NTNU

Kunnskap for en bedre verden

Department of Computer Science

TDT4259 - Applied Data Science

Electricity Spot Prices Forecasting

Author:
Pol Fonoyet - 128703 - polf@stud.ntnu.no
Gorka Parra - 128705 - gorkap@stud.ntnu.no
Ronja Steinfurth - 128730 - ronjacst@stud.ntnu.no
Ana Barrera Novas - 128680 - anabar@stud.ntnu.no
Pol Batallé Largo - 128691 - polba@stud.ntnu.no
Daniel Alcolea Aguado - 128695 - danieale@stud.ntnu.no

Autumn 2024

Table of Contents

1 Introduction and problem definition

This report is the result of a group assignment for the course Applied Data Science TDT4259 at the Norwegian University of Science and Technology. Throughout the course, we have explored and applied data science concepts critical to understanding and addressing real-world problems. Our project focuses on Day-ahead Electricity Spot Price Forecasting, a problem introduced by a research scientist from SINTEF, one of Europe's leading independent research organizations. Using the dataset provided and combining our previous experience in the domain with new information and research, our team aims to create a robust analytical model capable of forecasting electricity prices in the Nord Pool market with high accuracy. Now we will give a short overview of NordPool, the identified problem and the team, but the first two will be further developed in 2.Background.

1.1 NordPool

Nord Pool is the largest electricity market in Europe, facilitating the efficient trading of electricity across multiple countries. Established to enhance the coordination and optimization of electricity markets, Nord Pool plays a crucial role in balancing supply and demand in real-time. The market operates in a highly dynamic environment, where prices are determined through the interaction of buyers and sellers, ensuring transparency and competitiveness.

In this market, participants place bids for electricity for the next 24 hours. The prices are set one day before delivery, based on the equilibrium between supply and demand. This market mechanism allows participants to plan their production or consumption schedules efficiently, ensuring that electricity needs are met at the lowest possible cost.

Nord Pool's role goes beyond price-setting; by facilitating a competitive and transparent environment, it ensures that electricity is produced and consumed in alignment with real-time market conditions. This system is integral in reducing the likelihood of supply shortfalls or surpluses, ultimately supporting a resilient and responsive energy market.

1.2 Problem definition

Our aim is to develop a model to forecast electricity spot prices in the Nord Pool market using historical spot price and volume data.

This problem statement sets the groundwork for creating predictive models capable of accurately forecasting future electricity prices within the Nord Pool market. These forecasts, referred to as spot price predictions, are essential for enhancing market efficiency and stability.

Despite Nord Pool's role in facilitating electricity trading, the focus of our project is to leverage data analytics to improve price forecasting accuracy. The main benefit of these forecasts lies in aiding various stakeholders, including energy producers, traders, and large industrial consumers, in making informed decisions regarding electricity production, trading, and consumption, which could lead to contributing to more efficient energy usage and cost reduction.

In tackling this problem, we recognize several challenges intrinsic to electricity price forecasting. First of all, electricity prices are highly volatile due to factors such as weather variations, demand fluctuations, and energy source availability. The model must account for these complexities to provide reliable forecasts. Also, accurate forecasting depends on the quality of historical data. We need to assess and preprocess the dataset to handle any inconsistencies, missing values, or anomalies that could affect model performance. After that, choosing the right forecasting model is

crucial. We aim to experiment with various machine learning and statistical approaches to identify the model that best captures the data patterns. Model accuracy and stability will be evaluated based on metrics relevant to time-series forecasting. And lastly, different stakeholders have unique needs and risk tolerances. For instance, industrial consumers prioritize price stability to manage operational costs, while traders may seek to leverage price fluctuations. Our forecasting model must be adaptable to meet diverse stakeholder objectives.

1.3 Team description and responsibilities

Name	Background	Responsibilities
Pol Fonoyet	Exchange student from the Polytechnic University of Catalonia, pursuing a Bachelor's degree in Informatics Engineering with a major in Computing.	Preparing data, analytical research and model creation
Pol Batalle	Exchange student from the University of Barcelona, pursuing a Bachelor's degree in Telecommunication Electronic Engineering	Visualization and data analysis
Jacob Clements		Business insight and understanding of business needs
Ana Barrera	Exchange student from the University of A Coruña studying a bachelor degree in Computer Science, with a focus in Software Engineering	Business insight, project design, and video creation
Gorka Parra		Preparing data, analytical research and model creation
Ronja Steinfurth	Exchange student from RWTH Aachen University, studying Electrical Engineering and Business Administration at Master's level	Business insight, understanding business needs and project planning
Daniel Alcoeda		Visualization and data analysis

The team established since the beginning a proper dynamic to work efficiently while staying in contact and ensuring everyone was part of the most important decisions and project planning. We made several meetings and created checkpoints throughout the development of the project. This way we made sure everything was working at the right pace for everyone, and also everyone stayed tuned and could offer suggestions of modifications of the different parts done since the last checkpoint.

2 Background

Forecasting electricity spot prices is a vital tool for stakeholders in the energy sector, particularly energy producers, who need to make informed decisions on production scheduling, bidding strategies, and resource management. In this project, we aim to develop an accurate forecasting model for the Nord Pool market using historical data on demand, production, and spot prices. The model will enable energy producers to optimize operations by predicting price fluctuations and aligning production with demand trends, ultimately enhancing their revenue and operational efficiency.

This chapter is organized as follows: First, we outline the specific business and technical objectives of the project in Section 2.1, focusing on the needs of energy producers. Following this, we present the problem approach in Section 2.2, detailing how we preprocess the data, apply feature engineering, and select the forecasting model. Next, we introduce the CRISP-DM framework and Design Thinking methodology in Section 2.3, explaining how each methodology guides the project's execution. Finally, we discuss the integration of these methodologies and how they enhance our model development, evaluation, and deployment strategies.

2.1 Objectives

In our electricity spot price forecasting project, various stakeholders may benefit from accurate predictions. This section outlines the specific objectives that should be prioritized when considering energy producers as primary stakeholders. By addressing these objectives, the forecasting model will support energy producers in making informed decisions about production, bidding, and resource allocation in the Nord Pool market.

