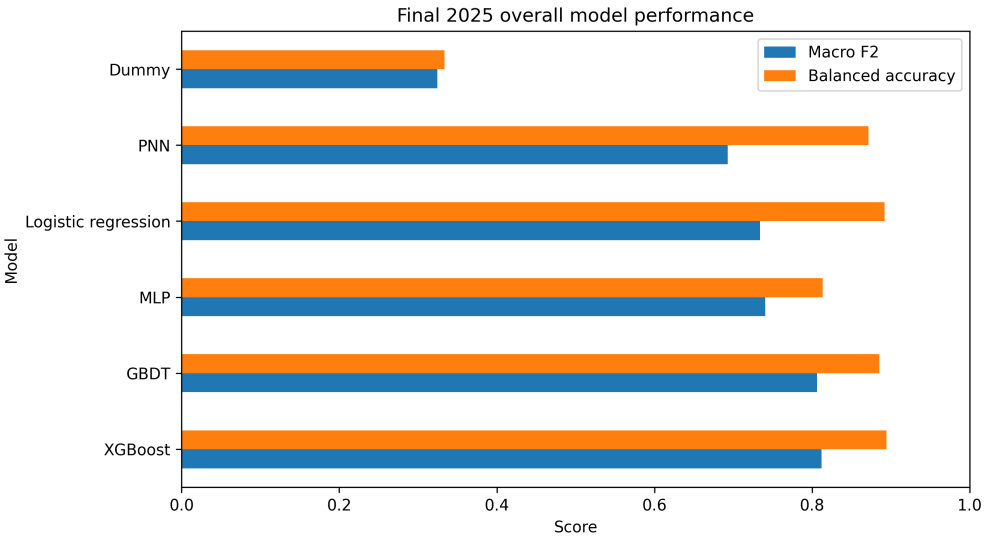

Figure 6.1: Final 2025 overall model performance

Table 6.3: Reverse-year macro F_2 performance

Model	Main test (2025)	Reverse-year test (2022)	Difference
XGBoost	0.812	0.716	-0.096
GBDT	0.806	0.642	-0.164
MLP	0.741	0.558	-0.183
Logistic regression	0.734	0.597	-0.137
PNN	0.693	0.487	-0.206
Dummy baseline	0.325	0.329	0.004

The model comparison is repeated with the year order reversed as a robustness check. In this setup, observations from 2023–2025 are used for model development and 2022 is used as the final test year. Table 6.3 shows that the macro F_2 scores are lower for all models in the reverse-year test, except for the dummy baseline. This indicates that the 2022 test period differs from 2025 in terms of event distribution and market conditions. However, the model ranking remains unchanged. The tree-based models remain the strongest model group, with XGBoost obtaining the highest reverse-year macro F_2 score and GBDT remaining a close second. Logistic regression performs below the tree-based models, while MLP and PNN show weaker performance. The check therefore supports the robustness of the model ranking. The chronological 2025 test remains the primary evaluation because it best reflects the intended forecasting setting.

6.3. Warning behaviour and error structure

The performance metrics in Section 6.2 show which models perform best overall, but they do not show how the models behave as early-warning systems. This section examines the confusion matrices and warning counts in the final 2025 test set. The focus lies on missed events and false warnings, since both affect the operational usefulness of an early-warning model.

Figure 6.2 shows the confusion matrices for the final model comparison. XGBoost correctly identifies 554 of 572 negative-price hours and 366 of 433 price-spike hours. The remaining errors for XGBoost mainly consist of event hours classified as normal: 18 negative-price hours and 67 price-spike hours are missed. GBDT shows a similar structure, with 533 correctly identified negative-price hours and 368 correctly identified price-spike hours. Compared with XGBoost, GBDT misses more negative-price hours but slightly fewer price-spike hours. Logistic regression detects the largest number of event hours, with 563 correctly identified negative-price hours and 405 correctly identified price-spike hours. This is due to more normal-price hours being classified as extreme events, especially for price spikes.

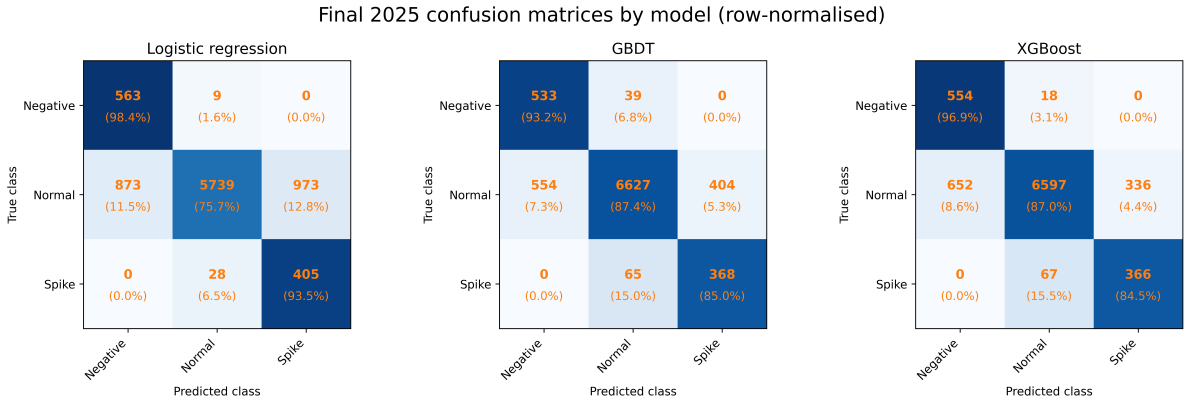

Figure 6.2: Final 2025 confusion matrices by model

For the models shown in Figure 6.2, the confusion matrices show that the two extreme classes are not confused with each other. Negative-price hours are not classified as price spikes, and price-spike hours are not classified as negative prices. The main trade-off is therefore not between the two types of extreme events, but between detecting extreme events and incorrectly assigning normal-price hours to an extreme class.

This trade-off becomes clearer in Figure 6.3. XGBoost issues 1,206 negative-price warnings and 702 price-spike warnings. GBDT issues fewer negative-price warnings, with 1,087 warnings, but slightly more price-spike warnings, with 772 warnings. Logistic regression issues substantially more warnings: 1,436 negative-price warnings and 1,378 price-spike warnings. This explains why logistic regression shows a high recall in Section 6.2, but is less attractive as the final early-warning model. Its additional event detection comes at the cost of many more false warnings, especially for price spikes.

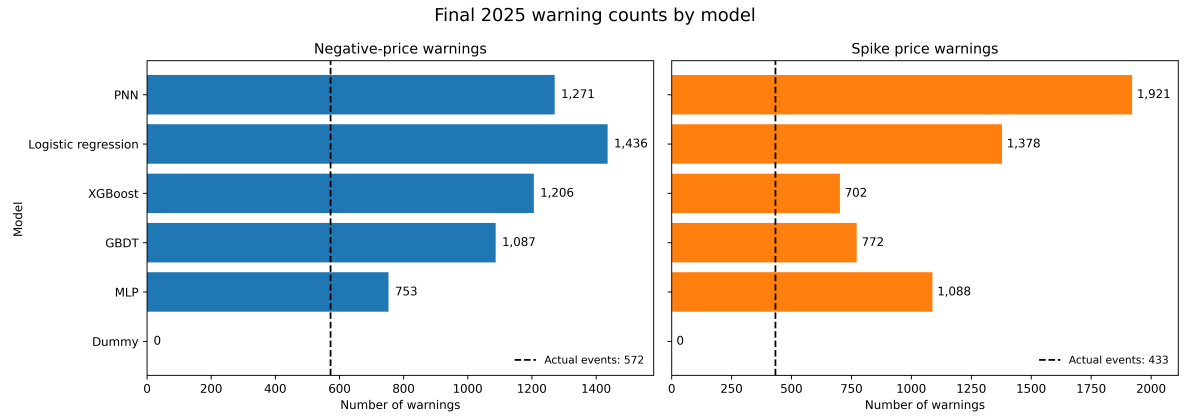

Figure 6.3: Final 2025 warning counts by model. The vertical line indicates the number of realised event hours.

The same pattern is visible for the PNN, which produces the largest number of price-spike warnings. Although this results in high price-spike recall, it also means that many normal-price hours are classified as price-spike warnings. Such behaviour can be problematic in an operational setting, because frequent false alarms reduce the practical value of a warning system and may lead users to ignore warnings over time.

Overall, the warning-behaviour results support the selection of XGBoost as the preferred final model. It detects most negative-price and price-spike events while keeping the number of false warnings lower than the more aggressive classifiers. GBDT remains a close benchmark, especially because its error structure is similar to XGBoost. Logistic regression remains useful as an interpretable baseline, but its warning behaviour shows the limitations of relying on a less flexible classifier for this prediction problem.

6.4. Target-definition sensitivity results

The previous sections evaluated the final models using the baseline target definition. This section tests whether the main modelling conclusions are sensitive to alternative definitions of negative-price and price-spike events. This assesses whether the ranking and warning behaviour of the main models depend strongly on the target-definition choice.

Table 6.4 summarises the main sensitivity results by reporting the number of lower-tail and price-spike events in the 2025 test set, together with the macro F_2 score for each model. The lower-tail class refers to negative prices in the main threshold scenarios, but is broadened to zero-negative and low-price hours in the final two scenarios. The price-spike definition remains based on the baseline month-year 95th-percentile threshold in these scenarios, resulting in an identical price-spike count to the baseline scenario.

Table 6.4: Target-definition sensitivity results on the 2025 test set

Scenario	Lower-tail events	Spike events	Log. reg.	GBDT	XGBoost
Baseline MY95	572	433	0.717	0.801	0.812
Broader spike MY90	572	863	0.747	0.818	0.810
Stricter spike MY97.5	572	219	0.679	0.790	0.792
Full-sample 95th	572	13	0.588	0.566	0.639
Rolling past 7d	572	524	0.713	0.790	0.787
Rolling past 14d	572	443	0.708	0.781	0.787
Rolling past 28d	572	423	0.719	0.765	0.772
Rolling centred 28d	572	412	0.673	0.798	0.793
Zero-negative	635	433	0.726	0.809	0.815
Low-price ≤ 5	863	433	0.736	0.821	0.813

Note: Macro F2 scores. MY95, MY90, and MY97.5 refer to month-year percentile spike thresholds. The lower-tail class refers to negative prices, except in the zero-negative and low-price sensitivity scenarios.

The results show that the main model ranking is broadly stable across the sensitivity scenarios. XGBoost and GBDT remain the strongest models in all cases, while logistic regression remains useful as a simple baseline with lower macro F_2 scores. In the baseline scenario, XGBoost obtains the highest macro F_2 score of the evaluated models. Combined with its stronger probability quality results in Section 6.2, its selection as the final model is supported. In some alternative definitions, GBDT performs similarly or slightly better, which indicates that the sensitivity analysis supports the robustness of tree-based models rather than the dominance of one specific implementation.

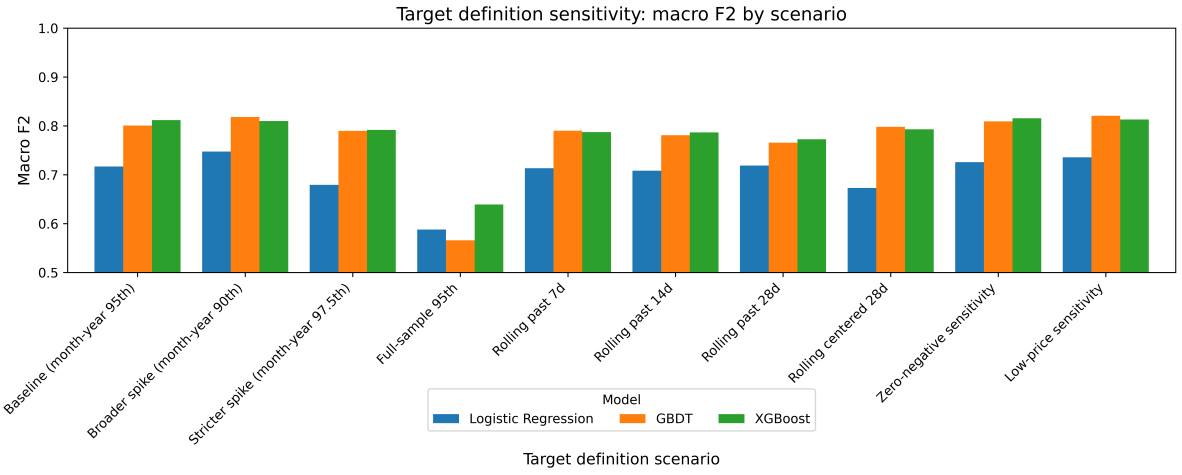

Figure 6.4: Target-definition sensitivity of macro F_2 score

Figure 6.4 confirms this pattern visually. The broader 90th-percentile price-spike definition and the zero-negative and low-price sensitivity definitions produce macro F_2 scores close to the baseline.

The stricter 97.5th-percentile price-spike definition leads to a slightly lower macro F_2 score, which is expected because the price-spike class becomes smaller and more difficult to detect.

The rolling-threshold scenarios provide a similar interpretation. Their macro F_2 scores are close to the baseline, especially for XGBoost and GBDT. From an operational perspective, these scenarios remain relevant because they can be calculated from recently observed prices, whereas the full month-year threshold is mainly suitable for retrospective labelling. However, they do not clearly improve performance compared with the baseline month-year threshold. This suggests that making the price-spike threshold more adaptive over time does not provide a clear predictive advantage for the final test year, although it may be more practical for real-time use early in a new month.

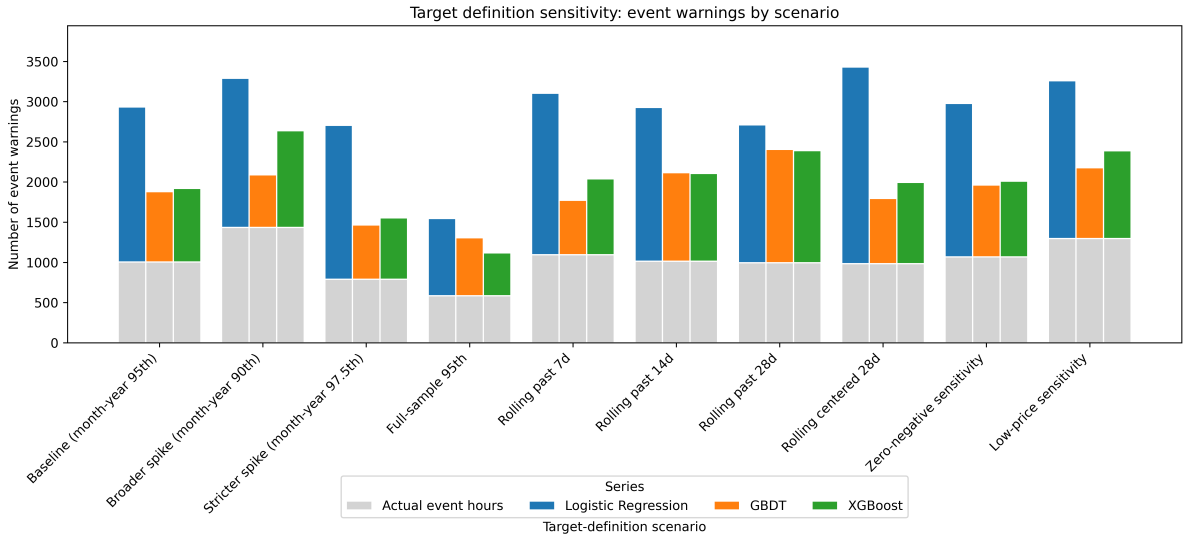

Figure 6.5: Target-definition sensitivity of event-warning volume. The grey part indicates the number of realised extreme hours in each scenario, while the total bar height indicates the number of model warnings

Figure 6.5 shows that target-definition choices also affect warning volume. Broader price-spike and low-price definitions generally increase both the number of realised event hours and the number of model warnings. Logistic regression produces the largest warning volume in most scenarios, while XGBoost and GBDT usually produce fewer warnings for similar or better macro F_2 performance.

