Forecasting 30-Year Power Prices using a Distillation Model

Columbia MSBA Capstone Project | Final Presentation

MAY 2025

Recap Midpoint Presentation

Historical Analysis

Open Questions from Midpoint Presentation

Baseline Model (XGBoost)

Model Performance

Observed Issues & Patterns

Attempted Fixes & Outcomes

Evaluation of Final Model

Alternative Models & Comparison

ARIMA

LSTM

Conclusion

We use input variables from Afry and enrich them with features to reflect grid behavior in the dispatch model

	Data	Engineered	Description of variables	Unit
Demand – Fuel Prices	Wholesale price	No	Average annual wholesale price in a market area (excl. marginal rent)	EUR / MWh
	Gas Price	Yes	Average gas prices over different gas types	EUR / MWh
	Coal Price	Yes	Average coal prices over different coal types	EUR / MWh
	H2 Price	No	Levelized costs of H2	EUR / MWh
	EUA Price	No	EUA Prices	EUR / MWh
Demand	Average Demand	No	Average annual daily demand	TWh
	Peak Demand	Yes	Daily peak demand / average annual daily demand	%
Supply	Nuclear Capacity	No	Installed nuclear production capacity	MWh
	Gas Capacity	No	Installed gas power plants production capacity	MWh
	Coal Capacity	Yes	Installed coal power plants production capacity (includes different types)	MWh
	H2 Capacity	No	Installed H2 power plants production capacity	MWh
	Solar Capacity	Yes	Installed solar modules production capacity (includes different types)	MWh
	Wind Capacity	Yes	Installed wind turbine production capacity (includes offshore, shore and land)	MWh
	Storage Capacity	Yes	Amount of electricity to be provided to the grid by different types of energy storage	MWh
Grid	Installed Trade Capacity	Yes	Amount of import/export capacity of one price area to all other price areas	MWh
	Trade Capacity Availability	Yes	Utilization (%) of available capacity	%

1. Calculated based on local capacity, local demand and import/export caps

The cleaned data set includes 384 unique price curves

Afry Model Releases (x16)

- Quarterly releases
- 2021 Q1 to 2024 Q4

Electricity Market Scenarios (x3)

- Low
- Central
- High

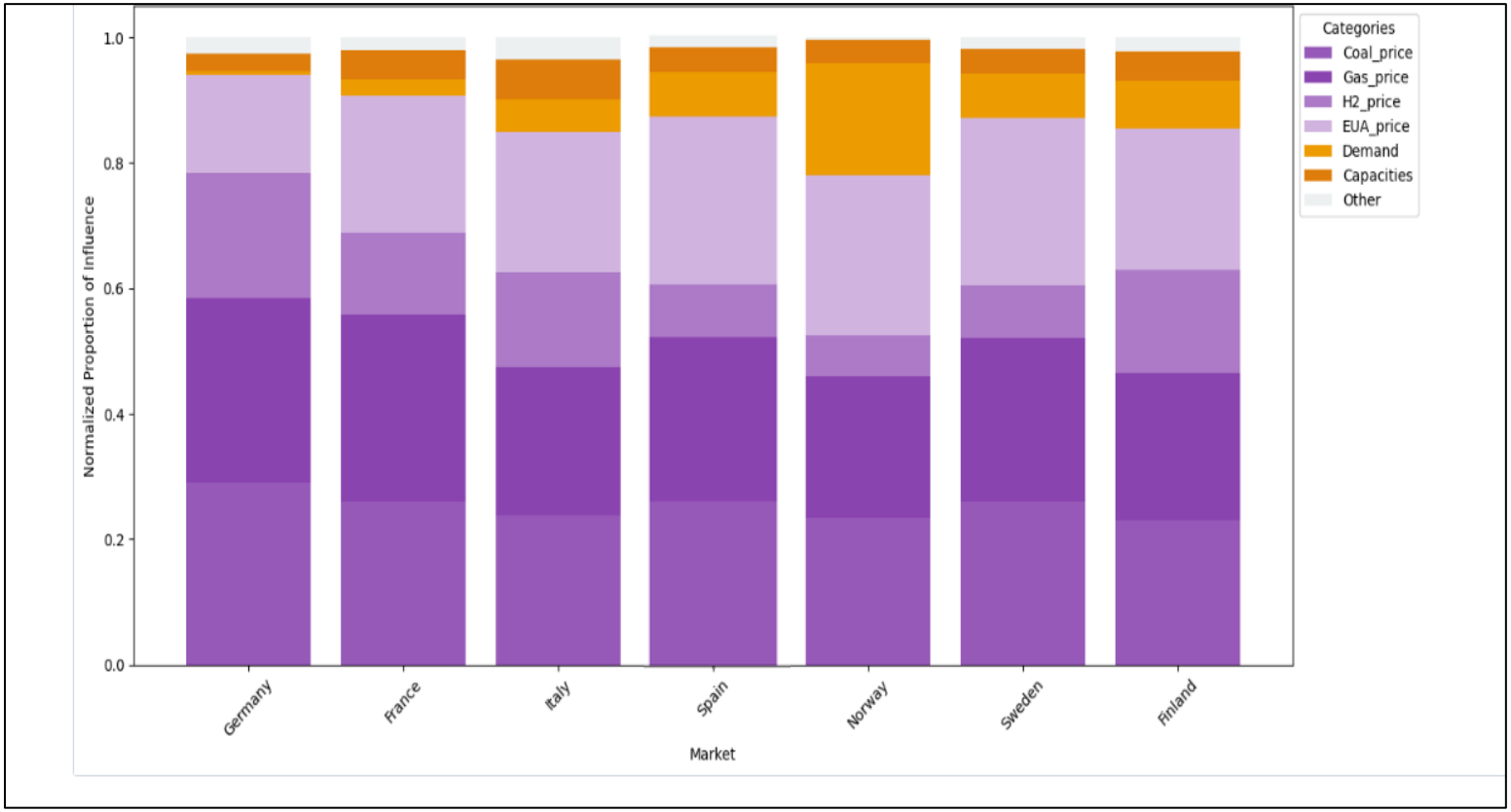
Electricity Price Areas (x10)

- Finland
- France
- Germany
- Italy (x3)
- Norway
- Spain
- Sweden (x2)

Note: Some releases or price areas excluded due to missing data/ questionable data quality

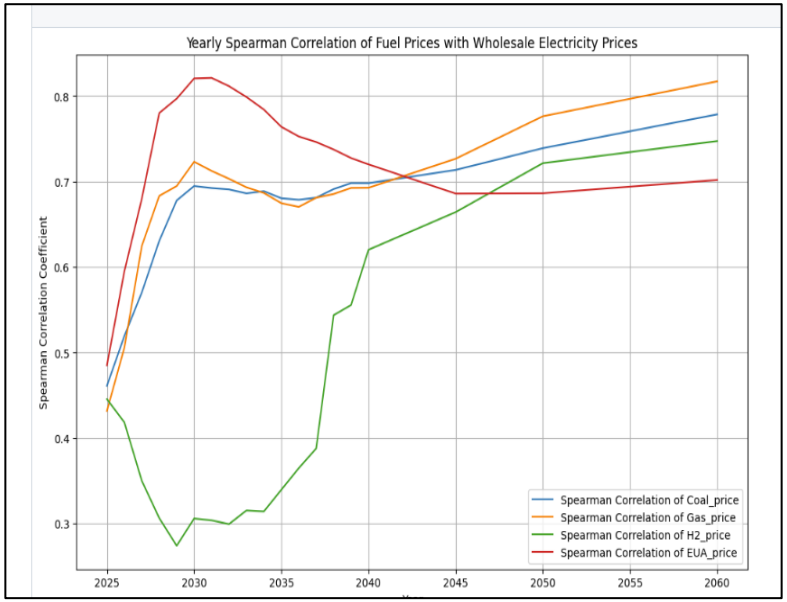
Changes in fuel prices (incl. EUA) are the most important factor when explaining changes in electricity price