Business Objectives

- **Enhance Demand-Supply Alignment**: By forecasting spot prices, energy producers can adjust production schedules to align closely with demand trends, maximizing revenue during high-demand, high-price periods.
- **Inform Strategic Bidding in Nord Pool Markets**: Leveraging price forecasts, producers can develop data-driven bidding strategies in the day-ahead markets, ensuring competitive and profitable positioning.
- **Improve Short-Term Planning and Risk Management**: Forecast insights allow producers to adapt schedules based on anticipated price fluctuations, reducing exposure to unprofitable hours and enhancing overall financial stability.
- **Optimize Resource Allocation for Cost Efficiency**: Accurate price predictions enable better resource allocation, minimizing costs related to overproduction or underproduction.
- **Increase Revenue from Flexible Assets**: For assets capable of rapid adjustments, such as gas or hydro plants, producers can use price trends to identify and capitalize on profitable adjustments in near real time.

Technical Objectives

- **Develop an Accurate Spot Price Forecasting Model**: Build a model to predict spot prices using historical trends in demand, production, and price data, capturing the critical relationships among these variables.
- **Integrate Temporal Patterns**: Utilize the datetime information to model daily, weekly, and seasonal price patterns, improving forecast accuracy by accounting for typical demand and supply fluctuations.

- **Model the Volume-Price Relationship**: Use `volume_demand` and `volume_production` data to understand how shifts in supply and demand impact prices, ensuring the model can respond to market dynamics.
- **Adapt for Day-Ahead Market Needs**: Design the forecasting model to incorporate updated data, enabling it to adjust predictions as new volume and price information becomes available, thereby supporting short-term decision-making.
- **Ensure Robust Model Evaluation and Performance Tracking**: Given the limited dataset, prioritize model accuracy through robust evaluation techniques to ensure the forecast remains reliable for immediate planning and decision-making.

2.2 Problem Approach

This project aims to forecast electricity spot prices in the Nord Pool market by leveraging historical data on demand, production, and spot prices. First, we will preprocess the dataset to ensure data quality by addressing any missing values, outliers, or inconsistencies. Next, feature engineering techniques will be applied to capture important patterns and seasonal trends that are likely to impact spot prices, such as daily, weekly, and seasonal demand fluctuations.

We will then develop a time series forecasting model tailored to the Nord Pool data, selecting an approach that best fits the characteristics of price data, such as machine learning-based methods. Evaluation metrics, including (???), will be used to assess model accuracy and guide improvements. By carefully validating and testing the model, we aim to deliver a reliable forecasting tool that helps energy producers optimize production schedules, bidding strategies, and resource allocation.

2.3 Project Management Strategy

This chapter outlines the CRISP-DM framework, a widely adopted methodology for managing data mining projects. The structure of this report is organized around the distinct phases of the CRISP-DM framework, highlighting the specific tasks and objectives within each phase. Additionally, Design Thinking serves as an important complement to CRISP-DM, enhancing the project management strategy by emphasizing user-centered approaches and iterative problem-solving throughout the data mining process.

2.3.1 Crisp-DM Framework

The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, developed in 1996, is a well-established framework for guiding data mining and data science projects. As an open standard, it provides a structured, field-proven approach that has been widely adopted across industries. CRISP-DM serves as a comprehensive process model, outlining the typical phases of a data project, detailing each phase's specific tasks, and clarifying the relationships between them, thus offering a clear overview of the entire data mining life cycle.

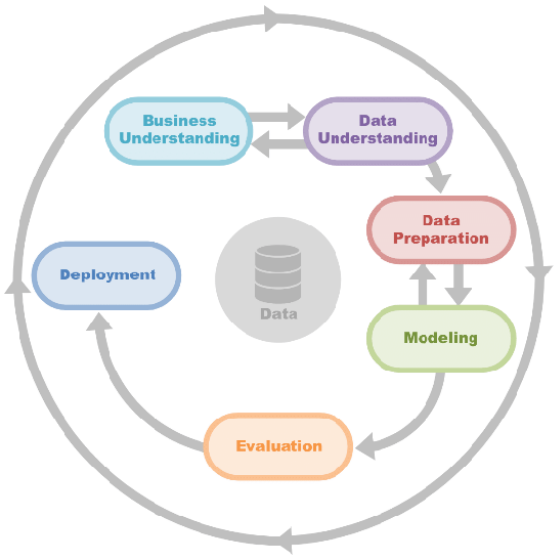

Figure 1: CRISP-DM Framework [TAH20]

This methodology divides the data mining process into six essential phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment, as illustrated in Figure 1. Each phase serves a distinct purpose and supports iterative improvement, enabling agile transitions that adapt to evolving project requirements. The CRISP-DM structure not only ensures alignment between technical efforts and business objectives but also facilitates effective project management by clearly defining each step's role in producing actionable results. A comprehensive overview of each phase is presented in the following sections, drawing on insights from [She00] for further detail.

Phase 1: Business Understanding

The Business Understanding phase sets the project's foundation by defining objectives from a business perspective and aligning them with strategic goals. Key steps involve setting objectives, assessing the current situation, and outlining a project plan to ensure relevant data aligns with business needs.

Phase 2: Data Understanding

Data Understanding begins with data collection and familiarization, identifying quality issues, and uncovering patterns. Steps include gathering relevant data, describing characteristics, exploring patterns, and verifying quality to ensure reliable analysis.

Phase 3: Data Preparation

Data Preparation transitions raw data to a final dataset suitable for modeling. Key tasks include selecting relevant data, cleansing inaccuracies, constructing new variables, integrating sources, and formatting for compatibility with modeling tools.

Phase 4: Modeling

In the Modeling phase, various techniques are applied, and parameters optimized. Key steps include selecting algorithms, designing tests, building models, and assessing effectiveness, generating actionable insights from the data.

Phase 5: Evaluation

The Evaluation phase critically assesses the model's performance, ensuring alignment with business objectives. Key actions include result evaluation, process review, and determining next steps for decision-making.