Overall, the sensitivity analysis supports the use of the baseline target definition for the main results. The month-year 95th-percentile price-spike definition gives a sufficiently populated price-spike class in the final test period, avoids the instability of the full-sample threshold, and leads to model rankings that are robust to alternative definitions. The zero-negative and low-price checks show that extending the lower-tail threshold does not substantially change the modelling conclusion. Therefore, the baseline definition remains the most appropriate target definition for the main analysis.

6.5. Added value of cross-border indicators

This section evaluates whether cross-border indicators improve the ex-ante prediction of Dutch negative prices and price spikes compared with models using only domestic features. The feature-group scenarios follow the removal design introduced in Section 5.9. The results are interpreted as added predictive value compared with the domestic baseline.

Table 6.5 summarises the macro F_2 scores for the main feature-group scenarios. The domestic baseline already provides a meaningful starting point, with macro F_2 scores of 0.700 for GBDT and 0.702 for XGBoost. However, the performance of both tree-based models improves when cross-border information is added. The core cross-border feature set increases the macro F_2 score to 0.808 for GBDT and 0.813 for XGBoost. This indicates that cross-border variables contain information that is not fully captured by Dutch price-history, domestic system, weather, and calendar variables.

The results also show that not all cross-border feature blocks contribute equally. Scheduled exchange variables provide a small improvement for GBDT and XGBoost, while transfer-capacity variables alone add little predictive value. Neighbouring prices show mixed results, improving for XGBoost

Table 6.5: Feature-group removal results on the 2025 test set

Feature scenario	Features	Log. reg.	GBDT	XGBoost
Domestic baseline	32	0.625	0.700	0.702
+ Scheduled exchange	40	0.601	0.720	0.729
+ Transfer capacity	42	0.621	0.699	0.708
+ Neighbouring system	54	0.717	0.789	0.791
+ Neighbouring prices	36	0.627	0.690	0.727
+ Core cross-border	76	0.698	0.808	0.813
Full feature set	102	0.734	0.806	0.812

Note: Macro F2 scores. Log. reg. refers to logistic regression.

but not for GBDT. The clearest single-block improvement comes from neighbouring-system features, including neighbouring load, generation, renewable-generation, and residual-load forecasts. This suggests that expected system conditions in connected bidding zones are more informative than cross-border capacity indicators on their own.

The core cross-border feature set gives the largest meaningful improvement for the tree-based models. For GBDT, the macro F_2 score increases from 0.700 to 0.808. For XGBoost, it increases from 0.702 to 0.813. The full feature set performs almost the same as the core cross-border scenario, with macro F_2 scores of 0.806 for GBDT and 0.812 for XGBoost. This suggests that the most relevant cross-border information is already captured by scheduled exchange, transfer capacity, neighbouring-system forecasts, and neighbouring lagged prices. Adding the remaining neighbouring weather and calendar variables does not provide a clear further improvement, as the full feature set performs slightly below the core cross-border set. This is consistent with the PCA results discussed in Chapter 4, which indicated redundancy within the feature set but did not provide clear evidence for removing variables before modelling.

The results for logistic regression are more sensitive to changes in the included variables. This is expected, because logistic regression estimates one linear contribution per variable. It is therefore less able to ignore redundant variables or non-linear feature effects than the tree-based models. Some cross-border variables contain useful information, but may also overlap with domestic or neighbouring-system variables. When these variables are added separately, the additional noise or overlap can reduce the performance of logistic regression. The logistic regression results should therefore be interpreted as a linear benchmark. They show that cross-border information is useful, but also that its value depends more strongly on feature selection for a less flexible model.



Figure 6.6: Change in macro F_2 after removing one neighbouring bidding zone from the core cross-border feature set

Figure 6.6 shows the macro F_2 change after removing one neighbouring bidding zone from the core cross-border feature set. Negative values indicate that performance decreases when that neighbour is removed. The largest decrease occurs when DE/LU is removed. The macro F_2 drops by 0.040 for GBDT, 0.027 for XGBoost, and 0.053 for logistic regression. Removing BE, NO2, or DK1 has smaller effects. This indicates that DE/LU contributes most consistently to the predictive value of the cross-border feature set. The full removal results, including class-specific F_2 scores, precision-recall area under the curve (PR-AUC) values, probability quality metrics, and warning counts, are reported in Appendix E.

Overall, the removal results show that cross-border indicators add predictive value beyond domestic features. The improvement is not only driven by transfer capacity, but also by neighbouring market and system conditions. The tree-based models show the most stable improvements, while logistic regression is more sensitive to overlapping feature blocks because of its linear model structure. The leave-one-neighbour results show that this added value is not evenly distributed across neighbouring bidding zones, since DE/LU provides the strongest contribution. This supports the relevance of cross-border information for the ex-ante prediction of Dutch extreme price events.

6.6. Calibration and reliability

This section assesses whether the predicted probabilities are useful early-warning risk estimates. Figure 6.7 shows the reliability curves for negative-price and price-spike probabilities. The diagonal line represents perfect calibration, where predicted probabilities correspond to observed event frequencies. The curves are interpreted together with the probability-quality metrics in Table 6.2. The reliability curves show calibration per class, while log loss and the multiclass Brier score focus on the quality of the full probability distribution.

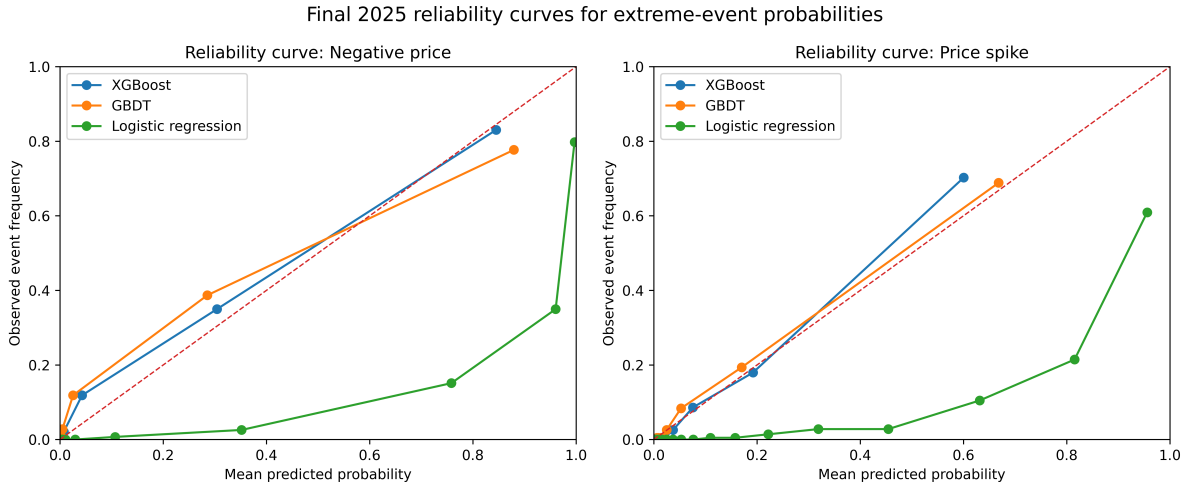

Figure 6.7: Reliability curves for negative-price and price-spike probabilities on the final 2025 test set

For negative-price probabilities, XGBoost and GBDT follow the diagonal more closely than logistic regression. XGBoost is closest to the diagonal at higher probabilities, while GBDT overestimates the highest probabilities. Logistic regression is less reliable, especially in the middle probability range.

For price-spike probabilities, XGBoost and GBDT again show the most useful reliability behaviour. Both models remain close to the diagonal in the lower and middle ranges, although XGBoost slightly underestimates the observed event frequency in the highest bin. Logistic regression shows clear overestimation across most probability bins, which is consistent with its high warning counts in Section 6.3.

The probability quality metrics in Table 6.2 support this visual interpretation. XGBoost obtains the lowest log loss and multiclass Brier score, with values of 0.161 and 0.095. GBDT performs close to XGBoost, with a log loss of 0.178 and a Brier score of 0.101. Logistic regression performs worse on both metrics, with a log loss of 0.533 and a Brier score of 0.314.

Overall, the results support XGBoost as the preferred final model. It combines strong event-detection performance with the best probability quality and useful reliability behaviour for both extreme-event classes.

6.7. Explainability insights

This section interprets the final XGBoost model using SHAP values. The SHAP analysis is performed on the final 2025 test-set predictions of the selected XGBoost model. The trained model is loaded from the stored model bundle and explained without retraining. Tree-based SHAP contributions are used, following the extension of SHAP methods to tree ensemble models (Lundberg et al., 2020). The purpose is to examine which variables contribute most to the model outputs for negative-price and price-spike risk. The values in this section are mean absolute SHAP values, so they indicate which features contribute most strongly on average within each event class.

Table 6.6 reports the highest-ranked SHAP features by mean absolute SHAP value for the final XGBoost model. The most important features for negative-price predictions are related to residual-load conditions, neighbouring generation, and scheduled cross-border exchange. The highest mean absolute SHAP value is from the DE/LU residual-load forecast, followed by the NO2 generation forecast and Dutch residual load. Residual load combines forecast demand with wind and solar generation, so its high importance suggests that the model uses expected surplus or tightness as an indicator for negative-price risk.

Table 6.6: Top SHAP features for the final XGBoost model

Rank	Negative-price feature	Mean |SHAP|	Price-spike feature	Mean |SHAP|
1	DE/LU residual load forecast	1.232	BE residual load forecast	0.746
2	NO2 generation forecast	0.585	DE/LU residual load forecast	0.505
3	NL residual load forecast	0.499	Week lagged price (EUR/MWh)	0.263
4	BE residual load forecast	0.437	BE solar forecast	0.225
5	NL scheduled export to NO2	0.182	Hour of the day	0.158
6	DK1 residual load forecast	0.123	NL residual load forecast	0.147
7	NL scheduled export to DK1	0.100	Day of the week	0.106
8	DK1 wind onshore forecast	0.091	DK1 solar forecast	0.097
9	NL load forecast	0.091	NL solar forecast	0.085
10	NL scheduled import from DK1	0.088	Same hour day before spike	0.078

For price-spike predictions, the highest-ranked features are residual-load indicators, lagged price behaviour, solar forecasts, and calendar timing. BE and DE/LU residual-load forecasts are the two most important price-spike features, followed by the week-lagged Dutch price. This supports modelling negative prices and price spikes as separate classes, since the two extremes rely on different feature patterns.

The SHAP results also support the cross-border findings from Section 6.5. DE/LU and BE residual-load forecasts appear among the most important features, while neighbouring-system and exchange variables are also included in the negative-price ranking. The highest-ranked features are concentrated in system, price-history, and cross-border indicators. This is in line with the PCA results in Chapter 4, which showed redundancy but did not support removing features before modelling. After keeping the full feature set, the SHAP results show which features the final model uses most strongly in its predictions.

Overall, the explainability results support the interpretation that negative-price and price-spike risks are driven by different model signals. The full SHAP summary plots for the final XGBoost model are provided in Appendix F.

6.8. Cost-sensitive decision interpretation

This section interprets the final model from a decision-support perspective. The results are interpreted using the cost-sensitive framework introduced in Section 5.11.

For negative-price risk, XGBoost detects most realised events. Table 6.2 shows a negative-price recall of 0.969 and a precision of 0.459. This means that only a small number of negative-price hours are missed, while a substantial share of negative-price warnings still occur during normal-price hours. Figure 6.2 shows the same trade-off: XGBoost misses 18 negative-price hours, but also classifies 652 normal-price hours as negative-price warnings. This warning behaviour can be useful when missed negative prices, missed charging opportunities, or missed export-adjustment opportunities are costly.

For price-spike risk, XGBoost has a recall of 0.845 and a precision of 0.521. This indicates a more balanced warning structure than for negative prices, although missed price spikes remain present. Figure 6.2 shows that XGBoost misses 67 price-spike hours and classifies 336 normal-price hours as price-spike warnings. This type of warning behaviour is relevant when missed high prices are costly, for example for consumers, retailers, or traders with open positions. If false alarms are costly, a higher operational warning threshold may be preferred.

The appropriate threshold depends on the user's cost ratio. If missed events are more costly than false warnings, a lower warning threshold is justified. If false warnings lead to costly or unnecessary actions, a higher threshold is more appropriate. The validation-tuned thresholds should therefore be adapted to the user context before being used operationally.

Overall, the cost-sensitive interpretation shows that the final XGBoost model is most useful as a probabilistic decision-support tool. Its probabilities can help identify hours that require closer attention before market closure, but the final action threshold should depend on the decision context and the relative costs of missed events and false warnings.

6.9. Conclusion of the results chapter

This chapter presented the results of the early-warning modelling framework. The final 2025 test set remains imbalanced, with normal-price hours forming the majority class and negative prices and price spikes forming smaller event classes. This confirms that overall accuracy alone is not suitable for model selection. Class-specific performance, probability quality, calibration, and warning behaviour are needed to evaluate the models.

Table 6.7 provides a compact qualitative synthesis of the model results across the four model-comparison KPIs. The table shows how each model performs on detection of relevant extreme events, limitation of unnecessary warnings, useful probability estimates, and support for actor-specific decision-making.

Table 6.7: Qualitative synthesis of model performance across the four model-comparison KPIs

KPI	Operationalised as	Dummy baseline	Logistic regression	PNN	MLP	GBDT	XGBoost
Detection of relevant extreme events	Macro F_2 Balanced accuracy Negative-price recall Price-spike recall	− −	+	0	0	+	+ +
Limitation of unnecessary warnings	Class-specific precision False positives Warning counts	− −	−	−	0	+	+
Useful probability estimates	Log loss Multiclass Brier score Calibration Reliability	− −	−	−	0	+	+ +
Support for actor-specific decision-making	Error structure Threshold behaviour Cost-sensitive interpretation	− −	0	−	−	+	+
Overall synthesis		− −	**0**	−	**0**	**+**	**+ +**

Note: − − very weak performance, − weak performance, 0 mixed performance, + good performance, and + + very strong performance.

XGBoost and GBDT provide the strongest overall results. As shown in Table 6.7, XGBoost is selected as the preferred final model because it provides the strongest overall combination of event detection, probability quality, calibration behaviour, and warning volume. Other models score high on underlying metrics, as logistic regression and PNN achieve high event recall, and MLP obtains strong negative-price precision. However, these models have weaker probability quality, lower precision for one or both event classes, or too many false warnings. GBDT remains a close benchmark, while logistic regression is useful as an interpretable baseline but produces substantially more false warnings.

The sensitivity, reverse-year, and removal analyses support the robustness of this conclusion. The target-definition sensitivity analysis shows that the main model ranking is stable across most alternative definitions. The reverse-year robustness check shows that the tree-based models remain the strongest model group when the year order is reversed. The feature-removal results show that cross-border indicators add predictive value beyond domestic features, especially through neighbouring-system and market information. The leave-one-neighbour results further show that DE/LU provides the strongest contribution.

Finally, the calibration, explainability, and cost-sensitive interpretation results show that the final XGBoost model is most useful as a probabilistic decision-support tool. Its probabilities can support early-warning decisions before market closure, but the appropriate operational warning threshold depends on the user's relative costs of missed events and false warnings. These findings form the basis for the discussion in Chapter 7.

7

Discussion

This chapter interprets the results from Chapter 6 in relation to the research objective, the Dutch day-ahead market context, and the literature reviewed in Chapter 2. Chapter 6 showed that tree-based models provide the strongest early-warning performance, that the selected target definition is broadly robust, and that cross-border indicators add predictive value beyond domestic features. The discussion chapter first explains what these findings mean for model performance and practical decision-making. It then compares the results with the literature, reflects on the main limitations, discusses potential feedback effects from wider model use, and identifies directions for future research and practical development. Together, these interpretations provide the basis for the direct answers to the sub-questions and main research question in Chapter 8.

7.1. Interpretation of model performance

The results show that the prediction of Dutch extreme day-ahead prices should not be evaluated as a standard classification problem. As shown in Section 6.1, the target variable is strongly imbalanced, with normal-price hours forming the large majority of the observations. Overall accuracy is therefore not sufficient for model selection, because a model can perform well on normal hours while still failing to detect negative prices or price spikes. The interpretation of model performance should therefore focus on class-specific recall and precision, macro F_2, balanced accuracy, probability quality, and warning behaviour, as defined in Chapter 5.