Share of variance in wholesale price explained by input factors:
All years, scenarios and releases

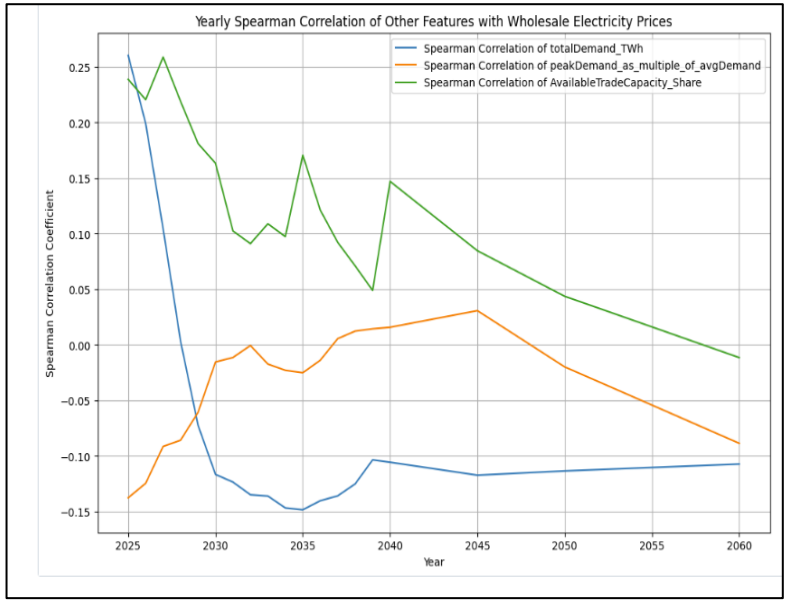


Input factor's importance varies over time

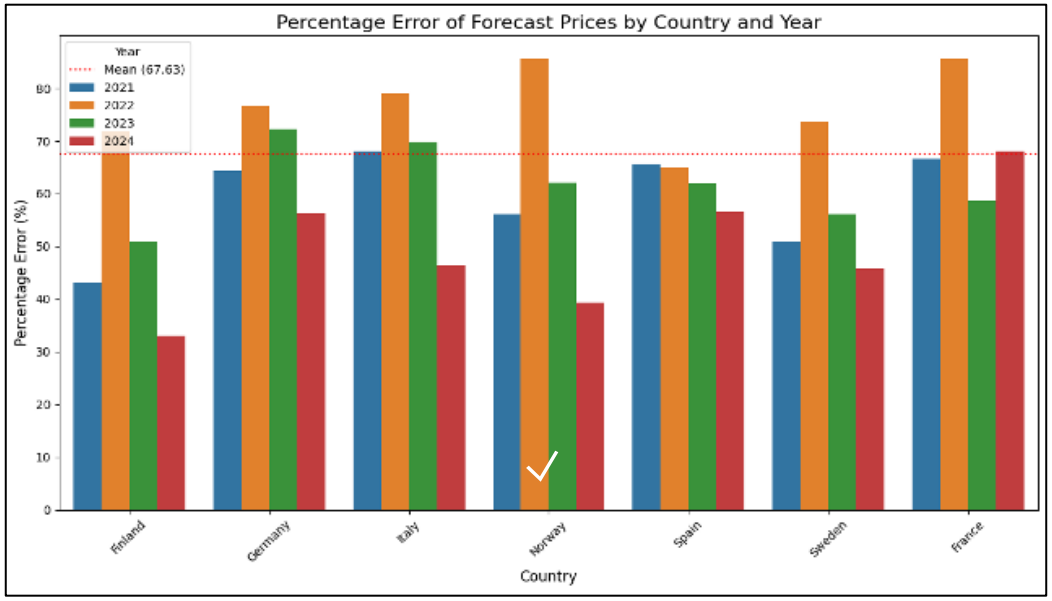
H2 price expected to become a determinant factor from 2040 onwards



Grid-related variables become less important in the long-run



We investigated your questions regarding forecast errors



Open questions from the mid-presentation

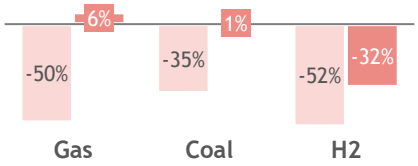
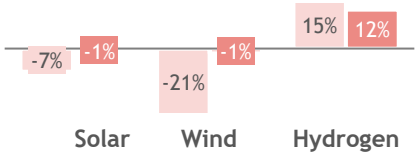
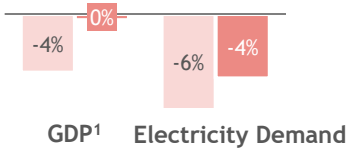
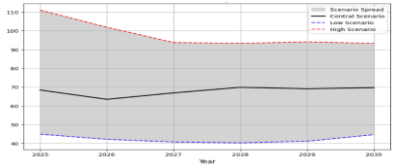
Based on our **previous work on the EDA and historical analysis**, we plan to make further revisions in the following areas:

1. Investigate the factors behind the high Afry forecast errors (especially in 2024, where the impact of the gas crisis is lower)
2. Gain deeper insights into market specifics, particularly focusing on France and Finland

1 Main input variables in Afry's model diverged from reality —fuel, renewables, and demand with largest deviations

Increase prices

Decrease prices

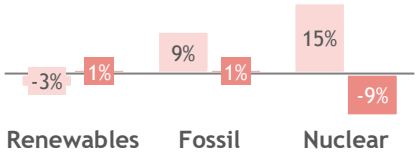
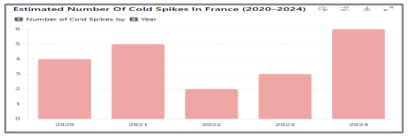
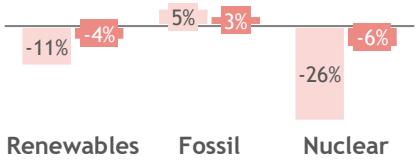
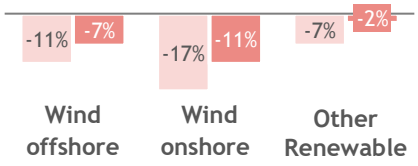
Category	Afry estimation vs. actual (%)	Most likely driver behind estimation error
Higher fuel prices		Gas markets stabilized faster than expected after the 2022 crisis, but AFRY's model likely assumed persistently high gas-driven power prices, leading to forecast overshooting. Source
Fewer green capacity		Renewable generation (wind) was stronger than expected across key markets in 2024, flooding the system with low-marginal-cost electricity and depressing prices beyond typical model assumptions. Source
Lower GDP growth		Structural demand shifts, such as sustained lower industrial electricity usage post-crisis and milder weather, caused load to be weaker than forecast models projected. Source
Higher complexity		AFRY's dispatch model likely struggles to anticipate complex post-crisis dynamics like cross-border flows, rapid capacity shifts, and non-linear demand recovery. Source