Phase 6: Deployment

Deployment translates insights into actionable formats, from reports to integrated "live" models. Key actions involve planning deployment, establishing monitoring, preparing reports, and conducting a project review to support practical application.

2.3.2 Overview of CRISP-DM Framework Integration

This section provides a concise summary of how each phase of the CRISP-DM methodology is applied across different sections of this report. Each phase contributes unique steps that are critical for addressing the project's objectives, guiding the analysis, and ensuring the model's practical application.

In Section 1.2 and 2.1, the Business Understanding phase is implemented, setting the foundation for defining project goals and aligning them with strategic objectives. The Data Understanding phase, outlined in Section 3.1 and 3.2, focuses on initial data exploration and quality assessment, laying the groundwork for informed analysis.

Section 3.3 documents the Data Preparation phase, detailing the transformation, cleaning, and structuring of data necessary for robust modeling. Modeling is covered in Section 3.2 and 4, where various techniques are applied, and their parameters calibrated to ensure optimal outcomes. In Section 4, the Evaluation phase rigorously tests the model's performance and ensures alignment with business requirements.

Finally, the Deployment phase is addressed in Section 5 and 6, highlighting how the model's insights are integrated into actionable formats to support decision-making and long-term project goals.

Each of these sections illustrates how the phases of the CRISP-DM framework contribute to achieving a structured, actionable data science project.

2.3.3 Design Thinking

Design Thinking, though defined in various ways, consistently highlights core elements. Some perspectives emphasize its focus on human-centered, needs-based problem-solving, while others underscore its capacity to drive entrepreneurial innovation. For example, [BU16] describes Design Thinking as a process that starts with human needs and harnesses technology to create value for both customers and businesses. Rather than following traditional "analytical" management methods, Design Thinking complements the structured CRISP-DM framework by encouraging creative, customer-focused solutions within data-driven projects.

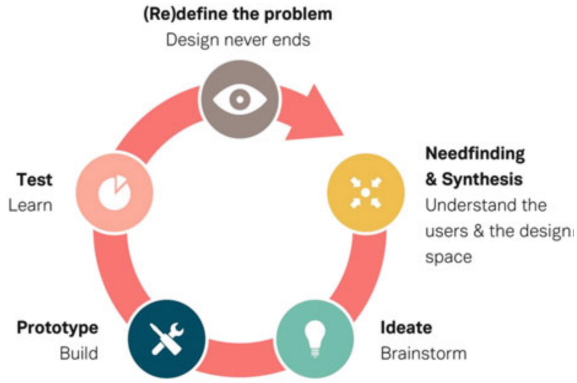

Figure 2: Design Thinking Micro-Process [BU16]

[BU16] describes the Design Thinking micro-process as an iterative cycle of six steps, as displayed in Figure 2:

1. Define the Problem: Formulate a focused yet open-ended challenge question to frame the project goals.
2. Needfinding and Synthesis: Investigate and interpret customer needs, capturing both obvious and hidden insights through research and direct interactions.
3. Ideate: Generate diverse ideas in response to the identified needs, encouraging innovative thinking grounded in real requirements.
4. Prototype: Develop tangible prototypes to explore solution possibilities, varying in complexity based on the project stage.
5. Test: Evaluate prototypes with actual users in real-world contexts, gathering critical feedback.
6. Learn: Reflect on testing outcomes to verify if the solution addresses the problem; adjust or redefine the challenge if needed, leading to new cycles.

This flexible, user-centered approach emphasizes continual refinement and responsiveness to insights throughout the process.

2.3.4 Integration of Design Thinking

This section outlines how the Design Thinking micro-process has been systematically applied throughout our project, enhancing creativity and ensuring a user-centered focus. Each step of the Design Thinking process is aligned with specific sections of the report to foster innovation and responsiveness to user needs.

In Section 1.2, during the “Define the Problem” phase, we articulated clear challenge questions that framed our project goals and directed our efforts toward addressing user expectations. This foundational step set the stage for a thorough exploration of the issues at hand.

In Section 2.1, the “Needfinding and Synthesis” phase involved engaging with stakeholders and conducting research to uncover both obvious and hidden customer needs. This step was crucial in informing our understanding of the context, ensuring that our project was grounded in real user insights.

The “Ideate” phase, which is reflected in our brainstorming sessions documented in Section 3.2, encouraged the generation of diverse ideas based on the synthesized needs. By fostering an environment that valued innovative thinking, we aimed to create solutions that were not only functional but also aligned with user desires.

During the "Prototype" phase, detailed in Section 3, we developed tangible representations of our ideas, allowing us to explore various solutions. These prototypes were subjected to rigorous testing with real users, as highlighted in Section 4, where the "Test" phase provided critical feedback on our concepts.

The iterative nature of Design Thinking allowed us to continuously learn and adapt our approach. Each testing phase generated insights that led us to revisit earlier sections, refining our solutions to ensure they effectively addressed the defined problems. By integrating the Design Thinking methodology with the CRISP-DM framework, we enhanced our project's adaptability and fostered a collaborative environment that prioritized user-centric innovation.

3 Method and analysis

This part outlines the methodologies and analytical approaches employed in our research. It covers data visualization, dataset characterization, data preprocessing procedures, analytical strategies, and machine learning techniques.

3.1 Dataset Description

3.1.1 Features and Attributes

The dataset includes four key features: datetime_utc, volume_demand, volume_production, and spot_price. The objective is to predict future electricity spot prices in the Nord Pool market using historical data on spot prices and volumes.

Datetime UTC

The datetime_utc feature provides the point in time of a single observation in the format YYYY-MM-DD HH:MM:SS+00:00. All datetimes are in UTC and are registered at the beginning of an hour. There are no missing time values. The first datetime is 2015-12-31 23:00:00+00:00, and the last is 2018-09-13 02:00:00+00:00, which adds up to 23,666 rows of data in total. As for the attribute type, although time itself is a continuous value, it is a discrete attribute in this dataset. This is due to the discrete nature of the measurements, where each observation pertains to a specific hour.