The strongest overall results are obtained by XGBoost and GBDT. These models perform better than the linear and neural-network alternatives in the final 2025 test year. Their stronger performance suggests that the prediction task benefits from models that can capture non-linear relations between domestic system variables, residual-load indicators, neighbouring market conditions, scheduled exchange, and calendar patterns. This fits the structure of the problem, since negative prices and price spikes arise from combinations of expected supply, demand, renewable generation, cross-border exchange possibilities, and temporal market conditions.

The comparison in Section 6.3 also shows a clear trade-off between event detection and warning volume. Logistic regression and PNN detect many extreme events, but issue more warnings and therefore produce more false positives. XGBoost provides a better balance between detecting realised events and limiting unnecessary warnings. For market use, this distinction is important, because a warning system that produces too many false alarms can become difficult to interpret.

The confusion matrices show that the main models do not confuse negative prices with price spikes. The main error is instead between the normal class and the two extreme classes. This supports the decision to model negative prices and price spikes as separate target classes. The remaining challenge is not to distinguish the two extremes from each other, but to separate both extremes from normal-price hours for early-warning purposes.

The results also indicate that price spikes are more difficult to predict than negative prices. Negative-price events are strongly detected by the final XGBoost model, while price-spike prediction remains more uncertain. One reason is that negative prices have a clear market-based definition: prices below 0 €/MWh. Price spikes depend on a chosen threshold that separates the upper tail from normal prices.

This makes the price-spike class sensitive to the labelling strategy and to changes in the price distribution over time. In addition, negative prices may be directly linked to observable surplus conditions, such as low residual load and high renewable generation. Price spikes may depend more on scarcity conditions, unexpected constraints, or market information that is not fully captured in the open ex-ante feature set.

The target-definition sensitivity analysis supports the robustness of the main model conclusion. XGBoost and GBDT remain the strongest models across most alternative target definitions. The full-sample 95th-percentile threshold is the exception, because it leaves only a few price-spike events in the final test year. The use of the month-year 95th-percentile definition provides a stable event class across changing market conditions, supporting its use as the main price-spike definition.

Overall, XGBoost is selected as the preferred final model because it combines strong event detection, good probability quality, and a manageable warning volume. GBDT remains a close benchmark, while logistic regression remains valuable mainly as an interpretable baseline. This interpretation is used in Chapter 8 to answer the sub-question on probabilistic model development and comparison.

7.2. Implications for Dutch market participants

The model outputs estimate the probability that an hour belongs to one of three classes: negative price, normal price, or price spike. They should therefore be interpreted as risk signals rather than exact price forecasts. This makes the model relevant for decisions that need to be prepared before day-ahead market closure, when market participants still have to act under uncertainty.

The practical implications differ per type of market participant. For an energy-intensive industrial consumer, such as a steel factory, high price-spike probabilities could indicate hours with increased exposure to high electricity prices. If part of the production process is flexible, the model could support decisions to shift non-critical electricity demand to other hours and increase internal risk buffers. Negative-price probabilities could indicate hours when flexible demand becomes more attractive, although the actual response would depend on operational constraints, production schedules, and contract structure.

For electricity producers, negative-price probabilities provide a supply-side warning. A high probability of a negative-price hour may indicate that continued generation could reduce revenues or cause losses, depending on marginal costs and contract positions. Producers with sufficient flexibility could use this signal to prepare production or bidding adjustments, such as reducing volumes or planning to ramp down output.

For a battery operator, the probabilities can be linked directly to charging and discharging decisions. High negative-price probabilities point to hours in which charging could become financially attractive, while high price-spike probabilities indicate hours in which stored electricity may have higher value. However, the model should not be interpreted as an optimisation model for battery dispatch. It should be combined with specific constraints before being used for actual battery dispatch decisions.

For a trader or retailer, the probabilities can support monitoring, bidding, and risk-management decisions before market closure. A high price-spike probability may justify closer inspection of open positions or additional hedging attention. A high negative-price probability may signal potential downside risk for generation positions or opportunities for flexible demand. The model is then useful as a prioritisation tool by helping to identify which hours require human attention or additional analysis before bids are submitted.

The target-definition sensitivity results also show that the most relevant event definition can be adapted to different market actors. The baseline definition is appropriate for the main modelling comparison because it provides a consistent lower-tail and upper-tail event structure, but alternative definitions can be more relevant in specific operational contexts. For renewable producers, zero or very low prices may already be important even when the price is not strictly negative, because these hours reduce revenues and may affect curtailment or bidding decisions. For battery operators and flexible consumers, the broader low-price definition may be useful because prices just above zero can still represent attractive charging or demand-shifting opportunities. For traders and retailers, the relevant price-spike definition may depend on portfolio exposure and risk tolerance: a broader price-spike threshold can support earlier monitoring, while a stricter threshold may be more relevant for high-impact scarcity events. The sensitivity results therefore show that the model framework can be adapted to different practical definitions of market risk.

These examples show why the cost-sensitive interpretation from Section 5.11 is important. The validation-tuned warning thresholds are useful for comparing models, but should not be treated as universal operational thresholds. A steel factory with limited flexibility may only act when the predicted probability is high, because unnecessary production changes are costly. A battery operator may use lower thresholds, because it can often respond more flexibly. A trader may use different thresholds depending on portfolio exposure and the cost of false alarms. The appropriate threshold therefore depends on the relative cost of false positives and false negatives for the user.

The practical contribution of the model is not a single fixed warning rule. Its value lies in providing a probability-based warning structure that can be adapted to different market actors. The same probability output can support different decisions, depending on the user's flexibility options and risk preferences. This interpretation is used in Chapter 8 to answer the sub-question on error costs and decision implications.

7.3. Comparison to literature

The results are in line with the literature reviewed in Chapter 2. First, they support the distinction between negative prices and price spikes as separate tail events. Negative prices are often linked to surplus conditions and limited downward flexibility, while price spikes are more related to scarcity conditions and limited upward flexibility (De Vos, 2015; Liu et al., 2022). The model results support this distinction, since negative-price and price-spike predictions rely on different feature patterns and show different levels of predictive difficulty.

The importance of residual-load variables is consistent with the price-formation mechanisms discussed in the literature. Residual load combines expected demand with expected wind and solar generation. Low residual load indicates that demand is small relative to renewable production, which can create surplus conditions and increase the probability of negative prices. High residual load indicates tighter system conditions, which can increase the probability of price spikes. The importance of residual-load features in the final model therefore supports the interpretation that both tails are connected to expected system stress.

This interpretation is also consistent with the merit-order effect, which describes how low-marginal-cost wind and solar generation can shift the supply curve, pushing out more expensive generation and lowering the market-clearing price (De Vos, 2015; Sensfuß et al., 2008). In coupled electricity markets, this effect is not limited to domestic conditions, since neighbouring renewable output and residual-load conditions can influence regional price formation through cross-border market coupling (De Blauwe et al., 2025). This mechanism is not estimated directly, as the full market-clearing process is not reproduced. However, the importance of residual-load and renewable-related features suggests that the model captures conditions under which the merit-order effect can contribute to low or negative prices. This is especially relevant for the lower-tail predictions. High residual load can indicate the opposite situation, where more expensive generation or imports are needed and price-spike risk increases.

The relevance of lagged prices and recent event indicators is in line with earlier work on extreme electricity prices. Previous studies show that extreme price events can show signs of persistence or clustering over time (Amjady & Keynia, 2011; Lu & Suthaharan, 2023). In this thesis, price-history variables are mainly relevant for price-spike prediction. This suggests that recent market conditions contain useful information for identifying extreme-price risk.

The results also support the use of probabilistic early-warning models rather than point forecasts. Probabilistic forecasts are useful in volatile electricity markets because they represent uncertainty and can be linked to risk-based decisions (Haben et al., 2021). This is reflected in the final model. The probability outputs are not only used for classification, but also for warning-threshold selection, calibration assessment, and cost-sensitive interpretation.

The model comparison also qualifies the role of PNN in the reviewed literature. Loizidis et al. (2025) find strong PNN performance for negative-price classification in the German day-ahead market. In this thesis, however, XGBoost performs better overall. This result can be explained by differences in the modelling setting. The evaluation in this thesis is based on a Dutch multiclass early-warning problem. In addition to event detection, it considers probability quality, calibration, and warning behaviour. With this broader evaluation, the strong performance of XGBoost is not unexpected. Its tree-based structure can capture non-linear interactions between domestic and cross-border indicators, while the results show that it also limits unnecessary warnings.

Finally, the cross-border results are consistent with the literature on market coupling and neighbouring market effects. Previous studies show that coupled electricity markets can share price drivers through renewable generation, load, interconnection capacity, and scheduled exchange (De Blauwe et al., 2025; Ovaere et al., 2023; Unger et al., 2018). The feature-removal results support this mechanism, since cross-border indicators improve prediction compared with a domestic baseline. The strongest contribution comes from neighbouring-system and market variables, especially from Germany/Luxembourg. Dutch extreme prices are therefore not only shaped by domestic conditions, but also by market conditions in connected bidding zones.

Overall, the findings support the main positioning of this thesis. Dutch extreme day-ahead prices can be interpreted as ex-ante tail-risk events shaped by domestic market stress, residual-load conditions, and cross-border market coupling. The results therefore connect the literature on extreme prices, probabilistic forecasting, and cross-border electricity markets in one modelling framework.

7.4. Limitations

This research has some limitations that should be considered when interpreting the results. The modelling period covers 2022–2025, with 2025 used as the final out-of-sample test year. This provides a useful test under unseen market conditions, but it still represents only one final test year. The period also includes market conditions associated with the European energy crisis and a rapid increase in renewable generation, meaning that future market regimes may differ from the patterns observed in the data.

Second, the target definition remains a modelling choice. Negative prices have a clear market-based definition, but price spikes require a chosen threshold. The month-year 95th-percentile threshold is supported by the sensitivity results, but alternative definitions may be more suitable for specific market participants. The treatment of exactly zero and low positive prices is also a limitation, since these hours are treated as normal prices in the main model even though they can reflect surplus conditions.

Third, the model is built at an hourly resolution. This matches the historical dataset used in this thesis, where all variables are aligned to hourly Dutch day-ahead delivery hours. However, EPEX SPOT introduced 15-minute market time units in SDAC on 1 October 2025 (EPEX SPOT, n.d.-a). An operational version of the model would therefore need to be redesigned for 15-minute products. Variables that are only available hourly may miss intra-hour variation, while interpolating these values could suggest a level of detail that is not actually present in the data. The current model should be interpreted as an hourly early-warning model, not as a directly usable 15-minute forecasting tool.

Fourth, the feature set is limited by open-data availability and the ex-ante constraint. The model only uses variables that could be obtained from open sources and aligned with publication timing before day-ahead market closure. This supports reproducibility, but excludes potentially relevant information that was not publicly available, not available before market closure, or not available in a consistent format. Not all intended Great Britain variables could be included. GB weather and calendar variables are included, but consistent GB price-history, system, scheduled-exchange, and transfer-capacity variables could not be obtained for the final feature set. The cross-border results therefore reflect the neighbouring-market information included in the dataset, rather than the influence of all Dutch interconnections.

Fifth, there is overlap between weather variables and energy forecast variables. Load, wind-generation, solar-generation, and residual-load forecasts already contain part of the expected weather effect. Separate weather variables may add limited independent information, even though weather remains an important underlying driver of demand and renewable generation. Low feature importance for individual weather variables should therefore not be interpreted as evidence that weather conditions are unimportant. Part of the weather effect may already enter the model through the system forecast variables, which makes the feature-importance results difficult to interpret causally.

Sixth, the analysis is limited to the day-ahead price outcome. The model identifies hours with increased negative-price or price-spike risk, but it does not estimate the financial consequences of these outcomes after day-ahead clearing. In practice, these financial consequences also depend on intraday and balancing-market outcomes, the participant's contracts, hedging strategy, flexibility options, and exposure to the day-ahead market. The predicted probabilities should therefore be interpreted as signals of day-ahead market risk, rather than as direct estimates of realised profit, loss, or imbalance costs.

Finally, the decision interpretation remains indicative. The validation-tuned warning thresholds are useful for model comparison, but they are not actor-specific operational thresholds. The calibration results show that the predicted probabilities should be interpreted as risk signals rather than guaranteed event frequencies. The examples of a steel factory, battery operator, and trader show how the model could support decision-making, but no economic assessment is performed for these actors. The model should therefore be seen as a decision-support tool, not as an automatic trading, hedging, or dispatch rule.

7.5. Potential feedback effects

The model is evaluated on historical data, where the predicted probabilities cannot influence the realised day-ahead price. However, this could change if the same type of early-warning tool became available to many market participants. The model output would then not only be an estimate of market risk. It could become a common signal that influences bidding, scheduling, hedging, and flexibility decisions before market closure. This is related to performative prediction, where predictions that are used for decision-making can influence the outcomes they are intended to predict (Perdomo et al., 2020).

This creates a potential market-feedback mechanism, because day-ahead prices are formed through supply and demand bids in the market-clearing process. Market participants with controllable generation assets already use analytical methods when preparing day-ahead bids, and adapted bids can change the resulting bid curves (Klæboe et al., 2022). If early-warning signals become part of these decision-making processes, they may influence bidding behaviour in similar directions. Resulting changes in bid curves could thereby contribute to changes in market outcomes. A high price-spike probability may lead to many actors taking action for the same expected scarcity hour. Traders may adjust bids, retailers may hedge open positions, batteries may reserve capacity for discharge, and flexible consumers may shift demand. Individually, these responses are rational. Together, they may concentrate behaviour around the same predicted risk period and change price pressure in that specific hour or neighbouring hours.

A similar mechanism can occur for negative-price warnings. A high negative-price probability may encourage battery charging, flexible demand, export-oriented bidding, or production adjustments. These responses can absorb a surplus, reducing the likelihood of a negative price. However, when many actors react in the same direction, the response can also shift demand or supply patterns and create new price effects around the warned period. The effect of a warning can therefore go in two directions: it can reduce the predicted event, but it can also contribute to new price pressure when many actors respond in the same way.

If the model is used to support market decisions, its outputs become part of the system they are intended to predict. Its value then depends not only on predictive performance, but also on how participants respond to the probabilities. Wider use would require additional monitoring to determine if the warnings continue to support better decisions, or if they start to create similar market responses. The model should then be treated as a decision-support tool within the market system, rather than as an external observer of that system.

7.6. Recommendations and future work

The results of this thesis show that probabilistic early-warning models can support the anticipation of Dutch day-ahead extreme prices, but several steps are needed before such a model can be used in practice. A first step is to evaluate the model over a longer time period. The current evaluation is based on one final out-of-sample test year, as discussed in Section 7.4. Testing the model on a longer period would make it possible to assess whether the results remain stable across different fuel-price regimes, renewable-generation levels, demand patterns, and cross-border market conditions.

Future work could extend the feature set by addressing remaining data limitations. The current model is based on open ex-ante data, which supports reproducibility, but limits the set of variables that can be included. More complete cross-border constraint data, intraday and balancing-market indicators, and Great Britain price, system, scheduled-exchange, and transfer-capacity variables would make the cross-border assessment more complete. Fossil-fuel and carbon-market variables, such as gas, coal, and CO_2 prices, could also be included if ex-ante publication timing can be verified. These variables may be especially relevant for price spikes, since they influence the marginal cost of dispatchable generation and therefore the conditions under which scarcity is priced (Gasparella et al., 2023).

Another important step is to adapt the modelling framework to the 15-minute day-ahead market structure discussed in Section 7.4. The current model is built around hourly delivery periods and should therefore not be treated as directly transferable to 15-minute products. Future research could assess which variables are available at 15-minute resolution, how hourly variables can be incorporated without overstating their level of detail, and whether the current target definition remains suitable for shorter market time units.