2021 estimation vs. 2024 actuals

2023 estimation vs. 2024 actuals

1. Not an Afry assumption, but likely same GDP assumption was used by Afry to model demand; Source IMF

2 Afry's 2024 forecast missed key supply and demand shifts in France and Finland —notably nuclear restarts, wind growth, and cold-driven demand volatility

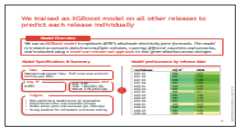
Category	Afry estimation vs. actual (%)	Most likely driver behind estimation error
More Nuclear Capacity		France's nuclear fleet returned strongly in 2024 after maintenance backlogs, adding low-cost capacity that pushed prices lower than models had predicted. Source
More Cold Spikes	 Cold spikes per year	France's demand profile remains extremely temperature-sensitive due to widespread use of electric heating, meaning that even in a low-price environment, cold spells can trigger sharp price spikes. Source
More Nuclear Capacity		Finland's commissioning of the Olkiluoto-3 reactor shifted its market from net importer to near self-sufficiency, significantly impacting regional price levels. Source
Higher complexity		Finland also experienced wind capacity growth, with wind now providing over 20% of production, leading to lower average prices but also increased variability and frequent negative pricing during high wind conditions. Source

Next steps after midpoint presentation

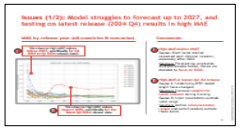
As we move into the **next (forward looking) phase of the project**, our priorities will include:

- The main objective remains to approximate the Afry dispatch model with a simpler, yet reliable model
- Our model should be able to reliably translate marginal changes in inputs into wholesale price chances
- Key Risk: The Afry model might not be reliably approximated by a simpler model. Should this become apparent during the course of the project, we will shift our focus and deliver a more insight-driven report.

Baseline model. We developed an XGBoost model to replicate AFRY's forecasts, trained on multiple scenario releases and evaluated using a leave-one-release-out method; strong overall performance, with some limitations

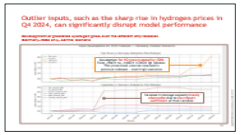


Limitations. The baseline model struggles in early years (pre-2028), varies by country, and performs poorly with extreme inputs outside its training range, such as in recent high-price or hydrogen-driven scenarios.

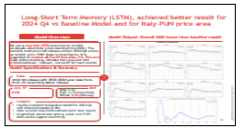


Solution finding. Weighting scenarios by recency showed minimal improvement, but might be beneficial for future use of the model; a country-clustering approach significantly improved accuracy

Validity range. We define a range for model use, focusing on moderate scenarios; model becomes unreliable under extreme price levels (e.g., >90-100 EUR/MWh) or outside the trained H2 costs input space.



Alternative methods. (ARIMA, LSTM) were explored, but did not outperform XGBoost or resolve key edge-case issues, confirming XGBoost as our primary modeling too mimic Afry's price forecasts



Recap Midpoint Presentation

Historical Analysis

Open Questions from Midpoint Presentation

Baseline Model (XGBoost)

Model Performance

Observed Issues & Patterns

Attempted Fixes & Outcomes

Evaluation of Final Model

Alternative Models & Comparison

ARIMA

LSTM

Conclusion

We trained an XGBoost model on all other releases to predict each release individually

Model Overview

We use an XGBoost model to replicate AFRY's wholesale electricity price forecasts. The model is trained on scenario data from multiple releases, covering different countries and scenarios, and evaluated using a leave-one-release-out approach to test generalization across vintages

Model Specifications & Summary

Data

Standard set except 'Italy - PUN' price area and with starting year 2024

Avg. R^2

0.953

MAE

Avg: 3.45
Top: 1.88 (2021 Q3)
Worst: 5.98 (2024 Q4)

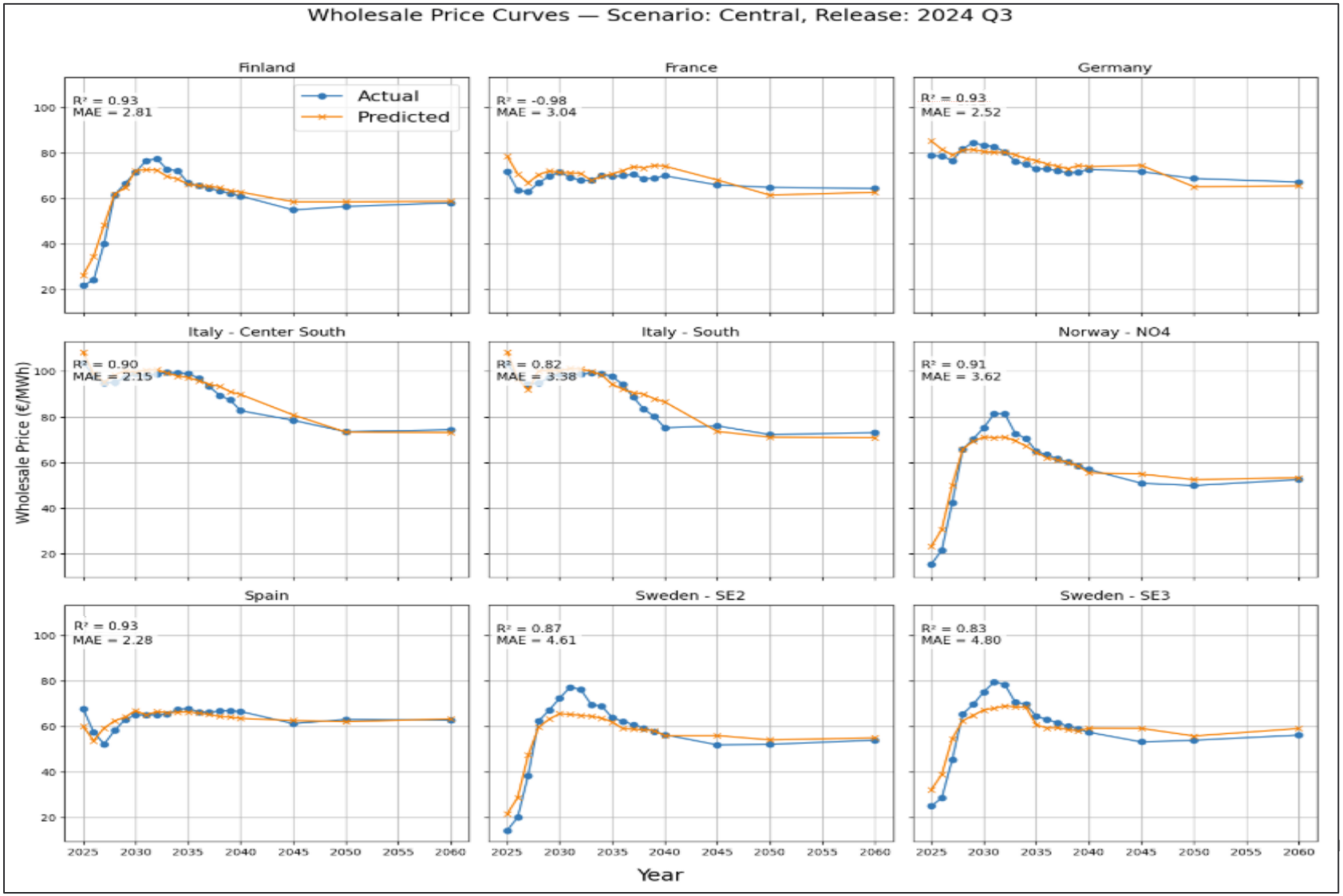
Insights

- Best-performing model across all approaches
- Outperforms linear and ensemble variants
- Stable across years and scenarios (ex. 2024 Q4)
- Strong baseline for refinement and stress testing