Volume Demand

The volume_demand is a floating point number representing the total electricity demand (in megawatts) at a specific time. This feature records the amount of electricity consumers require from the grid and can take any positive real number. As a continuous attribute, volume_demand can vary greatly depending on the time of day, season, and other factors influencing electricity consumption.

Volume Production

Similar to volume_demand, volume_production is a floating point number that indicates the total electricity produced (in megawatts) at a particular timestamp. This feature captures the supply side of the electricity market, detailing how much electricity is generated and fed into the grid. It is a continuous attribute, capable of taking any positive real value, and varies with production capacity, fuel availability, and other production factors.

Spot Price

The spot_price, which is the target variable for prediction, is a floating point number representing the price of electricity (in euros per megawatt-hour) at the specific timestamp. It reflects the

real-time cost of electricity in the market and is influenced by both demand and supply conditions. Being a continuous attribute, spot_price can assume any non-negative real value, indicating the dynamic nature of electricity pricing based on market conditions.

3.1.2 Sources

The dataset used for this project originates from Nord Pool, the leading electricity market in Europe. Nord Pool provides comprehensive data on electricity spot prices and market volumes, which allows us to analyze trends and patterns within the energy market. The data covers multiple years, offering an extensive look at fluctuations in electricity prices and volumes, which is critical for forecasting and model training.

In terms of volume data, Nord Pool records both production and demand volumes on an hourly basis. Each observation represents a sum of electricity demand or production at the start of each hour across various regions participating in the Nord Pool market. These records reflect aggregate data from multiple suppliers and consumers, ensuring a broad view of market dynamics but without individual participant details.

Additionally, Nord Pool's datasets include precise timestamps for each recorded entry, aligning with UTC and timestamped at the beginning of every hour. The data structure helps us develop accurate time-series models by ensuring thath every hour is accounted for without any missing entries.

3.2 Data Analysis Methods and Tools

This part outlines the methodologies and tools applied to examine the Nord Pool dataset, detailing their relevance to our research objectives. Here, we focus on identifying the data characteristics that could best support the predictive aspects of our study. Additionally, insights from this analysis informed our feature selection process, guiding us in choosing the most impactful variables for our objectives. We include descriptive statistics along with visualizations to showcase key findings from our analysis.

Seasonal and Trend decomposition using Loess

Understanding seasonality and trends is crucial in any time series analysis. Seasonality refers to a recurring pattern at a specific interval, while the trend represents a long-term upward or downward movement in the data. These elements are key in predicting future values based on historical data. When a trend is present, it signals a persistent change that is likely to continue, while seasonality allows for the prediction of values based on the time elapsed since previous observations.

To analyze these components, we employ the STL (Seasonal-Trend decomposition using LOESS) method as described by Cleveland et al. (1990). STL breaks down a time series into three parts: seasonality, trend, and residuals. By summing these components, we can reconstruct the original time series. This method is standard in time series analysis due to its robustness and effectiveness. For more details, refer to Cleveland et al. (1990). Figures 3 and 4 illustrate STL plots for the Nord Pool dataset, with Figure 3 covering the entire spot prices history and Figure 4 focusing on the first 300 hours.

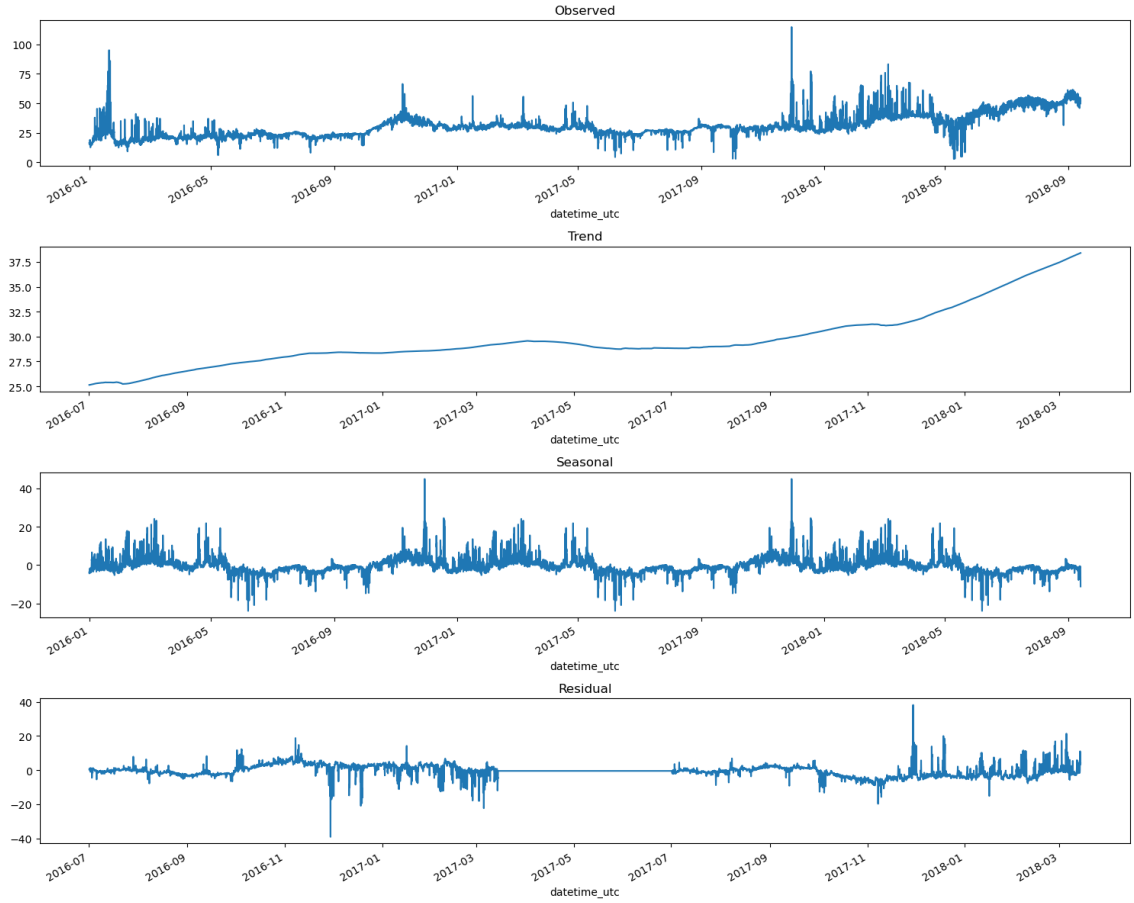

Figure 3: STL decomposition applied to spot price series for all hours

We observe a pronounced trend and seasonality within the dataset. The trend component shows a consistent upward movement throughout the period, indicating a general increase in spot prices over time. Although seasonality is evident, its frequency is too high for a qualitative analysis in figure 3. However, the STL decomposition in figure 4 provides a clearer representation:

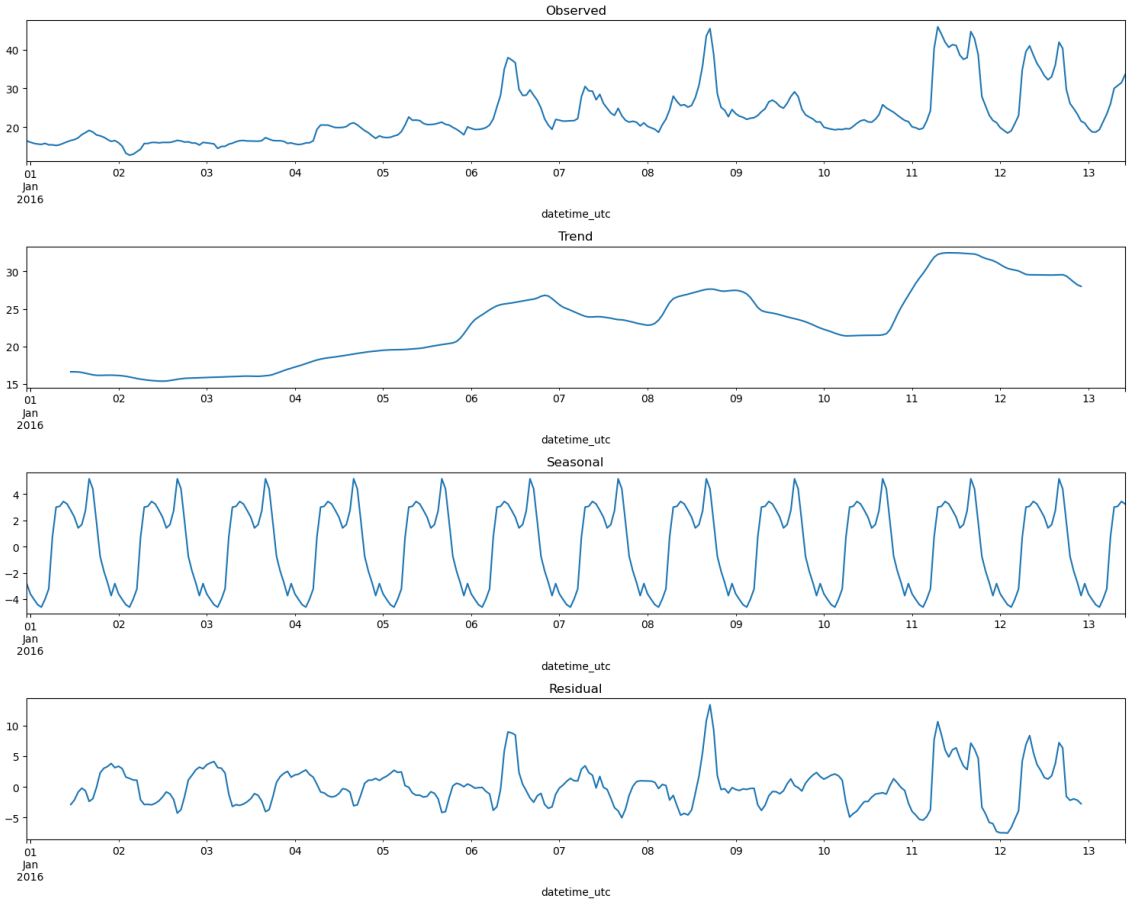

Figure 4: STL decomposition applied to the spot price series for the hours 0 to 300

From Figure 4, we observe a strong presence of intra-daily seasonality in the spot price. This makes intuitive sense, as energy demand varies with human activity. Since less energy is needed during the night, intra-daily seasonality is expected. We also hypothesized that spot prices could vary depending on the day of the week and the month. This hypothesis aligns with our understanding that energy demand—and thus spot prices—fluctuates with daily and monthly patterns in human activity. The analysis of the following figure supports these hypotheses, revealing significant variations across different days and months.