The probability outputs also require further development. Section 6.6 showed that XGBoost provides the strongest probability quality among the evaluated models, although its predicted probabilities do not always match observed event frequencies exactly. Additional calibration methods could therefore be tested, especially when the model is used to support actor-specific decisions. This should be combined with economic testing. Section 7.2 discussed how the same probability output could support different decisions for a steel factory, battery operator, or trader. However, this thesis does not test whether acting on the warnings would create financial value for these actors. Future work could therefore link the warnings to contracts, asset constraints, flexibility costs, and trading strategies, and compare the model with simpler strategies.

Finally, practical deployment should be tested in a real-time setting. This would require monitoring publication delays and API failures, and updating the model when market conditions change. It would also make it possible to evaluate the feedback effects discussed in Section 7.5. The next step is not only to improve the model technically, but also to test how it performs as part of an operational decision-support process, in which probabilistic warnings support expert judgement.

8

Conclusion

This chapter concludes the thesis by answering the research questions and summarising the main contribution of the developed early-warning framework. It first answers the five sub-questions separately, after which the main research question is answered. The chapter then summarises the key contributions, describes the practical deliverable, and closes with a final reflection on the value and limitations of probabilistic early-warning models for extreme Dutch day-ahead electricity prices.

8.1. Answer to the sub-questions

This section answers the five sub-questions that structured the thesis. Together, the answers show how the target definition, ex-ante feature set, model comparison, cross-border feature analysis, and cost-sensitive interpretation contribute to the final early-warning framework.

8.1.1. Sub-question 1 – Extreme event definition

Which threshold definition and labelling strategy is most appropriate for negative prices, normal prices, and price spikes in the Dutch day-ahead electricity market, taking price distributions, seasonality, and sensitivity to alternative definitions into account?

The answer depends on the conceptual meaning of the lower and upper price tails, as well as the robustness of the results under alternative definitions. Negative prices are most appropriately defined as hours in which the Dutch day-ahead price is below 0 €/MWh. Price spikes require a separate threshold, because there is no fixed boundary above which a price automatically becomes a price spike. Price spikes are therefore defined as prices above the month-year-specific 95th percentile of the Dutch day-ahead price. This definition accounts for seasonal and yearly differences in the price distribution while keeping the target variable transparent and reproducible.

The sensitivity analysis supports this choice. The main model ranking remains stable across most alternative definitions, and the rolling-threshold checks remain close to the baseline. This indicates that the results are not strongly dependent on the exact upper-tail threshold construction. At the same time, the rolling thresholds do not provide a clear improvement over the month-year 95th-percentile definition. The baseline definition is retained as the most suitable target definition for this thesis, since it provides a transparent and stable benchmark for comparing models across the full evaluation period. Rolling thresholds remain relevant as an operational robustness check, since they can be calculated from recently observed prices, whereas the full month-year threshold is mainly suitable for retrospective labelling. Only the full-sample 95th-percentile threshold leaves too few price-spike events in the 2025 test set for a stable evaluation. The zero-negative and low-price checks also show that broadening the lower-tail definition does not substantially change the modelling conclusion. The baseline definition is therefore suitable for the main analysis, while alternative lower-tail and rolling-threshold definitions may be useful for specific operational users.

8.1.2. Sub-question 2 – Ex-ante feature design

Which domestic, weather, cross-border, and calendar indicators can be operationalised as ex-ante predictors of Dutch day-ahead extreme price events, given the constraint that all features must be available before market closure?

The feature set was constructed by translating the mechanisms from the literature review into variables that were available before day-ahead market closure. The ex-ante feature set consists of Dutch price-history variables, domestic system variables, weather variables, cross-border variables, and calendar controls. These feature groups represent the mechanisms identified in the literature: domestic market stress, flexibility stress, and cross-border transmission of scarcity or surplus. The variables include lagged prices, recent event indicators, load and renewable-generation forecasts, residual-load indicators, weather forecasts, scheduled exchange, transfer capacity, neighbouring-system forecasts, and recurring hour and day of the week patterns.

The ex-ante feature design is therefore feasible, since all selected indicators can be operationalised using information available before day-ahead market closure. This means that historical Dutch prices are included only as lagged variables, while system conditions are represented through indicators available before the market outcome is known. Variables that would only be known after day-ahead market closure are excluded. The final modelling dataset contains 34,453 hourly observations and 102 explanatory features. This shows that an open-data feature set can be operationalised under the ex-ante constraint.

8.1.3. Sub-question 3 – Probabilistic model development and comparison

How can candidate probabilistic machine-learning models be evaluated and compared for estimating hourly probabilities of negative prices, normal prices, and price spikes in an ex-ante setting, and which model provides the most useful early-warning performance?

The model comparison was designed around the intended use of the model as an early-warning tool. Candidate models were developed as multiclass classifiers that estimate hourly probabilities for negative prices, normal prices, and price spikes. The models were trained, tuned, and evaluated using a chronological validation strategy. In this strategy, 2022–2023 are used for model development, 2024 is used for validation and threshold tuning, and 2025 is used as the final test year. Hyperparameters and warning thresholds were selected on the validation set. The selected model configurations were retrained on 2022–2024 and evaluated on the unseen test period. This ensures that the comparison reflects the ex-ante early-warning setting.

Because negative prices and price spikes are rare compared to normal-price hours, standard accuracy is not sufficient for model selection. A model could achieve high accuracy by mainly predicting the normal class, while still missing the events that are most relevant for decision-making. The candidate models were therefore evaluated on their ability to detect both extreme classes, limit unnecessary warnings, and produce probabilities that can be interpreted as risk signals. This required a combination of class-specific recall and precision, macro F_2, balanced accuracy, warning behaviour, probability-quality metrics, and calibration.

Based on this evaluation, XGBoost provides the strongest overall early-warning performance in the final 2025 test set. It offers the strongest balance between detecting negative-price and price-spike events, limiting the number of unnecessary warnings, and producing reliable probability estimates. GBDT performs similarly and remains a close benchmark, which suggests that tree-based models are well suited to this prediction problem. The tree-based models also remain the strongest when the year order is reversed, supporting the robustness of this conclusion. Logistic regression and PNN detect many event hours, but they do so by issuing substantially more false warnings. The MLP provides a more moderate alternative, but its lower overall event-detection performance and weaker probability quality make it less suitable than the tree-based models. XGBoost is therefore selected as the preferred final model.

8.1.4. Sub-question 4 – Added value of cross-border indicators

To what extent do cross-border coupling, directional scheduled exchange, day-ahead net transfer capacity, and neighbouring market features improve the ex-ante prediction of Dutch negative prices and price spikes compared with models that only use domestic features, and which neighbouring feature groups and bidding zones contribute most?

The added value of cross-border information was assessed by comparing domestic-only models with models that include different neighbouring-market and interconnector feature groups. Cross-border indicators improve the ex-ante prediction of Dutch negative prices and price spikes compared with a domestic-only feature set. For XGBoost, the macro F_2 score increases from 0.702 with the domestic baseline to 0.813 with the core cross-border feature set. For GBDT, it increases from 0.700 to 0.808. Dutch extreme-price prediction should therefore not be treated as a purely domestic problem.

The strongest cross-border contribution comes from neighbouring-system features, such as neighbouring load, generation, renewable-generation, and residual-load forecasts. This is supported by the explainability results, where residual-load forecasts are among the highest-ranked features for both negative-price and price-spike predictions for XGBoost. Transfer capacity alone adds little predictive value, while scheduled exchange provides a smaller improvement. The leave-one-neighbour analysis shows that DE/LU contributes most consistently to the predictive value of the cross-border feature set. Cross-border information, especially neighbouring-system and market information from DE/LU, should therefore be included in the final early-warning framework.

8.1.5. Sub-question 5 – Evaluation, error costs, and decision-support interpretation

How should the estimated class probabilities be interpreted as decision-support signals for market participants, considering asymmetric false-positive and false-negative costs?

The probabilities were interpreted by linking warning behaviour to the asymmetric consequences of missed events and false warnings. The estimated class probabilities should be interpreted as decision-support signals, not as automatic trading, hedging, or dispatch rules. The final model identifies hours with increased negative-price or price-spike risk before market closure, but the appropriate response depends on the user's position, flexibility options, contracts, and risk preferences.

The cost-sensitive interpretation indicates that the relevant warning threshold depends on the relative costs of false positives and false negatives. If missed events are more costly than false warnings, a lower threshold may be appropriate. If false warnings lead to costly or unnecessary actions, a higher threshold is more suitable. The validation-selected thresholds are therefore useful for model comparison and warning-behaviour analysis, but actor-specific operational thresholds require additional information about costs, constraints, and decision context.

8.2. Answer to the main research question

How can ex-ante probabilistic machine-learning models be developed and evaluated to estimate hourly probabilities of negative prices, normal prices, and price spikes in the Dutch day-ahead electricity market, using domestic, weather, and cross-border indicators available before market closure, and how can these probabilities support market decision-making under asymmetric error costs?

Ex-ante probabilistic ML models can be developed through a reproducible early-warning framework. The first step is to translate Dutch day-ahead prices into a three-class target variable. Negative prices are defined as prices below 0 €/MWh. Price spikes are defined as prices above the month-year-specific 95th percentile of the Dutch day-ahead price. Normal prices are all remaining prices. The second step is to link this target to a feature set that represents the mechanisms of domestic market stress, flexibility stress, and cross-border transmission of scarcity or surplus. This feature set includes Dutch price-history variables, domestic system forecasts, weather variables, cross-border indicators, neighbouring-system variables, and calendar controls. The third step is to filter these variables according to the ex-ante constraint, so that only information available before day-ahead market closure is included. Together, these steps define the ex-ante input and target structure for the development and evaluation of the models.

The models can then be evaluated as early-warning models rather than as standard classification models. This is done using a chronological train-validation-test design, with model tuning and threshold selection performed before final evaluation on an unseen test year. Because negative prices and price spikes are rare, overall accuracy is not sufficient. A model can perform well on normal-price hours while missing the events that matter most for decision-making. The evaluation therefore combines class-specific recall and precision, macro F_2, balanced accuracy, warning behaviour, probability-quality metrics, calibration, target-definition sensitivity, cross-border feature-removal analysis, and a reverse-year robustness check. Based on this evaluation, XGBoost provides the strongest overall

early-warning performance. It combines strong detection of negative-price and price-spike events with strong probability quality and a more limited warning volume than more aggressive models, such as logistic regression and PNN. GBDT remains a close benchmark, suggesting that tree-based models are well suited to this prediction problem.

The framework also tests whether Dutch extreme-price prediction benefits from including cross-border features. This is done by comparing models that only use domestic features with models that include cross-border indicators. The analysis assesses which neighbouring feature groups and bidding zones contribute most. Domestic system variables and price-history features provide a necessary baseline, while weather variables provide information on demand and renewable-generation conditions. At the same time, part of the weather effect is already captured in load, wind, solar, and residual-load forecasts. The results show that cross-border indicators improve performance beyond the domestic baseline, especially through system and market information from neighbours. DE/LU provides the strongest and most consistent contribution. The explainability results from the final XGBoost model indicate that it captures market-stress conditions linked to residual load, renewable generation, and cross-border coupling. This includes conditions under which the merit-order effect can contribute to low or negative prices.

Finally, the estimated probabilities can support market decision-making by identifying hours with increased negative-price or price-spike risk before market closure. The resulting risk signals indicate when closer monitoring or additional preparation may be required. This includes possible adjustments to bids, flexibility use, or risk buffers. However, the probabilities should not be interpreted as automatic trading or dispatch rules. Their operational value depends on the decision context, because the costs of missed events and false warnings differ between actors. The appropriate warning threshold should therefore be adapted to the user's flexibility options, contracts, and risk preferences.

8.3. Key contributions

This thesis contributes to the literature on electricity price forecasting by translating the Dutch extreme-price warning problem into a reproducible ex-ante modelling framework. The first step in this contribution is the construction of a dataset that combines Dutch price-history variables, domestic system forecasts, weather variables, cross-border indicators, and calendar controls. By documenting whether variables are available before day-ahead market closure, the dataset supports the development of early-warning models that do not rely on information that is only known after the market has closed.

Building on this dataset, the thesis contributes methodologically by modelling negative prices, normal prices, and price spikes within one multiclass probabilistic framework. This makes it possible to treat lower-tail and upper-tail price events as separate outcomes, while still evaluating them in the same early-warning model. The model comparison shows that overall accuracy is not a suitable basis for selecting an early-warning model, because it can hide poor detection of rare extreme-price events. Model evaluation should therefore combine class-specific performance with warning behaviour, probability-quality assessment, calibration, and robustness to alternative target definitions.

The modelling framework is then used to test whether Dutch extreme-price prediction should be treated as a domestic or regional problem. The feature-removal analysis shows that cross-border indicators add predictive value beyond domestic information, especially through neighbouring-system and market variables. DE/LU provides the strongest neighbouring contribution, supporting the conclusion that Dutch negative prices and price spikes are shaped by both domestic market conditions and regional market coupling.

Finally, the thesis connects probabilistic forecasting to decision-support interpretation. The predicted probabilities are treated as risk indicators that can be interpreted through the trade-off between false positives and false negatives. This links the modelling results to practical decision-making and shows how the same probability output can support different actions depending on an actor's exposure, flexibility options, contracts, and risk preferences.

From a CoSEM perspective, the contribution of this thesis is not limited to the predictive model itself. The thesis develops a decision-support framework for a complex electricity market in which technical system conditions, market arrangements, and actor-specific decisions are closely connected. The modelling pipeline translates open ex-ante market, weather, and cross-border data into hourly warning probabilities. The interpretation of these probabilities depends on the position, flexibility options, and risk preferences of different market participants. In this way, the thesis connects the technical forecast-

ing task to the broader decision context in which negative-price and price-spike risks are managed. The framework is relevant not only for individual market participants, but also for the broader challenge of decision-making in an increasingly renewable, congested, and interconnected electricity system.

8.4. Practical deliverable

The practical deliverable of this thesis is a reproducible modelling pipeline that can be used as the basis for an early-warning tool for Dutch day-ahead extreme prices. The pipeline combines open ex-ante data inputs, data retrieval and cleaning, target labelling, feature construction, model training and comparison, KPI-based evaluation, probability estimation, warning-threshold selection, feature explanation, and actor-specific interpretation. Its main output is an hourly probability estimate for each price class: negative price, normal price, and price spike. These probabilities can be translated into indicative warning levels for negative-price and price-spike risk before day-ahead market outcomes are known.

In practical use, this tool could present a next-day overview of hours with elevated negative-price or price-spike probabilities. To support decision-making, the most relevant model drivers for those hours could be shown to make the warning signals more interpretable. This would allow users to identify periods that require closer monitoring or additional preparation, such as reviewing bids, flexibility use, or risk buffers. The driver information can support interpretation by showing whether warnings are mainly related to residual load, renewable generation, cross-border conditions, or recent price patterns.

The performance and usefulness of the pipeline can be evaluated through the same KPI logic used in the model comparison. Event-detection metrics, probability-quality metrics, calibration, and warning behaviour indicate whether the model identifies relevant extreme-price hours while limiting unnecessary warnings. Before the pipeline can be used operationally, it would need to be tested in a live forecasting setting with automated data retrieval and continued monitoring over a longer out-of-sample period. Warning thresholds would also need to be adapted to the user context, since the costs of missed events and false warnings differ between market participants. Finally, the model would need to be linked to actor-specific decision rules and updated when market design or product structures change, especially if the application moves beyond the hourly day-ahead setting used in this thesis.

8.5. Final reflection

The increasing share of weather-dependent renewable generation makes the anticipation of extreme electricity prices more important. Negative prices and price spikes are not only unusual market outcomes, but also signals of moments in which the electricity system faces surplus, scarcity, or cross-border stress. In an interconnected electricity system, these situations are shaped by both domestic conditions and neighbouring market dynamics. This thesis shows that an ex-ante probabilistic early-warning model can help make these risks more visible before day-ahead market outcomes are known.

At the same time, the model should be interpreted as a tool to support decision-making, rather than as a replacement for expert market judgement. The probabilities indicate which hours may require closer attention, but they do not prescribe a single best action. Market participants still need to interpret the warning signals in relation to their own portfolio, contracts, flexibility options, and risk preferences. The value of the model therefore lies in structuring uncertainty and supporting preparation, not in removing the need for human interpretation.