Model performance by release date

Test Release	Adj. R^2	MAE
2021 Q1	0.94	4.21
2021 Q2	0.97	2.36
2021 Q3	0.98	1.88
2021 Q4	0.97	2.45
2022 Q1	0.95	3.39
2022 Q2	0.97	3.01
2023 Q1	0.92	4.1
2023 Q2	0.97	3.56
2023 Q3	0.97	3.39
2023 Q4	0.97	3.17
2024 Q1	0.90	5.07
2024 Q2	0.98	2.86
2024 Q3	0.97	3.25
2024 Q4	0.92	5.98

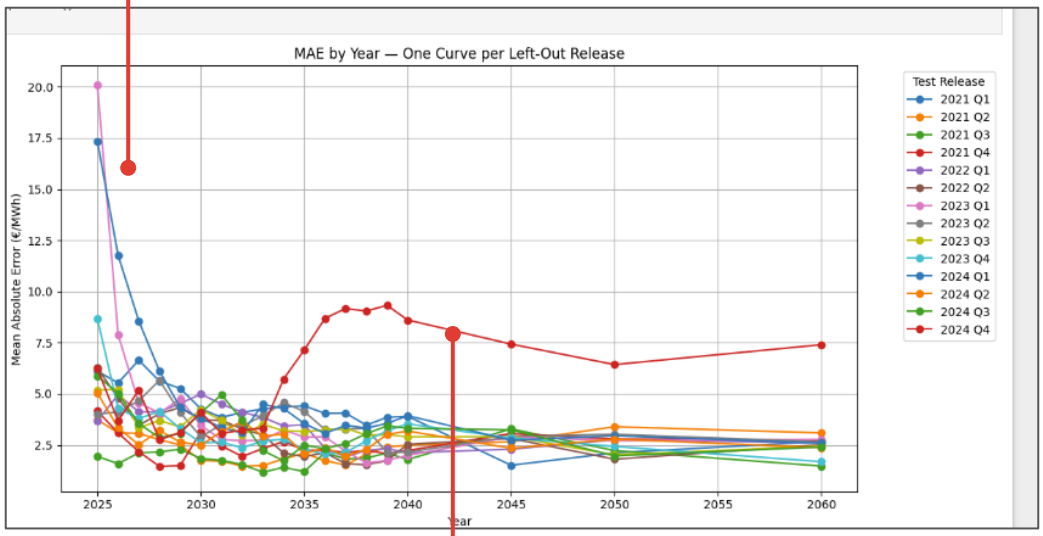
The model approximates one price curve per price area based on the respective input factors provided



Issues (1/2): Model struggles to forecast up to 2027, and testing on latest release (2024 Q4) results in high MAE

MAE by release year (all countries & scenarios)

1 We observe high MAE values before 2027, specifically for Q1 2023 to Q1 2024 release dates



2 We observe high MAE values beyond 2035, specifically for the latest Q4 2024 release date

Comments

1 ***High MAE before 2027***

Cause: Short-term market impacted post-Ukraine invasion, especially after 2022

Solution: Forecasting unreliable; multiple models failed. Hence we decided to focus on 2028+

2 ***High MAE in latest Q4 '24 release***

Cause 1: Underlying AFRY model might have changed

Solution: Increase weights for latest releases during training

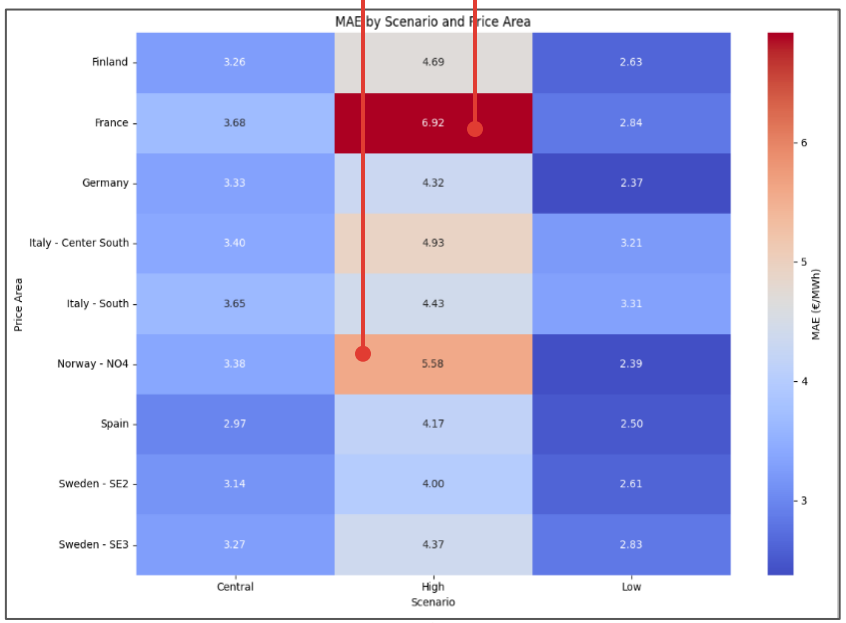
Cause 2: Input parameters outside valid range

Solution: Define valid parameter ranges and switch models outside those limits

Issues (2/2): The model performed significantly better in some countries and worse in others

MAE by release year (all countries & scenarios)

3 We observe higher MAE values for certain countries, e.g., France, Norway and Italy



4 The model performs worse in the higher-end range of the input values

Comments

3 ***Performance varies by country***

Cause: Different price areas might have different price-determining characteristics

Solution: Implement a clustering approach to address the individual market characteristics

4 ***Difficulties with extreme data***

Cause: Poor forecasting performance as soon as one input factor falls out of the range the model as been trained on

Solution: Define valid parameter ranges and switch models outside those limits

Improvements (1/2): Weighting releases by date to capture model shifts yields marginal performance gains

Model Overview

- **Linear weighting** reduces influence of releases gradually over time, ensuring stability
- **Exponential weighting** emphasizes recent data, making results more volatile and sensitive to short-term shifts

Model Specifications & Summary

Data

Standard set except Italy – PUN with starting year 2027

Adj. R^2

Avg: 0.961

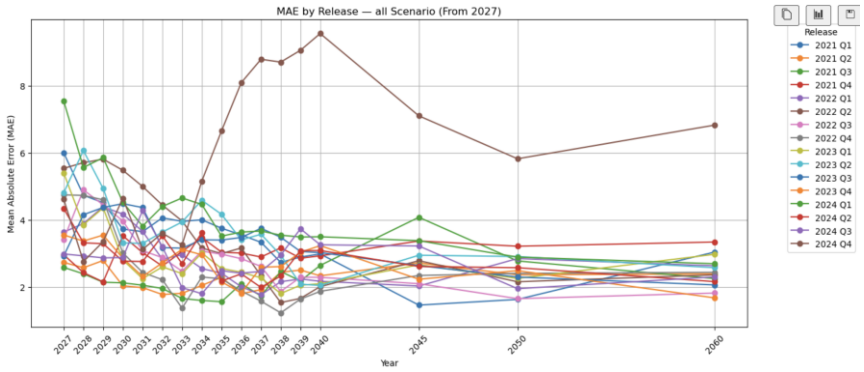
MAE

Avg: 3.40
Top: 2.25 (2021 Q3)
Worst: 5.90 (2024 Q4)

Insights

- Using exponential weights left performance essentially unchanged compared to linear.
- We resort to linear weights to allow for future flexibility and avoid spikes.