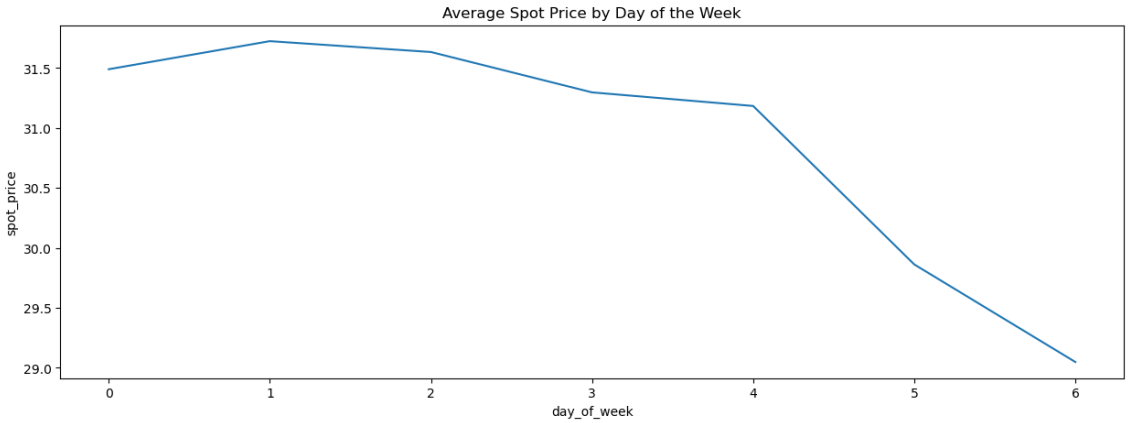

Figure 5: Average spot price by day of the week

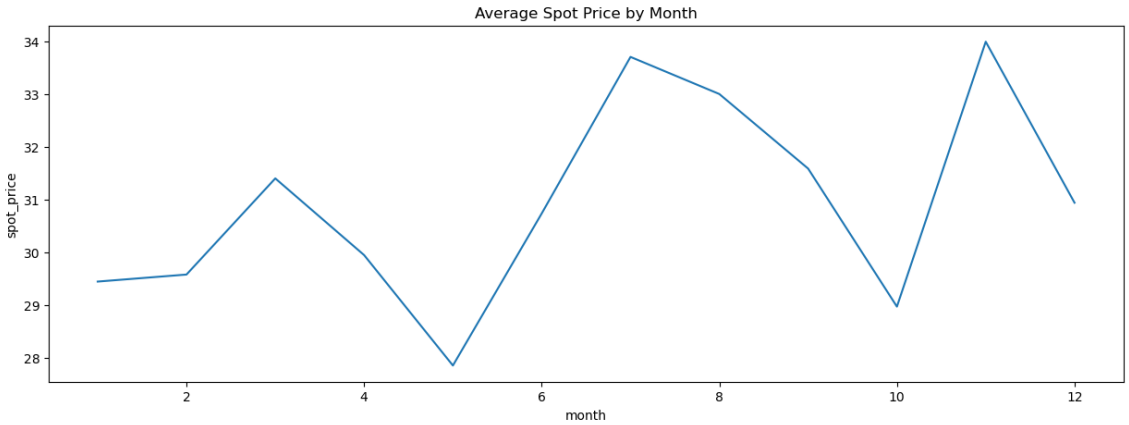

Figure 6: Average spot price by month

At the end of the week, we observe a tendency for spot prices to be lower. This trend can be attributed, in part, to decreased industrial electricity consumption. Monthly patterns may also be linked to temperature fluctuations, which significantly impact the demand for heating or cooling. We hypothesized that energy production and demand would influence its price. By creating correlation plots, we observed a slight positive correlation, indicating that when both consumption and production are higher, prices tend to increase. Additionally, when analyzing the correlation between the differences and the ratio of consumption and production with price, we found a similar slight positive correlation. This further supports our initial hypothesis.

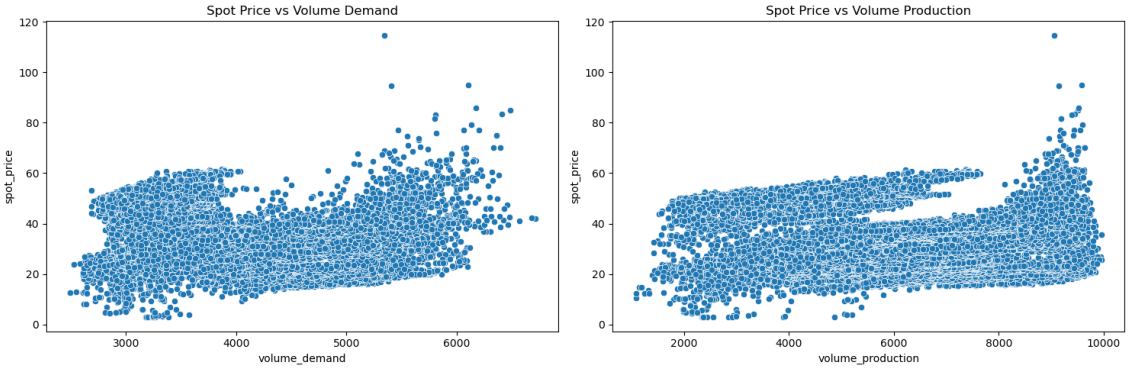

Figure 7: Demand and Production correlation plots over the entire duration of the dataset

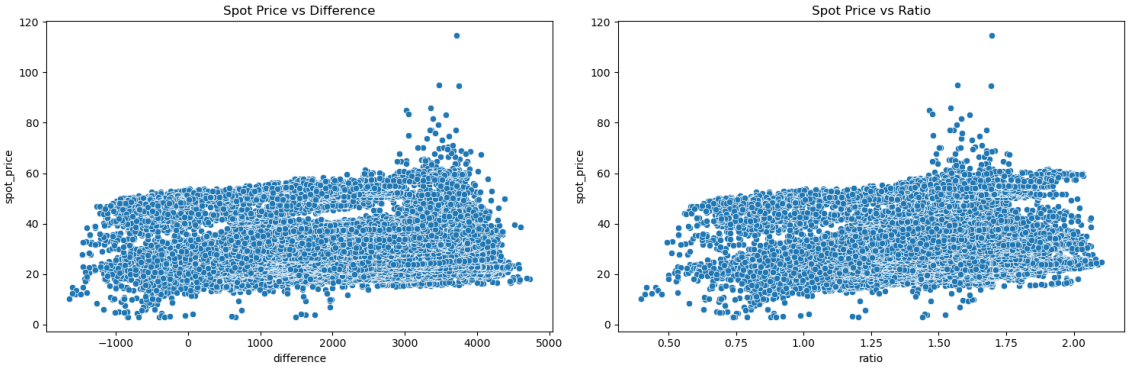

Figure 8: Difference and Ratio correlation plots over the entire duration of the dataset

Autocorrelation

Autocorrelation is a key metric in time series analysis, particularly useful for understanding seasonality and trends. It measures the correlation between a time series and a lagged version of itself over successive time intervals. This helps determine the relationship between past and future values of the series.

If X_t represents the time series at time t, and $\mu_t = E(X_t)$ is the expected value of X_t:

1. **Covariance** between X_t and X_{t+h} is given by:

$$\gamma_X(t+h,t) = \text{Cov}(X_{t+h}, X_t) = E((X_{t+h} - \mu_{t+h})(X_t - \mu_t)) \tag{1}$$

2. **Autocorrelation** for lag h is defined as:

$$\rho_X(h) = \text{Corr}(X_{t+h}, X_t) = \frac{\gamma_X(h)}{\gamma_X(0)} \tag{2}$$

where $\gamma_X(0)$ is the variance of X_t.