Several related aspects were considered during the research, but fall outside the scope of this thesis. These include the full market-clearing process, intraday and balancing-market outcomes, actor-specific trading, hedging, and dispatch decisions, the economic value of acting on warnings, and potential feedback effects if similar warning signals were used more widely. Addressing these aspects would require additional market, portfolio, and behavioural assumptions beyond the ex-ante prediction framework developed here. The contribution of this thesis is therefore more specific: it shows how open, ex-ante data can be translated into probabilistic warning signals for Dutch day-ahead extreme prices. This provides a basis for making extreme-price risk visible in advance, while leaving room for future work to connect these warnings more directly to operational decisions and market outcomes.

Bibliography

Agency for the Cooperation of Energy Regulators. (2023, January 10). *ACER Decision 01-2023 on the Harmonised Maximum and Minimum Clearing Price Methodology for Single Day-Ahead Coupling* (Decision). Agency for the Cooperation of Energy Regulators. https://www.acer.europa.eu/sites/default/files/documents/Individual%20Decisions/ACER%20Decision%2001-2023%20on%20HMMCP%20SDAC.pdf

Amjady, N., & Keynia, F. (2011). A new prediction strategy for price spike forecasting of day-ahead electricity markets. *Applied Soft Computing*, *11*(6), 4246–4256. https://doi.org/10.1016/j.asoc.2011.03.024

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The Balanced Accuracy and Its Posterior Distribution. *Proceedings of the 20th International Conference on Pattern Recognition*, 3121–3124. https://doi.org/10.1109/ICPR.2010.764

Cawley, G. C., & Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research*, *11*, 2079–2107.

Chatzigiannis, D. I., Dourbois, G. A., Biskas, P. N., & Bakirtzis, A. G. (2016). European day-ahead electricity market clearing model. *Electric Power Systems Research*, *140*, 225–239. https://doi.org/10.1016/j.epsr.2016.06.019

Christen, P., Hand, D. J., & Kirielle, N. (2023). A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives. *ACM Computing Surveys*, *56*(3), 1–24. https://doi.org/10.1145/3606367

Davi-Arderius, D., & Jamasb, T. (2026). Measuring a paradox: Zero-negative electricity prices. *Utilities Policy*, *98*, 102108. https://doi.org/10.1016/j.jup.2025.102108

De Blauwe, J., Deissenroth-Uhrig, M., Mantke, H., & Keles, D. (2025). Cross-border effects on electricity spot prices - a meta-study. *Renewable and Sustainable Energy Reviews*, *224*, 116094. https://doi.org/10.1016/j.rser.2025.116094

De Vos, K. (2015). Negative Wholesale Electricity Prices in the German, French and Belgian Day-Ahead, Intra-Day and Real-Time Markets. *The Electricity Journal*, *28*(4), 36–50. https://doi.org/10.1016/j.tej.2015.04.001

Dimitriadis, T., Gneiting, T., & Jordan, A. I. (2021). Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, *118*(8), e2016191118. https://doi.org/10.1073/pnas.2016191118

Elkan, C. (2001). The Foundations of Cost-Sensitive Learning. *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 973–978.

ENTSO-E. (n.d.). *Transparency Platform*. Retrieved February 16, 2026, from https://transparency.entsoe.eu/

EPEX SPOT. (n.d.-a). *15-minute products in Market Coupling*. Retrieved May 26, 2026, from https://www.epexspot.com/en/15-minute-products-market-coupling

EPEX SPOT. (n.d.-b). *Basics of the Power Market*. Retrieved February 16, 2026, from https://www.epexspot.com/en/basicspowermarket?#negative-preise-prix-negatifs

Gasparella, A., Koolen, D., & Zucker, A. (2023). *The Merit Order and Price-Setting Dynamics in European Electricity Markets* (tech. rep. No. JRC134300) (Accessed: 2026-05-26). European Commission, Joint Research Centre. https://publications.jrc.ec.europa.eu/repository/handle/JRC134300

Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Haben, S., Caudron, J., & Verma, J. (2021). Probabilistic Day-Ahead Wholesale Price Forecast: A Case Study in Great Britain. *Forecasting*, *3*(3), 596–632. https://doi.org/10.3390/forecast3030038

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75–105. https://doi.org/10.2307/25148625

Hughes, R. A., Heron, J., Sterne, J. A. C., & Tilling, K. (2019). Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *International Journal of Epidemiology*, *48*(4), 1294–1304. https://doi.org/10.1093/ije/dyz032

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Klæboe, G., Braathen, J., Eriksrud, A. L., & Fleten, S.-E. (2022). Day-ahead market bidding taking the balancing power market into account. *TOP*, *30*(3), 683–703. https://doi.org/10.1007/s11750-022-00645-1

Liu, L., Bai, F., Su, C., Ma, C., Yan, R., Li, H., Sun, Q., & Wennersten, R. (2022). Forecasting the occurrence of extreme electricity prices using a multivariate logistic regression model. *Energy*, *247*, 123417. https://doi.org/10.1016/j.energy.2022.123417

Loizidis, S., Kyprianou, A., & Georghiou, G. E. (2024). Electricity market price forecasting using ELM and Bootstrap analysis: A case study of the German and Finnish Day-Ahead markets. *Applied Energy*, *363*, 123058. https://doi.org/10.1016/j.apenergy.2024.123058

Loizidis, S., Venizelou, V., Kyprianou, A., & Georghiou, G. E. (2025). Integrating PNN classification and ELM-Bootstrap for enhanced Day-Ahead negative price forecasting. *Applied Energy*, *392*, 126013. https://doi.org/10.1016/j.apenergy.2025.126013

Lu, Y., & Suthaharan, N. (2023). Electricity price spike clustering: A zero-inflated GARX approach. *Energy Economics*, *124*, 106834. https://doi.org/10.1016/j.eneco.2023.106834

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, *2*, 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774.

Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, *110*, 63–73. https://doi.org/10.1016/j.jclinepi.2019.02.016

Open-Meteo. (n.d.). *Free Open-Source Weather API*. Retrieved February 19, 2026, from https://open-meteo.com/

Ovaere, M., Kenis, M., Van den Bergh, K., Bruninx, K., & Delarue, E. (2023). The effect of flow-based market coupling on cross-border exchange volumes and price convergence in Central Western European electricity markets. *Energy Economics*, *118*, 106519. https://doi.org/10.1016/j.eneco.2023.106519

Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Perdomo, J. C., Zrnic, T., Mendler-Dünner, C., & Hardt, M. (2020). Performative prediction. *Proceedings of the 37th International Conference on Machine Learning*, *119*, 7599–7609. https://proceedings.mlr.press/v119/perdomo20a.html

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Ruggles, T. H., & Caldeira, K. (2022). Wind and solar generation may reduce the inter-annual variability of peak residual load in certain electricity systems. *Applied Energy*, *305*, 117773. https://doi.org/10.1016/j.apenergy.2021.117773

scikit-learn developers. (2026a). *sklearn.metrics: Metrics API Reference* [scikit-learn 1.8.0 documentation]. Retrieved May 22, 2026, from https://scikit-learn.org/stable/api/sklearn.metrics.html

scikit-learn developers. (2026b). *sklearn.metrics.average_precision_score*. Retrieved May 22, 2026, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html

scikit-learn developers. (2026c). *sklearn.metrics.balanced_accuracy_score*. Retrieved May 22, 2026, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html

scikit-learn developers. (2026d). *sklearn.metrics.brier_score_loss*. Retrieved May 22, 2026, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.brier_score_loss.html

scikit-learn developers. (2026e). *sklearn.metrics.confusion_matrix*. Retrieved May 22, 2026, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html

scikit-learn developers. (2026f). *sklearn.metrics.fbeta_score*. Retrieved May 22, 2026, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.fbeta_score.html

scikit-learn developers. (2026g). *sklearn.metrics.precision_score*. Retrieved May 22, 2026, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html

scikit-learn developers. (2026h). *sklearn.metrics.recall_score*. Retrieved May 22, 2026, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html

Sensfuß, F., Ragwitz, M., & Genoese, M. (2008). The merit-order effect: A detailed analysis of the price effect of renewable electricity generation on spot market prices in Germany. *Energy Policy*, *36*(8), 3086–3094. https://doi.org/10.1016/j.enpol.2008.03.035

Shapley, L. S. (1953). A Value for n-Person Games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the Theory of Games II* (pp. 307–317, Vol. 28). Princeton University Press.

Sheng, V. S., & Ling, C. X. (2006). Thresholding for Making Classifiers Cost-sensitive. *Proceedings of the 21st National Conference on Artificial Intelligence*, 476–481.

Statistics Netherlands. (2025, March 10). *Half of electricity is produced from renewable sources*. Retrieved February 16, 2026, from https://www.cbs.nl/en-gb/news/2025/11/half-of-electricity-is-produced-from-renewable-sources

TenneT. (n.d.-a). *BritNed*. Retrieved February 16, 2026, from https://www.tennet.eu/nl/projecten/britned

TenneT. (n.d.-b). *COBRAcable*. Retrieved February 16, 2026, from https://www.tennet.eu/nl/projecten/cobracable

TenneT. (n.d.-c). *Grid maps*. Retrieved February 16, 2026, from https://www.tennet.eu/nl-en/our-high-voltage-grid/grid-maps

TenneT. (n.d.-d). *Market coupling*. Retrieved February 16, 2026, from https://www.tennet.eu/nl-en/markets/dutch-market/market-coupling

TenneT. (n.d.-e). *NorNed*. Retrieved February 16, 2026, from https://www.tennet.eu/nl/projecten/norned

TenneT. (2025, April 7). *TenneT provides 9 gigawatts of capacity on the high-voltage grid for customers on the waiting list*. https://www.tennet.eu/nl-en/news/tennet-provides-9-gigawatts-capacity-high-voltage-grid-customers-waiting-list

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

Zou, Q., Xie, S., Lin, Z., Wu, M., & Ju, Y. (2016). Finding the Best Classification Threshold in Imbalanced Classification. *Big Data Research*, *5*, 2–8. https://doi.org/10.1016/j.bdr.2015.12.001

A

Literature search strategy

This appendix documents the literature search strings and database filtering criteria used in the scoping review. The searches were conducted in Scopus and Web of Science using combinations of core concepts related to day-ahead electricity markets, extreme price events, probabilistic forecasting, and cross-border market mechanisms.

A.1. Search strings

Table A.1: Search strings used in Scopus and Web of Science.

Label	Date of search	Results	Search string
1A	09/02/2026	15	TITLE-ABS-KEY((”day-ahead” OR ”day ahead” OR dayahead OR ”day-ahead market” OR ”spot market” OR ”DA price”) AND (electricity OR power) AND price* AND (”negative price*” OR (price* W/3 spike*) OR (”extreme” W/2 price*) OR (”tail” W/2 event*) OR ”scarcity pricing” OR ”price volatility” OR ”negative electricity price*”) AND (”probabilistic forecast*” OR ”probability forecast*” OR ”quantile forecast*” OR ”interval forecast*” OR ”early warning” OR ”risk forecast*” OR ”event prediction” OR classification))
1B	09/02/2026	16	TITLE-ABS-KEY((”day-ahead” OR ”day ahead” OR dayahead OR ”day-ahead market” OR ”spot market” OR ”DA price”) AND (electricity OR power) AND price* AND (”market coupling” OR ”price coupling” OR SDAC OR ”Single Day-Ahead Coupling” OR EUPHEMIA OR ”market clearing” OR auction) AND (”flow-based” OR NTC OR ”net transfer capacity” OR interconnect* OR ”cross-border” OR ”cross border” OR ”cross-border flow*” OR ”price convergence”) AND (congestion OR ”transmission constraint*” OR ”transmission capacity”))

Label	Date of search	Results	Search string
1C	09/02/2026	27	TITLE-ABS-KEY((”day-ahead” OR ”day ahead” OR dayahead OR ”day-ahead market” OR ”spot market” OR ”DA price”) AND (electricity OR power) AND price* AND (”negative price*” OR ”negative wholesale price*” OR (price* W/3 spike*) OR (”extreme” W/2 price*) OR ”scarcity pricing” OR ”price volatility” OR ”negative electricity price*”) AND (”load forecast*” OR ”demand forecast*” OR ”wind forecast*” OR ”solar forecast*” OR ”renewable* forecast*” OR ”forecast error” OR ”net load” OR ”residual load” OR outage* OR curtailment))
2A	10/02/2026	18	TS=((”day-ahead” OR ”day ahead” OR dayahead OR ”day-ahead market” OR ”spot market” OR ”DA price”) AND (electricity OR power) AND price* AND (”negative price*” OR (price* NEAR/3 spike*) OR (”extreme” NEAR/2 price*) OR (”tail” NEAR/2 event*) OR ”scarcity pricing” OR ”price volatility”) AND (”probabilistic forecast*” OR ”probability forecast*” OR ”quantile forecast*” OR ”interval forecast*” OR ”early warning” OR ”risk forecast*” OR ”event prediction” OR classification))
2B	10/02/2026	24	TS=((”day-ahead” OR ”day ahead” OR dayahead OR ”day-ahead market” OR ”spot market” OR ”DA price”) AND (electricity OR power) AND price* AND (”market coupling” OR ”price coupling” OR SDAC OR ”Single Day-Ahead Coupling” OR EUPHEMIA OR ”market clearing” OR auction) AND (”flow-based” OR NTC OR ”net transfer capacity” OR interconnect* OR ”cross-border” OR ”cross border” OR ”cross-border flow*” OR ”price convergence”) AND (congestion OR ”transmission constraint*” OR ”transmission capacity”))

Label	Date of search	Results	Search string
2C	10/02/2026	23	TS=((”day-ahead” OR ”day ahead” OR dayahead OR ”day-ahead market” OR ”spot market” OR ”DA price”) AND (electricity OR power) AND price* AND (”negative price*” OR ”negative wholesale price*” OR (price* NEAR/3 spike*) OR (”extreme” NEAR/2 price*) OR ”scarcity pricing” OR ”price volatility”) AND (”load forecast*” OR ”demand forecast*” OR ”wind forecast*” OR ”solar forecast*” OR ”renewable* forecast*” OR ”forecast error” OR ”net load” OR ”residual load” OR outage* OR curtailment))

A.2. Database filtering criteria

Table A.2 reports the database-specific filters applied after running the search strings.

Database	Criteria
Scopus (1A–1C)	PUBYEAR > 2005 AND PUBYEAR < 2027 AND (LIMIT-TO (SUBJAREA,”ENER”) OR LIMIT-TO (SUBJAREA,”ENGI”) OR LIMIT-TO (SUBJAREA,”ECON”) OR LIMIT-TO (SUBJAREA,”COMP”) OR LIMIT-TO (SUBJAREA,”MATH”) OR LIMIT-TO (SUBJAREA,”ENVI”) OR LIMIT-TO (SUBJAREA,”BUSI”) OR LIMIT-TO (SUBJAREA,”MULT”) OR LIMIT-TO (SUBJAREA,”EART”)) AND (LIMIT-TO (DOCTYPE,”re”) OR LIMIT-TO (DOCTYPE,”ar”) OR LIMIT-TO (DOCTYPE,”ch”) OR LIMIT-TO (DOCTYPE,”bk”)) AND (LIMIT-TO (LANGUAGE,”English”))
Web of Science (2A–2C)	Refined by: • Years: 2006–2026 • Document type: Article, Proceeding Paper, Review Article, Data Paper, Book, Book Chapter • Language: English

Table A.2: Filtering criteria used for database searches.

B

PRISMA flow diagram and literature overview

B.1. PRISMA flow diagram

The literature selection was documented using a PRISMA-based flow diagram, to make the screening process transparent and reproducible. The purpose is not to cover all literature on electricity price forecasting, but to select academic studies that support the conceptualisation, feature selection, and modelling design of this thesis.