MSE by release date after weighting



Improvements (2/2): Implementing clustering solutions slightly improves stability of error vs. baseline model

Model Overview

- **KMeans clustering** applied to structural energy features (e.g. capacity, demand, fuel prices) to segment markets with similar dynamics
- Separate models trained within each cluster → improves **local fit** and captures **heterogeneous market behavior**
- Ensures **consistent train/test splits** and enables **targeted sensitivity analysis** within each cluster

Model Specifications & Summary

Data

Standard set except Italy – PUN with starting year 2027

Adj. R^2

Avg: **0.963**

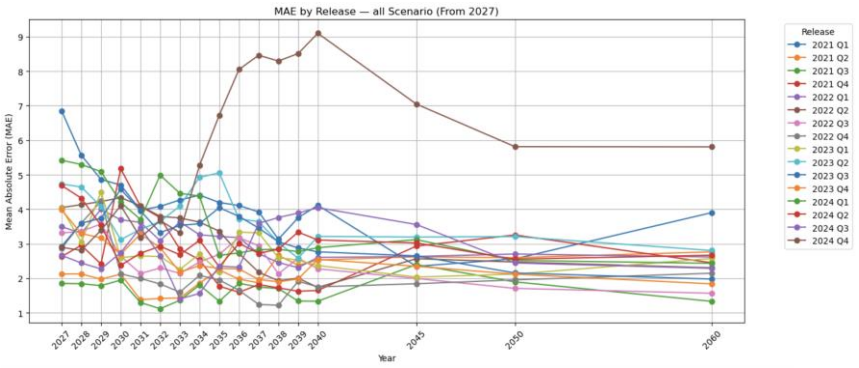
MAE

Avg: **3.26**
Top: 1.63 (2021 Q3)
Worst: 5.70 (2024 Q4)

Insights

- Maintain same Train routine
- Linear Weighting
- Helpful for stability and accuracy
- Defined 3 clusters

MSE by release date after weighting & clustering



Clustering approach splits data into three clusters

Cluster 1: 'Baseline'

- Covers all nine price areas - 60 % of dataset
- Use as reference cluster when explaining overall model fit

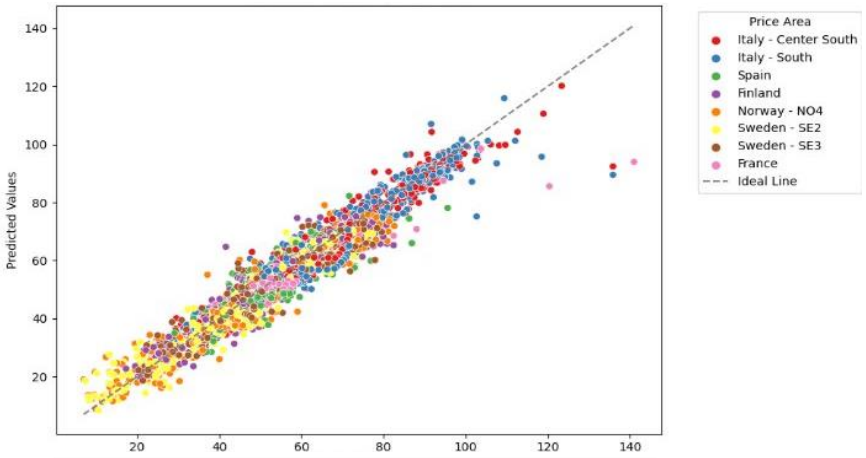
Cluster 2: 'High outliers'

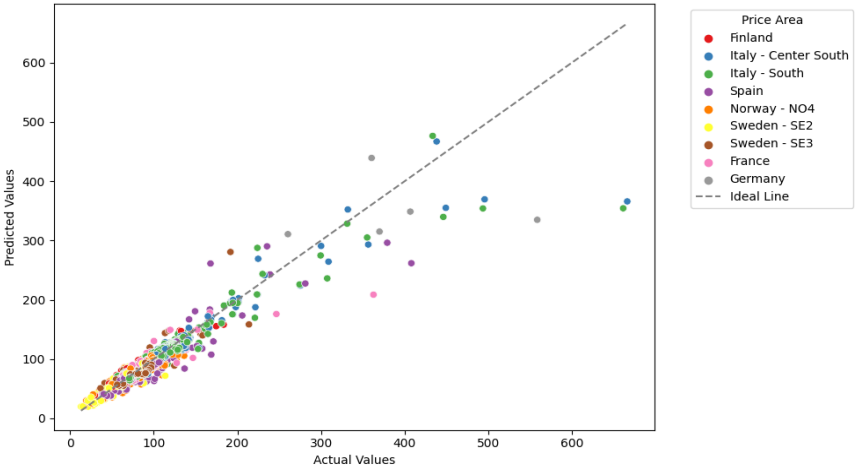
- Consists of few extreme actuals
- Extremely high data points (300–650 €/MWh) tend to be under-predicted
- Mix of all countries with Italy & Nordics being the most frequent
- High scenario MSEs spike for Italy - South & Center South
- **Action:** treat as “edge-of-distribution”

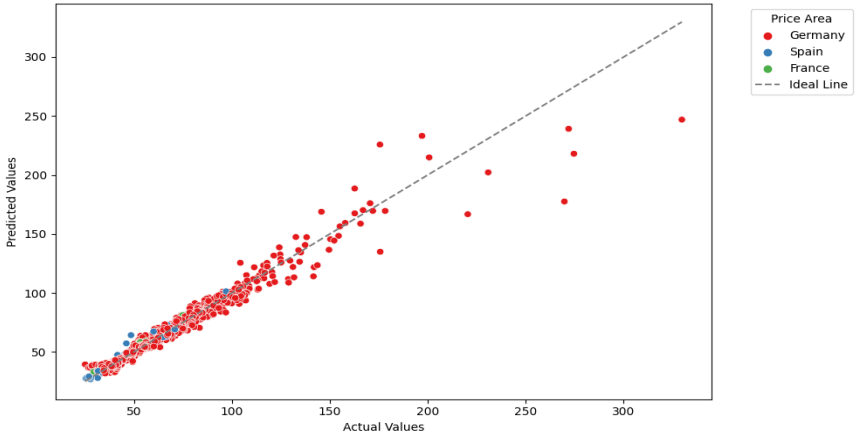
Cluster 3: 'Germany'

- Germany makes up 94% of data points
- Systematically under-predicts at high prices → steeper spread above 200 €/MWh