Autocorrelation provides insight into how past values of a time series influence future values. For data with a strong trend, values at nearby time points are likely to be similar, resulting in high positive autocorrelation at small lags. If the data exhibits seasonality, autocorrelation will show a repeating pattern at lags corresponding to the seasonal period. For example, if data has a yearly seasonality, autocorrelation will be higher at lags of 12 months, 24 months, etc.

High positive autocorrelation indicates that high values tend to follow high values, and low values follow low values, which is typical for trending data. High autocorrelation at specific lags suggests seasonality, as similar patterns repeat at these intervals.

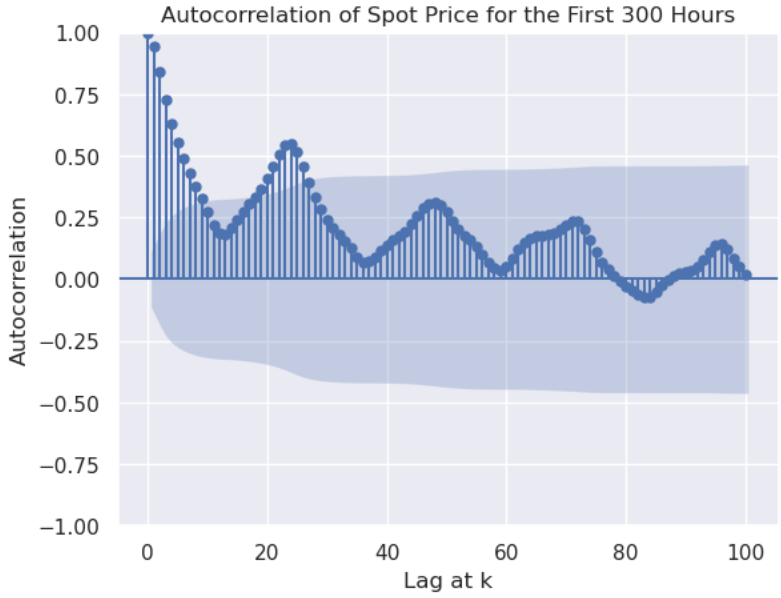

Figure 9: Autocorrelation of the spot price from hour 0 to 300

In Figure 9, it is evident that autocorrelation diminishes for time points further back, with peaks at intervals corresponding to the seasonal frequency. This behavior implies the presence of both trend and seasonality in the observed data, supporting the hypothesis of recurring patterns in the spot price series. Moreover, the analysis reveals a notable correlation between spot price values and their historical counterparts, which suggests that adding lag features could improve the model's capacity to capture these temporal dependencies. This understanding provides a basis for determining an effective range for lag features. The shaded blue area in the plot highlights the statistically significant correlations, indicating that these observed relationships are unlikely due to random noise. Autocorrelation appears to lose significance after roughly 30 time points, offering practical insights for designing our feature engineering strategy, as detailed in the next section.

3.3 Data Preprocessing

Feature Engineering

To enhance the predictive utility of the time variable, we broke it down into five distinct components: day of the week, day of the year, month, hour, and year. This decomposition was intended to aid in capturing seasonal patterns, as prior analysis confirmed the presence of seasonality. By representing time with these granular features, the model is better positioned to identify and leverage patterns that recur on an hourly, daily, weekly, monthly, or annual basis.

In the analysis from the previous phase of the project, we demonstrated the presence of a statistically significant autocorrelation pattern in the data, indicating that adding lag features could be beneficial. However, when we implemented these lag features in the model, we observed that they did not significantly improve performance. For this reason, we ultimately decided not to include them in the final version of the model.

Data Splitting: Input and Output

The final inputs to the model consist of both static features volume_demand, volume_production, temporal features. The output of the model is the predicted spot_price_tomorrow, a continuous variable indicating the anticipated spot price per MWh one day ahead.

For evaluation, the dataset was split into training and testing sets, and models were assessed using Mean Squared Error (MSE) to quantify prediction accuracy. This setup allows us to compare baseline predictions to those enhanced by including lagged features, helping to identify the most effective approach for forecasting the next day's electricity prices.

Preprocessing

The datetime_utc column, which contains timestamps, was converted to a proper datetime format using pandas.to_datetime(). This allowed us to easily extract time-based features like the year, month, week, day, and hour. The spot_price_tomorrow column was created by shifting the spot_price column by 24 hours, aligning the next day's price with the corresponding data from the previous day. No missing values were detected in the dataset, ensuring that no rows were dropped during preprocessing.

Data Cleaning

During the data cleaning process, we first checked for missing values in the dataset and confirmed that there were no missing entries across any of the columns, including volume_demand, volume_production, and spot_price. This is essential to ensure the reliability of our analysis and model predictions.

Additionally, we dropped the datetime_utc and date columns after extracting relevant features from the datetime information. These columns were no longer necessary for our analysis, streamlining the dataset and reducing its complexity. By focusing only on the essential features, we prepared a clean dataset ready for further analysis and modeling.

3.4 Machine Learning Methods

Model Choice and Motivation

We used Random Forest regression as our forecasting model. The reasons are that Random Forest is considered a robust and well-performing method and is quick to implement with the convenient scikit-learn library. We initially experimented with a simple baseline model that creates predictions based on the direct past. A comparison between the methods, as well as an empirical justification for our selection of Random Forest, is provided in the results section.

Baseline: Naive Forecasting

A naive forecast approach for predicting the next day's spot price would simply replicate the most

recent day's data available. In our context, the baseline model predicts based on the spot price data from the previous day, embodying the most straightforward prediction technique. Concretely, if $\hat{y}_i$ is the prediction and y_i is the spot price at time i:

$$\hat{y}_i = y_{i-24}$$

The model is evaluated in the results section.

Random Forest

Random Forest is an ensemble model that consists of many decision trees. Each tree is trained on a random subset of the data, and the final prediction is made by averaging the predictions from all individual trees. For brevity, we will include a condensed mathematical description of Random Forest. The model prediction is calculated by averaging the predictions from T trees:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(x)$$

Where $\hat{y}$ is the final prediction and $f_t(x)$ represents the prediction from the t-th tree.