The initial database search was conducted in Scopus and Web of Science, using the search strings reported in Appendix A. The search strings were developed around the focus of the review: day-ahead electricity price formation, extreme price events, probabilistic forecasting, and cross-border market coupling. The database searches returned 123 records in total: 58 from Scopus and 65 from Web of Science. After duplicate removal based on titles, author information, and DOI numbers, 72 records remained for screening.

The screening was conducted in two steps. First, titles and abstracts were screened for relevance to the scope of the thesis. Studies were kept when they addressed at least one of the following review aspects: day-ahead electricity price formation, negative prices or price spikes, probabilistic electricity price forecasting, or cross-border effects in electricity markets. This step excluded 29 records that were outside the scope of the review. Second, the remaining 43 full texts were assessed for eligibility.

Reports were excluded during full-text assessment when they did not contribute directly to the thesis objective. The main exclusion reasons were: not being a journal article, focusing mainly on optimisation or market participation, not being relevant to early warning of day-ahead extreme prices, focusing mainly on market design or regulation, using a market context that was not transferable to the Dutch or European day-ahead setting, focusing on unrelated markets, addressing general price forecasting without an extreme-price component, or being redundant with a more relevant included study. This resulted in 12 journal articles included in the final literature review, which are presented in Table B.1.

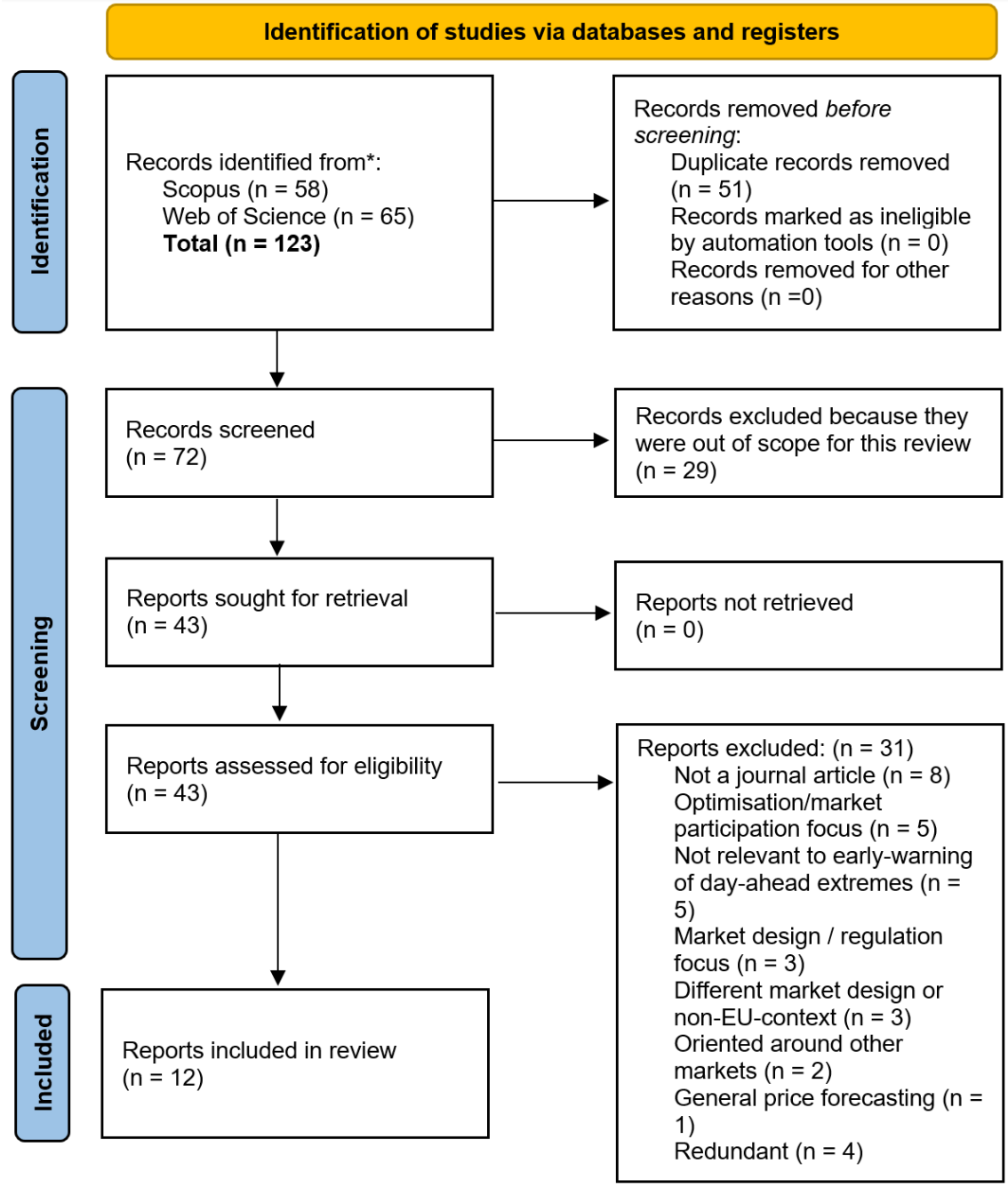

Figure B.1: PRISMA 2020 flow diagram for the scoping review.

B.2. Extended literature overview

Table B.1 provides a more detailed overview of the selected articles than the shorter synthesis table in Chapter 2. It documents the focus, market context, method, and key takeaway of each study.

Table B.1: Selected literature overview.

Article	Focus	Market	Method	Key takeaway
Loizidis et al. (2025)	Negative-price prediction	Germany, 2020–2023	PNN classifier; model comparison	Cost-sensitive modelling improves negative-price detection, especially by reducing expensive false negatives.
Lu & Suthaharan (2023)	Spike clustering	Australia, 2016–2022	GARX time-series model	Spikes are rare, clustered, and temporally dependent, so models must account for imbalance and persistence.
Amjady & Keynia (2011)	Spike occurrence and magnitude	PJM	Event-oriented spike forecasting	Separating spike occurrence from spike magnitude improves prediction and highlights the need for targeted spike features.
Liu et al. (2022)	Extreme-event probabilities	Wholesale day-ahead market	Multivariate logistic regression	Logistic regression offers an interpretable probabilistic baseline, but captures non-linearity and interactions less well.
Haben et al. (2021)	Probabilistic day-ahead forecasting	Great Britain	Probabilistic forecasting framework	Probabilistic forecasts are more useful than point forecasts for decision-making under extreme-price risk.
Loizidis et al. (2024)	Uncertainty forecasting	Germany and Finland	ELM with interval analysis	Extreme prices can be represented in probabilistic outputs, although the focus is on intervals rather than cross-border effects.
De Vos (2015)	Drivers of negative prices	Germany, France, Belgium	Comparative market analysis	Negative prices reflect oversupply and downward inflexibility rather than noise, and arise through identifiable system mechanisms.
Davi-Arderius & Jamasb (2026)	Meaning of negative prices	Europe	Policy and system-cost interpretation	Negative day-ahead prices often coincide with broader system stress and balancing costs, which supports the value of early warning.
De Blauwe et al. (2025)	Cross-border price effects	EU markets	Meta-study	Cross-border variables matter for price formation, but their effects are heterogeneous and often seasonal.
Ovaere et al. (2023)	Flow-based market coupling	CWE and SDAC/Flow-Based Market Coupling (FBMC)	Empirical evaluation of FBMC	Flow-based coupling affects exchanges and price convergence, showing that capacity and coupling signals can shape extremes.
Unger et al. (2018)	Wind and cross-border pricing	Denmark	Empirical analysis of wind, flows, and prices	Forecast wind output and planned flows influence cross-border price differences and can serve as early-warning signals.
Chatzigiannis et al. (2016)	Day-ahead market clearing	EU day-ahead market	Market-clearing model	Pre-closure information and clearing constraints are crucial for understanding how extreme day-ahead prices emerge.

C

Feature dictionary and ex-ante availability

This appendix provides the detailed feature dictionary used to support the data and feature-engineering choices in Chapter 4. The tables document the variables used in the dataset, their operationalisation, units, literature or modelling support, expected importance based on literature, source type, and ex-ante availability status.

The variable names in the tables follow the column names used in the modelling dataset. Variables included in the main modelling framework must be available before day-ahead market closure. Variables for which publication timing cannot be verified are excluded from the main model or reserved for robustness checks. For the main feature groups, the documented ex-ante status was supported by manual retrieval checks for upcoming delivery periods, in addition to source documentation and historical data inspection.

C.1. Target variable

The target variable is constructed from the Dutch hourly day-ahead market-clearing price. The realised price itself is not used as a same-day explanatory variable, since it is only known after market clearing. Instead, it is used to construct the three-class label used in the probabilistic classification models.

Table C.1: Target variable dictionary.

Variable	Definition	Unit	Support	Importance	Source	Ex-ante
Date	Hourly timestamp of the delivery hour.	Date-time	Needed for alignment and temporal validation.	Support	Dataset	Known before modelling; not a predictor by itself.
Day-ahead price (EUR/MWh)	Dutch hourly day-ahead market-clearing price.	€/MWh	Used as price outcome in day-ahead price studies.	Target	ENTSO-E	Not ex-ante; used to construct labels.
Price category	Three-class label: negative, normal, or spike.	Class	Supported by negative-price and spike literature.	Target	Derived	Target label; known after clearing.

C.2. Dutch price-history variables

Dutch price-history variables are derived from historical Dutch day-ahead prices and the assigned price category labels. They capture price persistence, recent volatility, and the clustering of extreme price events. These variables satisfy the ex-ante constraint since they only use information available before the prediction moment.

Table C.2: Dutch price-history variable dictionary.

Variable	Definition	Unit	Support	Importance	Source	Ex-ante
1-day lagged price	Same-hour Dutch price one day earlier.	€/MWh	Price persistence.	High	Derived	Past value; ex-ante.
2-day lagged price	Same-hour Dutch price two days earlier.	€/MWh	Price persistence.	Medium	Derived	Past value; ex-ante.
3-day lagged price	Same-hour Dutch price three days earlier.	€/MWh	Price persistence.	Medium	Derived	Past value; ex-ante.
Week-lagged price	Same-hour Dutch price one week earlier.	€/MWh	Weekly price pattern.	High	Derived	Past value; ex-ante.
Same-hour day-before spike	Indicator for a spike in the same hour of the previous day.	Binary	Extreme-event clustering.	High	Derived	Derived from past label; ex-ante.
Same-hour day-before negative	Indicator for a negative price in the same hour of the previous day.	Binary	Extreme-event clustering.	High	Derived	Derived from past label; ex-ante.
Spike count day before	Number of spikes during the previous day.	Count	Spike clustering.	High	Derived	Derived from past label; ex-ante.
Negative count day before	Number of negative prices during the previous day.	Count	Negative-event persistence.	High	Derived	Derived from past label; ex-ante.
Spike count last week	Number of spikes during the previous week.	Count	Spike clustering.	High	Derived	Derived from past label; ex-ante.
Negative count last week	Number of negative prices during the previous week.	Count	Negative-event persistence.	High	Derived	Derived from past label; ex-ante.

C.3. Domestic system variables

Domestic system variables represent expected supply-demand conditions in the Dutch electricity market. They operationalise domestic market stress and flexibility stress before day-ahead market clearing. Forecast variables are used instead of realised variables because the model is designed as an ex-ante early-warning tool.

Table C.3: Domestic system variable dictionary.

Variable	Definition	Unit	Support	Importance	Source	Ex-ante
NL load forecast	Forecasted Dutch electricity demand.	MW	Demand and scarcity signal.	Very high	ENTSO-E	Forecast; pre-closure.
NL generation forecast	Forecasted Dutch total generation.	MW	Supply condition.	Medium	ENTSO-E	Forecast; pre-closure.
NL solar forecast	Forecasted Dutch solar generation.	MW	Surplus and low-price signal.	High	ENTSO-E	Forecast; pre-closure.
NL wind offshore forecast	Forecasted Dutch offshore wind generation.	MW	Renewable surplus signal.	High	ENTSO-E	Forecast; pre-closure.
NL wind onshore forecast	Forecasted Dutch onshore wind generation.	MW	Renewable surplus signal.	High	ENTSO-E	Forecast; pre-closure.
NL renewable generation forecast	Forecasted Dutch wind and solar generation.	MW	Renewable pressure.	High	Derived / ENTSO-E	Forecast; pre-closure.
NL renewable ramping forecast	Change in forecasted renewable generation.	MW/h	Flexibility stress.	Medium	Derived	Derived from pre-closure variables.
NL mFRR down	Downward manual reserve-capacity indicator.	MW	Downward flexibility.	Medium	ENTSO-E	Pre-closure.
NL mFRR up	Upward manual reserve-capacity indicator.	MW	Upward flexibility.	Medium	ENTSO-E	Pre-closure.
NL aFRR down	Downward automatic reserve-capacity indicator.	MW	Downward flexibility.	Medium	ENTSO-E	Pre-closure.
NL aFRR up	Upward automatic reserve-capacity indicator.	MW	Upward flexibility.	Medium	ENTSO-E	Pre-closure.
NL unavailable capacity	Planned unavailable generation capacity.	MW	Supply constraint.	Medium	ENTSO-E	Pre-closure.
NL total import capacity	Sum of day-ahead NTC from DE/LU, BE, DK1, and NO2 to NL.	MW	Cross-border import capacity.	High	Derived	Derived from pre-closure variables.
NL total export capacity	Sum of day-ahead NTC from NL to DE/LU, BE, DK1, and NO2.	MW	Cross-border export capacity.	High	Derived	Derived from pre-closure variables.

C.4. Weather variables

Weather variables are included as supporting predictors of demand-side and supply-side conditions. Temperature can affect load demand, while wind and solar conditions affect renewable generation. Weather variables are included as historical forecasts, since that is consistent with the ex-ante modelling setup.

Table C.4: Weather variable dictionary.

Variable	Definition	Unit	Support	Importance	Source	Ex-ante
NL temperature	Dutch temperature.	°C	Demand/weather signal.	Medium	Open-Meteo	Historical forecast; pre-closure.
NL sunshine duration	Dutch sunshine duration.	Hours	Solar condition.	Low–med.	Open-Meteo	Historical forecast; pre-closure.
NL direct radiation	Dutch direct radiation.	W/m²	Solar condition.	Low–med.	Open-Meteo	Historical forecast; pre-closure.
NL wind speed at 100m	Dutch wind speed at 100m.	m/s	Wind condition.	Medium	Open-Meteo	Historical forecast; pre-closure.
DE/LU temperature	DE/LU temperature.	°C	Neighbour demand/weather.	Medium	Open-Meteo	Historical forecast; pre-closure.
DE/LU sunshine duration	DE/LU sunshine duration.	Hours	Neighbour solar condition.	Low–med.	Open-Meteo	Historical forecast; pre-closure.
DE/LU direct radiation	DE/LU direct radiation.	W/m²	Neighbour solar condition.	Low–med.	Open-Meteo	Historical forecast; pre-closure.
DE/LU wind speed at 100m	DE/LU wind speed at 100m.	m/s	Neighbour wind condition.	Medium	Open-Meteo	Historical forecast; pre-closure.
GB temperature	GB temperature.	°C	Neighbour demand/weather.	Medium	Open-Meteo	Historical forecast; pre-closure.
GB sunshine duration	GB sunshine duration.	Hours	Neighbour solar condition.	Low–med.	Open-Meteo	Historical forecast; pre-closure.
GB direct radiation	GB direct radiation.	W/m²	Neighbour solar condition.	Low–med.	Open-Meteo	Historical forecast; pre-closure.
GB wind speed at 100m	GB wind speed at 100m.	m/s	Neighbour wind condition.	Medium	Open-Meteo	Historical forecast; pre-closure.
BE temperature	Belgian temperature.	°C	Neighbour demand/weather.	Medium	Open-Meteo	Historical forecast; pre-closure.
BE sunshine duration	Belgian sunshine duration.	Hours	Neighbour solar condition.	Low–med.	Open-Meteo	Historical forecast; pre-closure.
BE direct radiation	Belgian direct radiation.	W/m²	Neighbour solar condition.	Low–med.	Open-Meteo	Historical forecast; pre-closure.
BE wind speed at 100m	Belgian wind speed at 100m.	m/s	Neighbour wind condition.	Medium	Open-Meteo	Historical forecast; pre-closure.
NO2 temperature	Norwegian temperature.	°C	Neighbour demand/weather.	Low–med.	Open-Meteo	Historical forecast; pre-closure.
NO2 sunshine duration	Norwegian sunshine duration.	Hours	Neighbour solar condition.	Low	Open-Meteo	Historical forecast; pre-closure.
NO2 direct radiation	Norwegian direct radiation.	W/m²	Neighbour solar condition.	Low	Open-Meteo	Historical forecast; pre-closure.
NO2 wind speed at 100m	Norwegian wind speed at 100m.	m/s	Neighbour wind condition.	Medium	Open-Meteo	Historical forecast; pre-closure.
DK1 temperature	Danish temperature.	°C	Neighbour demand/weather.	Medium	Open-Meteo	Historical forecast; pre-closure.
DK1 sunshine duration	Danish sunshine duration.	Hours	Neighbour solar condition.	Low–med.	Open-Meteo	Historical forecast; pre-closure.
DK1 direct radiation	Danish direct radiation.	W/m²	Neighbour solar condition.	Low–med.	Open-Meteo	Historical forecast; pre-closure.
DK1 wind speed at 100m	Danish wind speed at 100m.	m/s	Neighbour wind condition.	Medium	Open-Meteo	Historical forecast; pre-closure.