Actual vs Predicted per cluster







'Outliers' cluster 2 shows highest prediction error

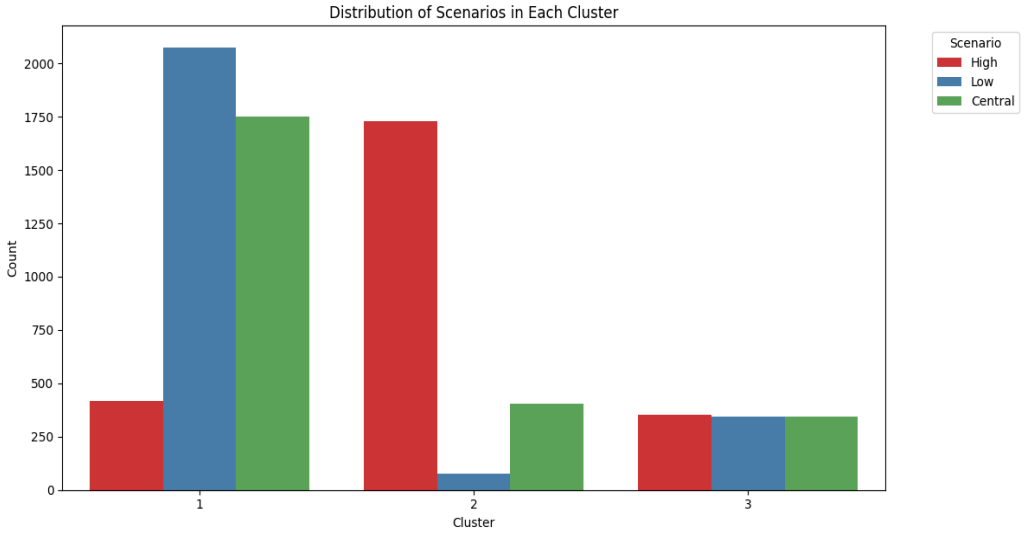
Comments

Cluster 1 (4 238 pts) – bulk of data, mostly Low/ Central; meets KPI by 2028

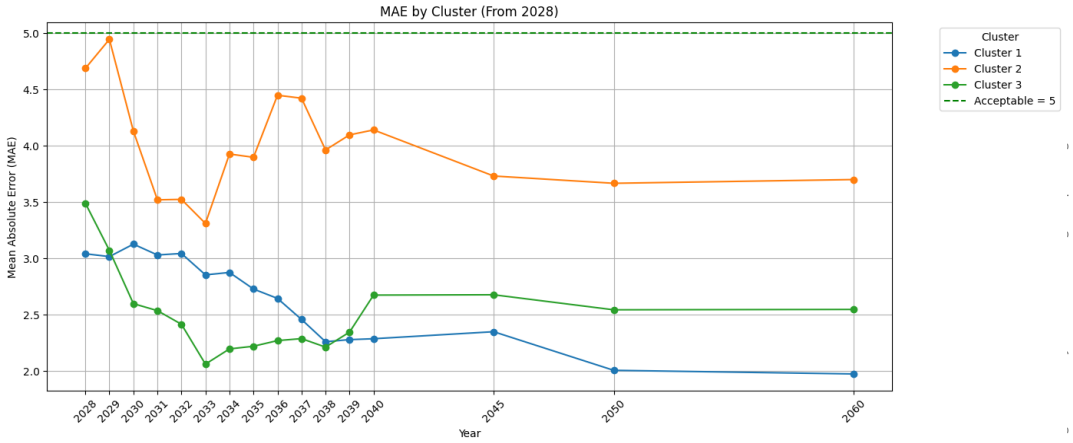
Cluster 2 (2 210 pts) – 90 % High; large early-horizon error

Cluster 3 (1 040 pts) – Germany-dominated; good accuracy but monitor long-term drift.

Distribution of Scenarios Each Cluster



MAE of each year by Cluster



Recap Midpoint Presentation

Historical Analysis

Open Questions from Midpoint Presentation

Baseline Model (XGBoost)

Model Performance

Observed Issues & Patterns

Attempted Fixes & Outcomes

Evaluation of Final Model

Alternative Models & Comparison

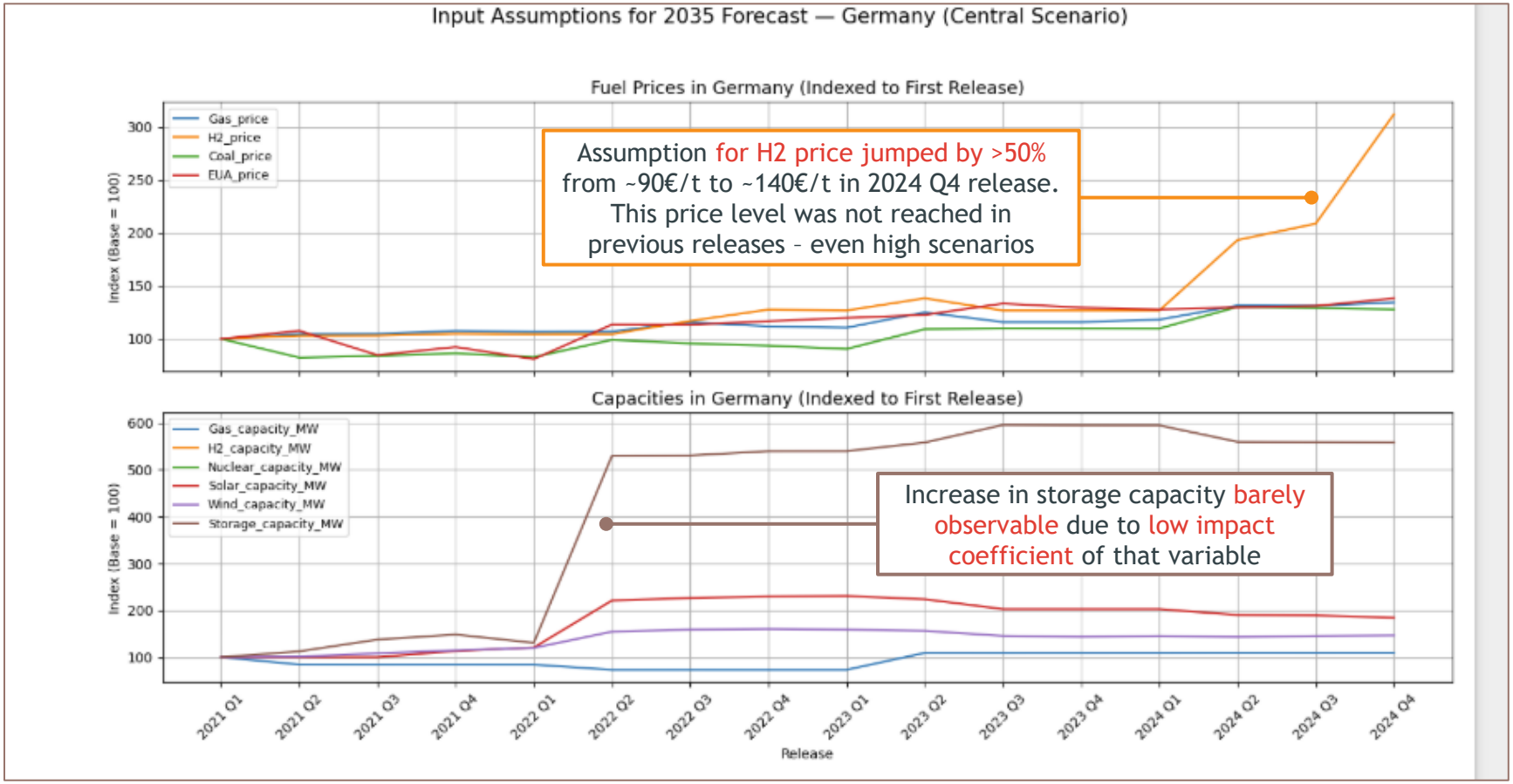
ARIMA

LSTM

Conclusion

Outlier inputs, such as the sharp rise in hydrogen prices in Q4 2024, can significantly disrupt model performance