The objective of the Random Forest algorithm is to minimize the mean squared error, given by:

$$L(\phi) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Where N is the number of observations, y_i is the actual value, and $\hat{y}_i$ is the predicted value.

In summary, many weak learners (decision trees) are combined to create a strong learner, and the model is trained to be generalizable. Consequently, the model is well suited for the problem statement task, as it captures complex patterns while maintaining generalizability across different time intervals.

3.5 Model Evaluation and Metrics

4 Evaluation and interpretation

4.1 Evaluation of analysis and visualizations

In our analysis of the Nord Pool electricity market data, we identified critical patterns and relationships that inform our forecasting model. Below, we outline the key analyses and corresponding visualizations that provide insights into the dynamics of the market.

4.1.1 Variable Distributions

We examined the distribution of each variable-demand, production, and spot price, to understand the spread, potential skewness, and outliers within the dataset. This analysis helps to reveal baseline patterns and any unusual behavior in the data. Box plots for each variable display the distributions, allowing for a clear interpretation of the typical range and any significant deviations that may impact forecasting.

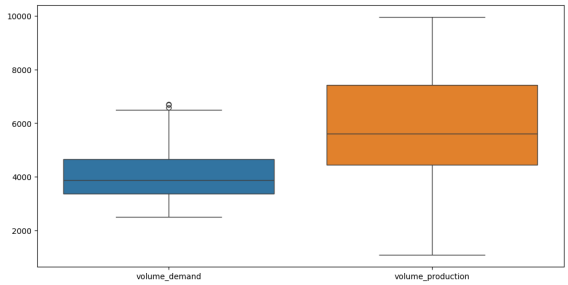

Figure 10: Volume demand and volume production box-plot

For the volume demand, the box plot shows a central range with a few outliers above the upper whisker, indicating some unusually high values in demand. The median line is near the middle of the box, suggesting a relatively symmetric distribution of demand values.

In orange, the production volume data has a broader range compared to demand, with a higher median (the line inside the box). The whiskers extend widely, indicating more variability in production compared to demand.

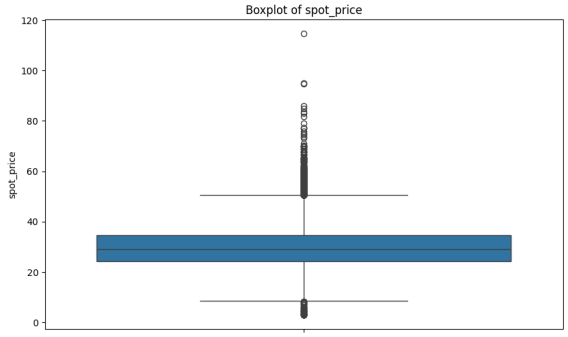

Figure 11: Spot price box-plot

The box plot illustrates the distribution of spot prices, highlighting a concentration of values within the IQR and the presence of significant outliers, which suggest variability and potential anomalies in the data. The y-axis, labeled spot price, ranges from 0 to 120. The blue box represents the inter-quartile range (IQR), spanning approximately from 20 to 40, with a median spot price around 30. The whiskers extend from about 0 to 60, indicating the range of the main data distribution. There are several outliers above 60, with the highest around 110, as well as a few outliers below 0.

4.1.2 Correlation analysis between variables

To explore the relationships between demand, production and spot prices, we conducted a correlation analysis. The heatmap provides a quick way to identify both strong and weak relationships between these variables.

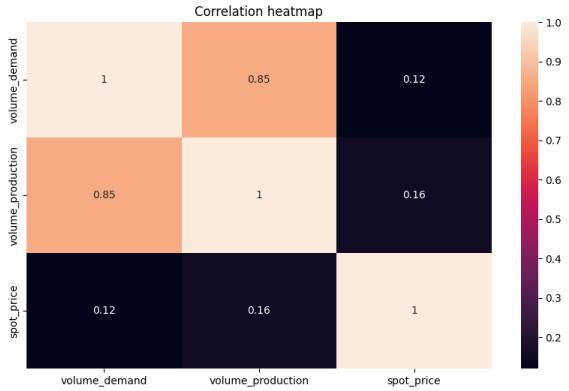

Figure 12: Correlation heat map

This correlation heatmap visualizes the relationships between three variables: volume demand, volume production, and spot price. It uses color intensity to represent correlation values, ranging from -1 to 1, with a color bar on the right as a reference. The correlation between volume demand and volume production is 0.85, indicating a high positive correlation, which suggests that as demand volume increases, production volume also tends to increase, or vice versa. The correlation between volume demand and spot price is 0.12, showing a very weak positive relationship between demand volume and spot price. For volume production and spot price, the correlation is 0.16, also weak, implying that production volume has little impact on the spot price.

4.2 Business value interpretation

The observed seasonal and trend patterns, correlations and temporal dependencies contribute directly to understanding and predicting spot price fluctuations. An accurate forecast would enable Nord Pool to minimize the spot price minimizing costs by adjusting their activity based on expected price fluctuations.

5 Deployment and recommendations

6 Monitoring and maintenance

Bibliography

[BU16] Walter Brenner and Falk Uebernickel, eds. *Design Thinking for Innovation: Research and Practice*. Library of Congress Control Number: 2016933122. Cham, Switzerland: Springer, 2016. ISBN: 978-3-319-26098-3. DOI: 10.1007/978-3-319-26100-3. URL: https://doi.org/10.1007/978-3-319-26100-3.

[She00] C. Shearer. 'The CRISP-DM model: the new blueprint for data mining'. In: *Journal of Data Warehousing* 5.4 (2000), pp. 13–22.

[TAH20] Youssef Tounsi, Houda Anoun and Larbi Hassouni. *CSMAS: Improving Multi-Agent Credit Scoring System by Integrating Big Data and the new generation of Gradient Boosting Algorithms*. Mar. 2020. DOI: 10.1145/3386723.3387851.