C.5. Cross-border variables

Cross-border variables represent neighbouring market pressure and the transmission of scarcity or surplus through interconnectors and market coupling. The variables cover directly connected bidding zones for which scheduled-exchange and day-ahead transfer-capacity variables are available in the final dataset. Import and export variables are defined from the Dutch perspective. Table C.5 reports the operationalisation and ex-ante status of each cross-border variable.

Table C.5: Cross-border variable dictionary.

Variable	Definition	Unit	Importance	Source	Ex-ante
NL scheduled export to DE/LU	Scheduled export from NL to DE/LU.	MW	High	ENTSO-E	Planned; pre-closure.
NL scheduled import from DE/LU	Scheduled import from DE/LU to NL.	MW	High	ENTSO-E	Planned; pre-closure.
NL import capacity from DE/LU	Day-ahead NTC from DE/LU to NL.	MW	High	ENTSO-E	Pre-closure.
NL export capacity to DE/LU	Day-ahead NTC from NL to DE/LU.	MW	High	ENTSO-E	Pre-closure.
DE/LU generation forecast	DE/LU total generation forecast.	MW	Medium	ENTSO-E	Forecast; pre-closure.
DE/LU load forecast	DE/LU demand forecast.	MW	High	ENTSO-E	Forecast; pre-closure.
DE/LU residual load forecast	DE/LU load minus wind and solar forecasts.	MW	High	Derived	Derived from pre-closure variables.
DE/LU lagged day-ahead price	Previous DE/LU day-ahead price.	€/MWh	High	ENTSO-E	Past value; ex-ante.
NL scheduled export to BE	Scheduled export from NL to BE.	MW	High	ENTSO-E	Planned; pre-closure.
NL scheduled import from BE	Scheduled import from BE to NL.	MW	High	ENTSO-E	Planned; pre-closure.
NL import capacity from BE	Day-ahead NTC from BE to NL.	MW	High	ENTSO-E	Pre-closure.
NL export capacity to BE	Day-ahead NTC from NL to BE.	MW	High	ENTSO-E	Pre-closure.
BE generation forecast	Belgian total generation forecast.	MW	Medium	ENTSO-E	Forecast; pre-closure.
BE load forecast	Belgian demand forecast.	MW	High	ENTSO-E	Forecast; pre-closure.
BE residual load forecast	Belgian load minus wind and solar forecasts.	MW	High	Derived	Derived from pre-closure variables.
BE lagged day-ahead price	Previous Belgian day-ahead price.	€/MWh	High	ENTSO-E	Past value; ex-ante.
NL scheduled export to DK1	Scheduled export from NL to DK1.	MW	Medium	ENTSO-E	Planned; pre-closure.
NL scheduled import from DK1	Scheduled import from DK1 to NL.	MW	Medium	ENTSO-E	Planned; pre-closure.
NL import capacity from DK1	Day-ahead NTC from DK1 to NL.	MW	Medium	ENTSO-E	Pre-closure.
NL export capacity to DK1	Day-ahead NTC from NL to DK1.	MW	Medium	ENTSO-E	Pre-closure.
DK1 generation forecast	DK1 total generation forecast.	MW	Medium	ENTSO-E	Forecast; pre-closure.
DK1 load forecast	DK1 demand forecast.	MW	Medium	ENTSO-E	Forecast; pre-closure.
DK1 residual load forecast	DK1 load minus wind and solar forecasts.	MW	Medium	Derived	Derived from pre-closure variables.
DK1 lagged day-ahead price	Previous DK1 day-ahead price.	€/MWh	Medium	ENTSO-E	Past value; ex-ante.
NL scheduled export to NO2	Scheduled export from NL to NO2.	MW	Medium	ENTSO-E	Planned; pre-closure.
NL scheduled import from NO2	Scheduled import from NO2 to NL.	MW	Medium	ENTSO-E	Planned; pre-closure.
NL import capacity from NO2	Day-ahead NTC from NO2 to NL.	MW	Medium	ENTSO-E	Pre-closure.
NL export capacity to NO2	Day-ahead NTC from NL to NO2.	MW	Medium	ENTSO-E	Pre-closure.
NO2 generation forecast	NO2 total generation forecast.	MW	Medium	ENTSO-E	Forecast; pre-closure.
NO2 load forecast	NO2 demand forecast.	MW	Medium	ENTSO-E	Forecast; pre-closure.
NO2 residual load forecast	NO2 load minus wind and solar forecasts.	MW	Medium	Derived	Derived from pre-closure variables.
NO2 lagged day-ahead price	Previous NO2 day-ahead price.	€/MWh	Medium	ENTSO-E	Past value; ex-ante.

C.6. Calendar controls

Calendar controls capture recurring temporal patterns. They are treated as supporting variables that help separate predictable time patterns from domestic, weather, and cross-border stress signals. Since these variables are known before delivery, they satisfy the ex-ante constraint by construction.

Table C.6: Calendar control variable dictionary.

Variable	Definition	Unit	Support	Importance	Source	Ex-ante
NL_holiday_dummy	Dutch public-holiday indicator.	Binary	Calendar demand pattern.	Medium	Calendar	Known ex-ante.
DE_holiday_dummy	German public-holiday indicator.	Binary	Neighbour demand pattern.	Low–med.	Calendar	Known ex-ante.
GB_holiday_dummy	British public-holiday indicator.	Binary	Neighbour demand pattern.	Low–med.	Calendar	Known ex-ante.
BE_holiday_dummy	Belgian public-holiday indicator.	Binary	Neighbour demand pattern.	Low–med.	Calendar	Known ex-ante.
NO_holiday_dummy	Norwegian public-holiday indicator.	Binary	Neighbour demand pattern.	Low–med.	Calendar	Known ex-ante.
DK_holiday_dummy	Danish public-holiday indicator.	Binary	Neighbour demand pattern.	Low–med.	Calendar	Known ex-ante.
LU_holiday_dummy	Luxembourg public-holiday indicator.	Binary	Neighbour demand pattern.	Low–med.	Calendar	Known ex-ante.
Day of the week	Day-of-week indicator.	Index	Weekly demand pattern.	Medium	Calendar	Known ex-ante.
Weekend	Weekend indicator.	Binary	Working/non-working day.	Medium	Calendar	Known ex-ante.
Month	Month indicator.	Index	Seasonal variation.	Medium	Calendar	Known ex-ante.
Hour of the day	Delivery-hour indicator.	Index	Intraday demand pattern.	Medium	Calendar	Known ex-ante.

Together, these tables document the final feature set used in Chapter 4. The modelling chapter uses these variables as the starting point for model comparison, while later feature reduction or exclusion is based on model-based evidence rather than on the feature dictionary itself.

D

Model specifications and tuning details

This appendix provides the detailed model rationale and reproducibility information for the candidate models used in Chapter 5. The main text only summarises the role of each model, while this appendix gives the extended motivation, fixed implementation settings, hyperparameter grids, and final selected settings used for the 2025 evaluation.

D.1. Candidate models

The dummy classifier is included as a naive benchmark. It does not learn a relationship between the input features and the target classes, but provides a lower reference point for model comparison. This is important because the target variable is imbalanced: the dataset consists largely of normal prices, while negative prices and price spikes are rare. A model can therefore appear to perform well when assessed only through overall accuracy, even if it fails to detect the extreme-price classes. The inclusion of a simple benchmark is consistent with probabilistic electricity-price forecasting literature, where proposed models are compared against reference models rather than evaluated in isolation (Haben et al., 2021). The event-focused metrics used to assess whether models improve beyond this benchmark are defined in Section 5.6.

Multinomial logistic regression is included as an interpretable probabilistic baseline. Logistic regression has been used to estimate the occurrence probability of extremely low and high electricity prices based on historical prices and market variables (Liu et al., 2022). Its main advantage in this thesis is transparency: the model estimates class probabilities through a linear combination of the explanatory variables. This makes it useful as a reference point for assessing whether more flexible models improve performance because they capture non-linearities and interactions. At the same time, logistic regression is not expected to capture all relevant price-formation mechanisms. Extreme prices may arise from combinations of demand, renewable generation, weather conditions, cross-border constraints, and recent price dynamics, making them non-linear.

GBDT is included as a non-linear tree-based model. Boosting combines multiple weak decision trees into an ensemble. Multiple individual trees are used to create a more stable predictive model. Each subsequent tree focuses on correcting errors made by the previous trees. This makes the model suitable for capturing non-linear relationships and interaction effects between feature groups. Such interactions are relevant in the context of this thesis. Negative prices may occur when high renewable generation, low demand, export limitations, and recent low-price conditions coincide. Price spikes may occur when residual load is high, renewable generation is low, imports are constrained, or neighbouring markets also experience tight conditions. A tree-based ensemble can capture such patterns more naturally than a linear model.

XGBoost is included as a computationally efficient implementation of gradient-boosted trees. It is suited to structured tabular datasets with many explanatory variables, such as the 102-feature dataset used in this thesis. XGBoost is also relevant because recent negative-price forecasting literature includes it as a benchmark against models such as PNN, SVM, and LSTM (Loizidis et al., 2025). In this thesis, XGBoost is included as a flexible candidate model because it combines non-linear predictive capacity with practical explainability. In particular, it can be analysed with feature-importance

and SHAP-based methods, which is important because the thesis aims not only to compare predictive performance, but also to interpret which domestic and cross-border indicators contribute to warning signals.

A multilayer perceptron is included as a neural-network benchmark. Neural-network approaches are common in electricity-price forecasting because they can identify complex non-linear relationships between inputs and outputs (Loizidis et al., 2024). Liu et al. (2022) also compare multivariate logistic regression with neural-network alternatives, including an MLP, for the forecasting of extremely low and high electricity prices. The MLP included here is not intended to represent the full range of deep-learning possibilities. Instead, it provides a standard neural-network comparison against the linear and tree-based models. This makes it possible to test whether a relatively flexible neural classifier improves the detection of Dutch negative prices and price spikes compared with the other model families. Since MLPs are sensitive to the scale of input variables, scaling is included in this model.

Finally, a PNN model is included as a probabilistic-neighbourhood reference inspired by probabilistic neural-network classification. PNN-type models are relevant for this thesis because they have been used in both price-spike and negative-price forecasting. Amjady and Keynia (2011) include a PNN component in a price-spike forecasting strategy, while Loizidis et al. (2025) use a PNN classifier for negative day-ahead price forecasting and compare it with alternative models such as XGBoost and LSTM. The implementation in this thesis is therefore used as a PNN-inspired reference model rather than as the main final model. Its purpose is to test whether probability estimates based on neighbourhood similarity in the feature space are competitive with the linear, tree-based, and neural-network alternatives.

D.2. Initial model configuration

This section reports the initial model definitions and fixed implementation settings used when the candidate models were first specified. These settings describe the modelling pipelines, preprocessing steps, and fixed parameters that were not the main focus of the hyperparameter search. Parameters not shown in Table D.1 follow the default settings of the corresponding Python packages used in the modelling pipeline.

The table is included for reproducibility. The relevance of the initial settings is assessed through the model-comparison results, warning behaviour, probability-quality assessment, sensitivity analysis, and feature-removal results reported in Chapter 6.

Table D.1: Initial model definitions and fixed implementation settings

Model	Preprocessing	Initial and fixed settings
Dummy classifier	None	strategy = most_frequent random_state = 12
Logistic regression	StandardScaler	max_iter = 1000 random_state = 12 L2 penalty lbfgs solver
GBDT	None	GradientBoostingClassifier initial n_estimators = 100 random_state = 12
XGBoost	None	objective = multi:softprob eval_metric = mlogloss initial n_estimators = 100 random_state = 12 tree_method = hist n_jobs = 1
MLP	StandardScaler	activation = relu solver = adam initial hidden_layer_sizes = (64, 32) alpha = 0.0001 max_iter = 500 early_stopping = False random_state = 12
PNN	StandardScaler	class_prior = uniform initial bandwidth = 1.0 kernel = gaussian

D.3. Hyperparameter grid

Table D.2 reports the hyperparameter grids used during model tuning. These grids define the values used during the model-development stage described in Section 5.4. Parameters not shown in the table were kept fixed during tuning or follow the default settings of the corresponding Python packages.

Table D.2: Hyperparameter grids used during model tuning

Model	Tuned hyperparameters	Values tested
Logistic regression	Regularisation strength	$C \in \{0.01, 0.1, 1, 10\}$
	Class weighting	{none, balanced}
GBDT	Number of trees	$n_{\text{estimators}} \in \{100, 200\}$
	Learning rate	$\{0.03, 0.05, 0.1\}$
	Maximum tree depth	$\{2, 3\}$
	Minimum samples per leaf	$\{1, 5, 10\}$
	Subsample share	$\{0.8, 1.0\}$
	Maximum features	{all, $\sqrt{p}$}
XGBoost	Number of trees	$n_{\text{estimators}} \in \{100, 200, 400\}$
	Learning rate	$\{0.03, 0.05, 0.1\}$
	Maximum tree depth	$\{2, 3, 4\}$
	Row subsampling	$\{0.8, 1.0\}$
	Column subsampling	$\{0.8, 1.0\}$
	Minimum child weight	$\{1, 5\}$
	L2 regularisation	$\lambda \in \{1, 5\}$
	Minimum loss reduction	$\gamma \in \{0, 1\}$
MLP	Hidden-layer structure	$\{(32), (64), (64, 32), (128, 64)\}$
	L2 regularisation	$\alpha \in \{10^{-5}, 10^{-4}, 10^{-3}\}$
	Initial learning rate	$\{0.0005, 0.001, 0.003\}$
	Batch size	$\{128, 256\}$
PNN	Bandwidth	$\{0.5, 1.0, 2.0\}$
	Kernel	{Gaussian}

Note: The full initial model definitions and final selected settings are reported in Sections D.2 and D.4.

D.4. Final selected model settings

Table D.3 reports the final selected hyperparameters and validation-selected warning thresholds used for the 2025 out-of-sample evaluation. The hyperparameters were selected during the model-development stage described in Section 5.4. The warning thresholds were selected on the 2024 validation period using the threshold-selection procedure described in Section 5.5.

After these settings had been fixed, the models were refitted on the combined 2022–2024 period and evaluated once on the 2025 test set. The table is included for reproducibility. The selected hyperparameters and warning thresholds are not interpreted as substantive results by themselves; their relevance is assessed through the model-comparison, warning-behaviour, probability-quality, sensitivity, and feature-removal results in Chapter 6.