Development of predicted Hydrogen price over the different Afry releases
Germany, 2035 only, central scenario



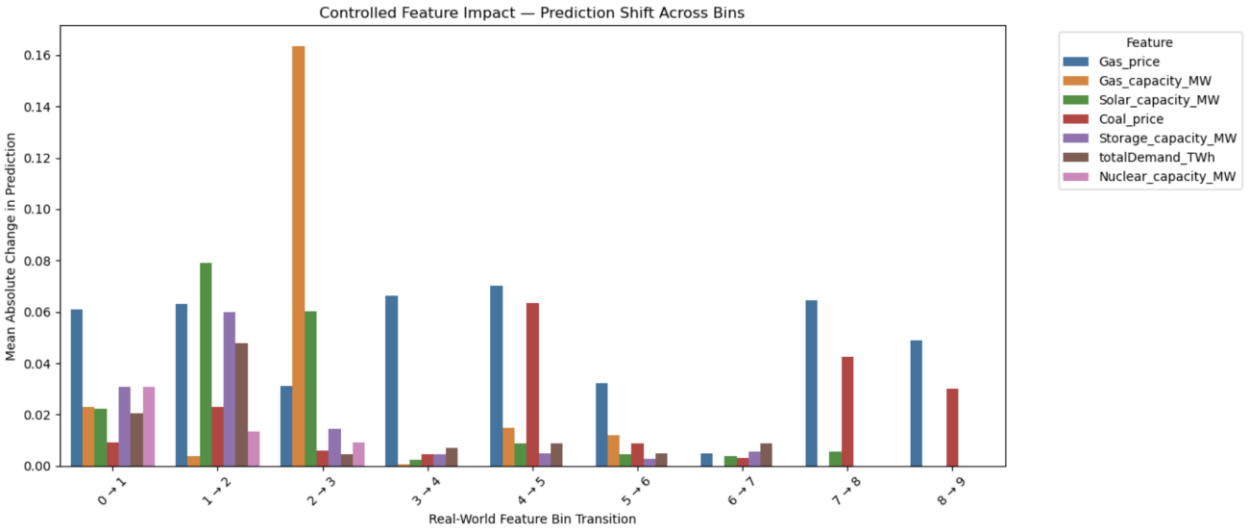
Stress Testing (1/3): Input changes in gas prices, gas capacity, and solar capacity lead to largest MAE increase

Overview

For each feature, we binned historical values into 10 quantiles and measured the **mean change in predicted electricity Price.** Then only that feature was perturbed from one bin to the next, while holding all other inputs constant.

Key Takeaways

- **Gas capacity** remains the most sensitive feature: even controlled, bin-to-bin changes can shift predicted prices by **over 16%**, suggesting high model responsiveness.
- **Gas price**, **solar capacity**, and **Coal Price** also show consistent predictive sensitivity.
- In contrast, **nuclear capacity** has limited effect on predictions, the model remains **robust even under large input changes.**
- These results inform **which inputs can be stressed** when simulating system shocks or validating model resilience



Stress Testing (3/3): Bootstrapped CI prediction shows that model evolves accurately over time

Overview

- Evaluates whether the model's uncertainty estimates align with actual forecast accuracy

Method

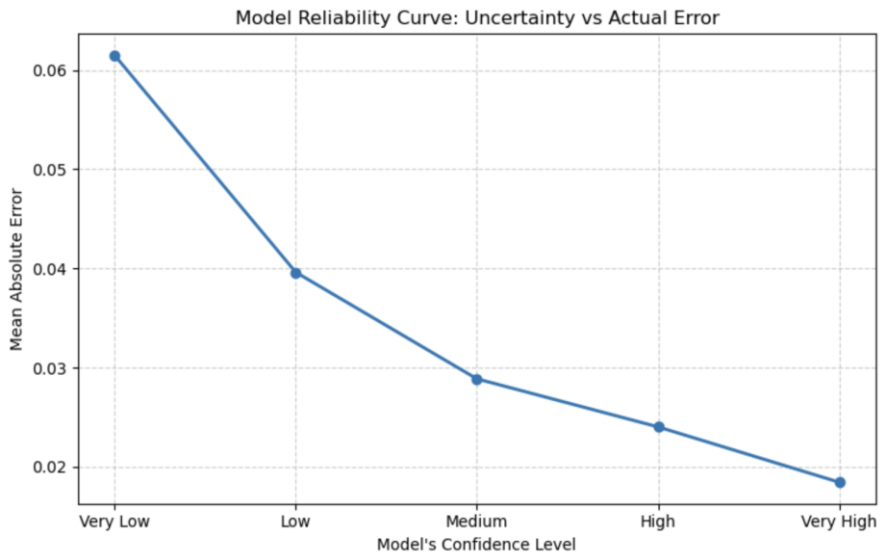
- Use bootstrapped prediction intervals through repeated resampling and retraining
- Compute confidence level based on interval width
- Group predictions by confidence bin and track mean absolute error

Key Insight

- Strong inverse relationship: predictions with narrower intervals (higher confidence) show lower error
- Indicates the model is well-calibrated and uncertainty estimates are meaningful

Implication

- The model can identify when its predictions are trustworthy, supporting confidence-weighted decisions



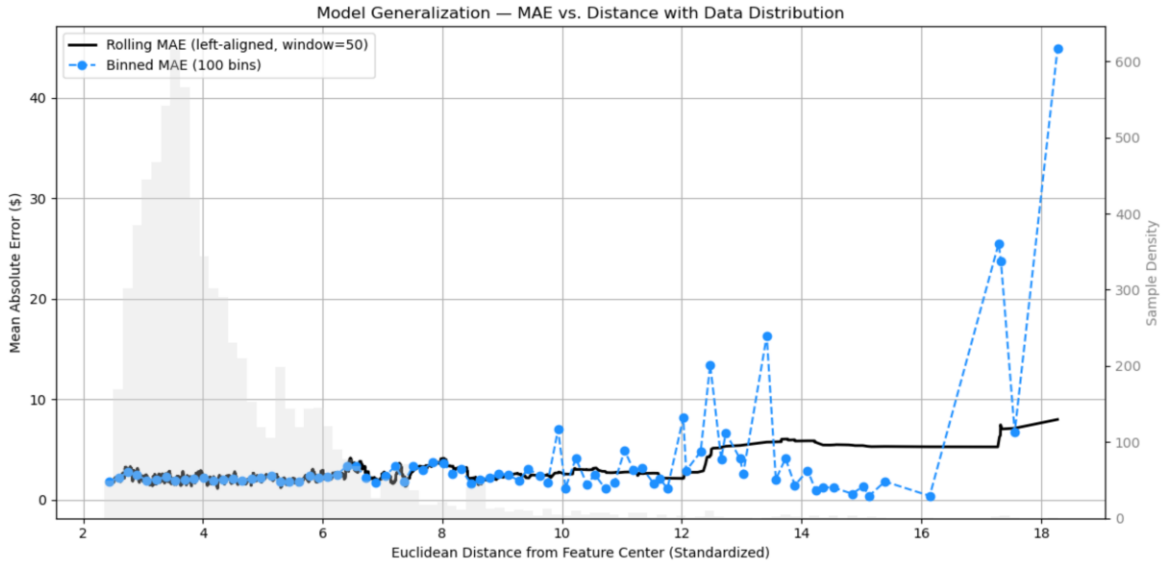
Stress Testing (2/3): Euclidean distance-based approach shows that the MAE grows with more unfamiliar inputs

Overview

We computed the Euclidean distance of each sample from the feature space center (using standardized features), then analyzed how **mean absolute error (MAE)** varies with distance. Two complementary methods were used: a **rolling window** average and a **binned aggregation**. They show how prediction accuracy deteriorates as inputs drift.

Key Takeaways

- Prediction error remains stable across most of the input space, but **rises sharply beyond a distance of ~12**, indicating the model struggles on out-of-distribution inputs.
- Central, well-sampled regions exhibit **low and consistent errors**, confirming model performance on typical scenarios.
- Results define a **trust boundary** : Inputs far from the center should be flagged or handled with fallback logic.



Recap Midpoint Presentation

Historical Analysis

Open Questions from Midpoint Presentation

Baseline Model (XGBoost)

Model Performance

Observed Issues & Patterns

Attempted Fixes & Outcomes

Evaluation of Final Model

Alternative Models & Comparison

ARIMA

LSTM

Conclusion

Long-Short Term Memory (LSTM), achieved better result for 2024 Q4 vs Baseline Model and for Italy-PUN price area

Model Overview

We use a **4-period LSTM** sequence to predict wholesale electricity price starting from 2026. The model is trained on all release except 2024 Q4 where we tested upon. With feature importance, it is suggested to **remove all Fuel Prices (Gas, H2, EUA and Coal)** before training. All data then grouped into 'priceAreaName', 'release', 'scenario' for each model

Model Specifications & Summary

Data

2024 Q4 release with 2022-2024 year data from their corresponding latest release.

Adj. R^2

0.92

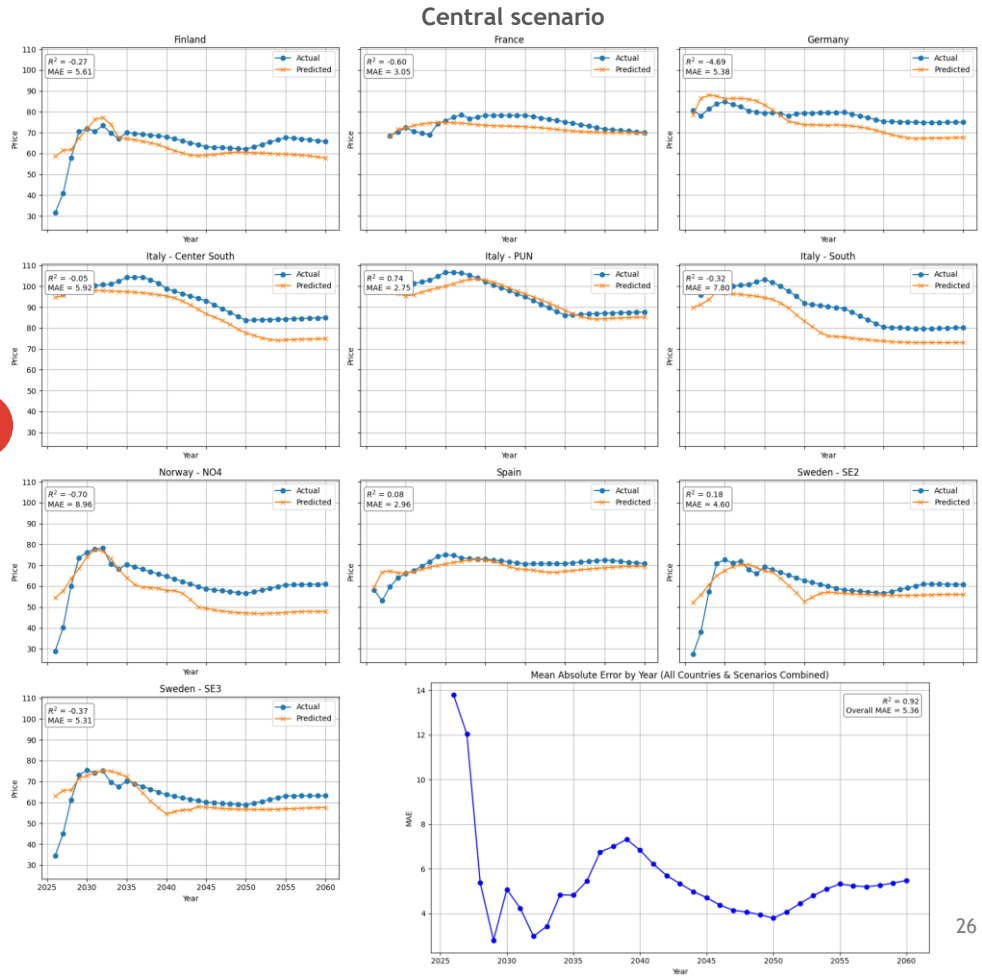
MAE

Avg: 5.36
Top: 2.75 (Italy-PUN)
Worst: 8.96 (Norway)

Insights

- Slightly outperform baseline model for 2024 Q4 and improved model (5.90)
- Able to solve Italy-PUN problem (with best result)
- Unsatisfied robustness testing result with 7.95 MAE, could suggest overfitting

Model Output: Overall MAE lower than baseline model



Recap Midpoint Presentation

Historical Analysis

Open Questions from Midpoint Presentation

Baseline Model (XGBoost)

Model Performance

Observed Issues & Patterns

Attempted Fixes & Outcomes

Evaluation of Final Model

Alternative Models & Comparison

ARIMA

LSTM

Conclusion

Project recap: XGBoost-based model delivers accurate local forecasts and is now being integrated into Databricks

- Final Model: **XGBoost approach** with linear input weighting and country-based clustering, achieving **high accuracy** (MAE ~ €3/MWh) in replicating AFRY price curves for 2028–2050
- The model is **validated using bootstrapped confidence intervals**, which assess how uncertainty estimates align with actual forecast errors
- Still, the model performance can **degrade under outlier assumptions** — for example, extreme input shifts like the hydrogen price spike in Q4 2024
- Hence, model performs best as a **local approximation tool**, meaning it is most reliable when inputs remain within the range of historical data used during training
- We are **currently implementing** the model code in **Ardian Databricks**, and it will soon be available for testing

Overview: Columbia Analytics in Practice Capstone Team

Mark Tassanasunthornwong

- B.Sc. In **Business Economics** (Bangkok, TH)
- Work experience in **e-Commerce** and in the **energy industry**
- *wt2378@columbia.edu*

Filippo Di Fazio

- B.Sc. in **Finance** (Bayes)
- Internship experience as an **M&A** and **private equity** analyst
- *fd2577@columbia.edu*

Martina Paez Berru

- B.Sc. In **Math and Economics** (École Polytechnique de Paris)
- Internship experience in **statistical economic research**
- *mp4395@columbia.edu*

Simon Marchart

- B.Sc. and M.Sc. in **Financial Economics** (Mannheim, HSG)
- Work experience in **venture capital** and in **consulting**
- *spm2194@columbia.edu*

Jenny Chao

- B.Sc. in **Quantitiative Economics** (Emory)
- Internship experience as a **quantitative trading** analyst
- *rc3695@columbia.edu*