Table D.3: Final selected model settings used for the 2025 evaluation

Model	Selected hyperparameters	τ_{negative}	τ_{spike}
Dummy classifier	Most-frequent class benchmark No tuned hyperparameters	–	–
Logistic regression	Regularisation strength: $C = 0.01$ Class weighting: balanced	0.400	0.500
GBDT	Number of trees: $n_{\text{estimators}} = 200$ Learning rate: 0.1 Maximum tree depth: 3 Minimum samples per leaf: 5 Subsample share: 0.8 Maximum features: all	0.020	0.100
XGBoost	Number of trees: $n_{\text{estimators}} = 100$ Learning rate: 0.1 Maximum tree depth: 3 Row subsampling: 0.8 Column subsampling: 1.0 Minimum child weight: 5 L2 regularisation: $\lambda = 1$ Minimum loss reduction: $\gamma = 1$	0.020	0.150
MLP	Hidden-layer structure: $(32,)$ L2 regularisation: $\alpha = 0.001$ Initial learning rate: 0.0005 Batch size: 128	0.001	0.001
PNN	Bandwidth: 2.0 Kernel: Gaussian	0.500	0.500

Note: The thresholds τ_{negative} and τ_{spike} are validation-selected warning thresholds, not model hyperparameters. The selected hyperparameter names follow the grid reported in Table D.2. Fixed implementation settings, such as the logistic-regression penalty and solver, are reported in Table D.1. Parameters not shown here follow the fixed implementation settings or the package defaults.

E

Extended feature-removal results

This appendix shows the extended results of the feature-group removal analysis used in Section 6.5. The main text focuses on macro F2 scores and the interpretation of the cross-border contribution. Table E.1 provides the additional metrics for the same analysis, including class-specific F_2 scores, PR-AUC values, probability-quality metrics, warning counts, and the leave-one-neighbour scenarios.

The table contains two types of feature-removal scenarios. The first group compares feature-block scenarios with the domestic baseline. These scenarios show which types of cross-border information improve performance beyond domestic features. The second group removes one neighbouring bidding zone at a time from the core cross-border feature set. These scenarios support the interpretation in Section 6.5 that the predictive value of cross-border information is not evenly distributed across neighbouring zones.

Table E.1: Extended feature-removal results on the 2025 test set

Scenario	Model	Features	Macro F2	Δ Macro F2	Neg. F2	Spike F2	Warnings	Brier	Neg. PR-AUC	Spike PR-AUC
Domestic baseline	Log. reg.	32	0.625	+0.000	0.653	0.548	3,787	0.460	0.754	0.447
Domestic baseline	GBDT	32	0.700	+0.000	0.671	0.640	2,774	0.121	0.776	0.562
Domestic baseline	XGBoost	32	0.702	+0.000	0.651	0.683	2,944	0.116	0.764	0.592
+ Scheduled exchange	Log. reg.	40	0.601	-0.024	0.633	0.536	4,120	0.465	0.659	0.449
+ Scheduled exchange	GBDT	40	0.720	+0.020	0.695	0.659	2,647	0.151	0.578	0.595
+ Scheduled exchange	XGBoost	40	0.729	+0.027	0.693	0.694	2,727	0.146	0.592	0.592
+ Transfer capacity	Log. reg.	42	0.621	-0.005	0.648	0.544	3,813	0.464	0.753	0.446
+ Transfer capacity	GBDT	42	0.699	-0.001	0.671	0.656	2,978	0.121	0.750	0.570
+ Transfer capacity	XGBoost	42	0.708	+0.006	0.652	0.693	2,895	0.117	0.762	0.592
+ Neighbouring system	Log. reg.	54	0.717	+0.091	0.744	0.634	3,016	0.359	0.810	0.624
+ Neighbouring system	GBDT	54	0.789	+0.089	0.762	0.737	2,116	0.100	0.814	0.749
+ Neighbouring system	XGBoost	54	0.791	+0.089	0.753	0.758	2,176	0.093	0.845	0.746
+ Neighbouring prices	Log. reg.	36	0.627	+0.002	0.660	0.541	3,730	0.452	0.760	0.432
+ Neighbouring prices	GBDT	36	0.690	-0.010	0.671	0.629	2,948	0.121	0.759	0.566
+ Neighbouring prices	XGBoost	36	0.727	+0.024	0.685	0.692	2,691	0.116	0.769	0.592
+ Neighbouring weather/calendar	Log. reg.	58	0.612	-0.014	0.629	0.551	3,952	0.427	0.769	0.461
+ Neighbouring weather/calendar	GBDT	58	0.700	-0.000	0.677	0.624	2,680	0.117	0.806	0.566
+ Neighbouring weather/calendar	XGBoost	58	0.706	+0.004	0.655	0.687	2,934	0.115	0.779	0.603
+ Core cross-border	Log. reg.	76	0.698	+0.072	0.716	0.626	3,184	0.351	0.767	0.616
+ Core cross-border	GBDT	76	0.808	+0.108	0.784	0.745	1,855	0.101	0.772	0.756
+ Core cross-border	XGBoost	76	0.813	+0.111	0.795	0.752	1,898	0.094	0.818	0.740
Full feature set	Log. reg.	102	0.734	+0.108	0.756	0.651	2,814	0.314	0.779	0.645
Full feature set	GBDT	102	0.806	+0.106	0.790	0.735	1,859	0.101	0.759	0.758
Full feature set	XGBoost	102	0.812	+0.110	0.793	0.752	1,908	0.095	0.810	0.747
Core without total capacity	Log. reg.	74	0.698	+0.000	0.716	0.627	3,180	0.352	0.765	0.617
Core without total capacity	GBDT	74	0.810	+0.002	0.799	0.736	1,866	0.099	0.775	0.754
Core without total capacity	XGBoost	74	0.811	-0.002	0.797	0.745	1,900	0.094	0.818	0.742
Core without DE/LU	Log. reg.	63	0.645	-0.053	0.661	0.581	3,649	0.409	0.727	0.553
Core without DE/LU	GBDT	63	0.768	-0.040	0.772	0.649	1,893	0.121	0.723	0.669
Core without DE/LU	XGBoost	63	0.786	-0.027	0.790	0.695	2,045	0.119	0.726	0.663
Core without BE	Log. reg.	63	0.684	-0.013	0.723	0.596	3,339	0.369	0.782	0.593
Core without BE	GBDT	63	0.817	+0.009	0.781	0.754	1,625	0.100	0.775	0.762
Core without BE	XGBoost	63	0.803	-0.010	0.773	0.753	1,964	0.097	0.801	0.727
Core without NO2	Log. reg.	64	0.715	+0.017	0.738	0.634	2,992	0.361	0.799	0.619
Core without NO2	GBDT	64	0.805	-0.003	0.799	0.735	2,003	0.093	0.819	0.754
Core without NO2	XGBoost	64	0.805	-0.008	0.785	0.750	2,023	0.089	0.846	0.738
Core without DK1	Log. reg.	64	0.710	+0.012	0.722	0.641	3,039	0.326	0.776	0.623
Core without DK1	GBDT	64	0.807	-0.001	0.790	0.746	1,973	0.101	0.777	0.748
Core without DK1	XGBoost	64	0.813	+0.000	0.796	0.752	1,900	0.095	0.821	0.749

Note: Δ Macro F2 reports the change relative to the relevant comparison scenario. For feature-block scenarios, it is calculated relative to the domestic baseline of the same model. For the core-without-total-capacity and leave-one-neighbour scenarios, it is calculated relative to the core cross-border scenario of the same model. Lower Brier scores indicate better probability quality.

F

Explainability outputs

This appendix reports the SHAP summary plots for the final XGBoost model. The main text reports the highest-ranked features by mean absolute SHAP value. The figures below add the direction and spread of feature effects across the final 2025 test sample. In these plots, points to the right increase the model output for the corresponding event class, while points to the left decrease it. The colour indicates whether the feature value is relatively high or low. As in the main text, these plots are interpreted as model-specific feature attributions and not as causal effects.

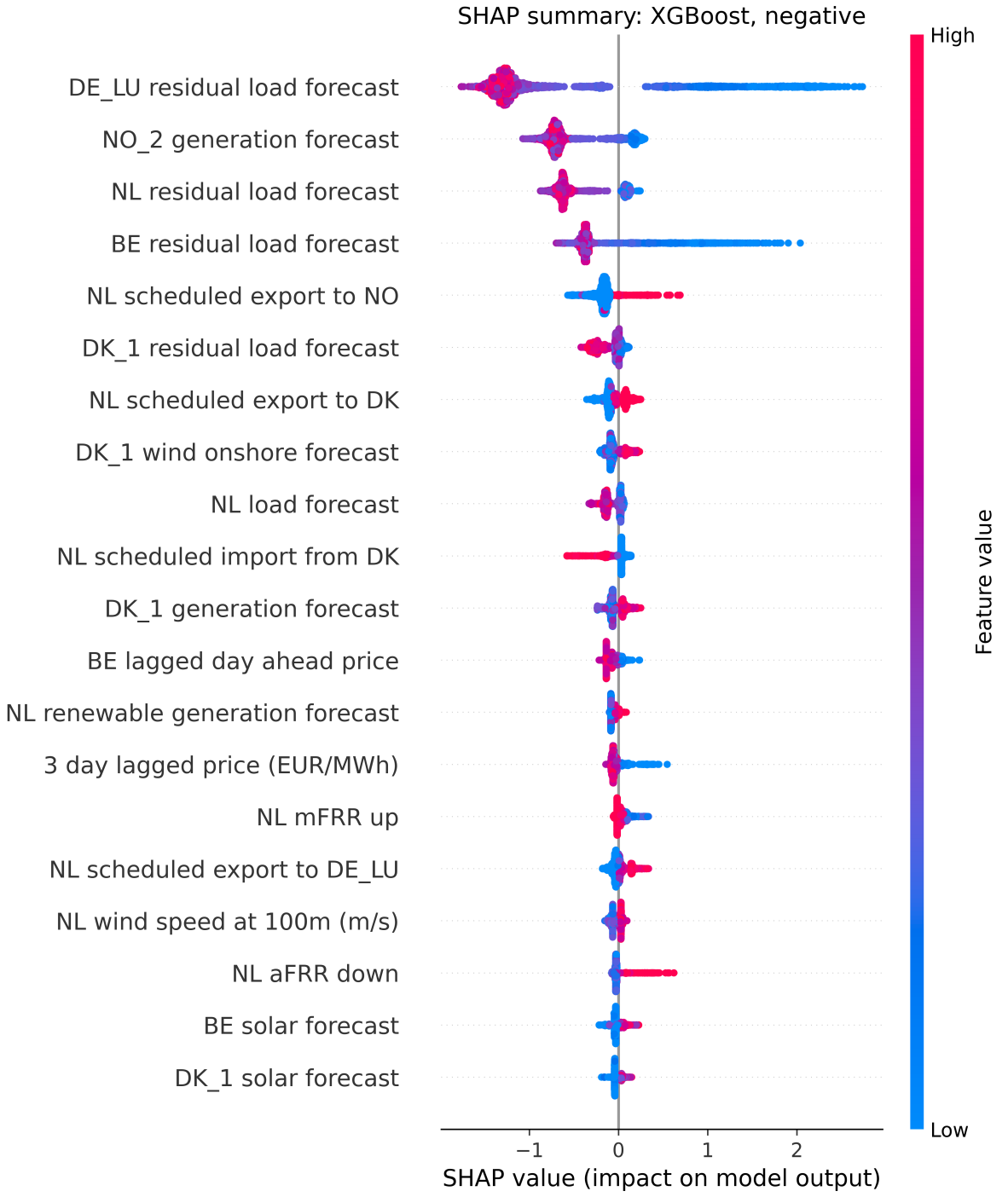

Figure F.1: SHAP summary plot for negative-price predictions of the final XGBoost model

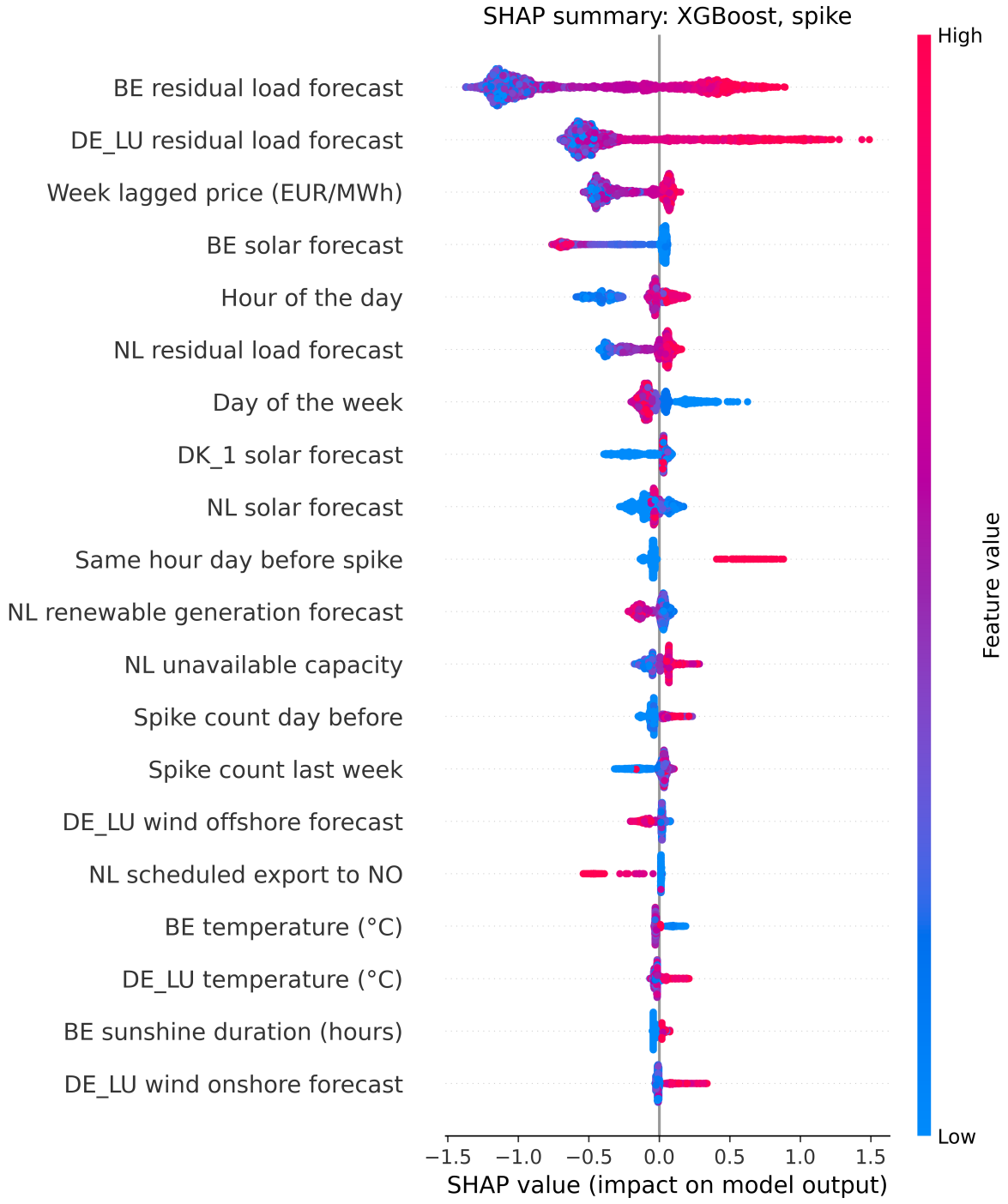

Figure F.2: SHAP summary plot for price-spike predictions of the final XGBoost model

G

AI statement

During this thesis, AI-based tools were used to support the research and writing process. ChatGPT was used for brainstorming, refining the formulation of the main research question and sub-questions, improving grammar and flow, cleaning the reference list, and reviewing sections or code for unnecessary repetition. ChatGPT and GitHub Copilot were both used to support debugging and code clean-up, for example when identifying possible causes of modelling errors, data mismatches, or inconsistencies. ChatGPT was also used to generate the cover image for the thesis.

The use of AI tools did not replace the research process or academic judgement. Suggestions generated by AI tools were assessed, edited, and verified before being implemented. Code-related suggestions were checked and tested before they were included in the final pipeline. AI tools were not used as independent academic sources. References and factual statements made by AI were checked against the original sources used in the thesis.

The final responsibility for the content, analysis, results, and conclusions of this thesis remains with the author.