**Abstract**

The growing reliance on renewable energy has underscored the need for accurate forecasting and monitoring of variable sources like wind power. This project focuses on Germany’s wind energy market and aims to develop machine learning models that can forecast next-day wind generation, detect shortfalls before they occur, and identify anomalies that may signal operational issues. Using high-resolution data collected at 15-minute intervals from 2015 to 2018, we applied a structured CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology to guide the analysis.

Feature engineering was a key component, leveraging lag values, rolling averages, and time-based attributes to capture temporal patterns and production behavior. A Random Forest Regressor was employed for generation forecasting, achieving strong predictive performance with low error rates. Shortfall prediction was addressed through classification models, where the Random Forest Classifier demonstrated high precision and solid recall. Unsupervised learning with DBSCAN successfully isolated rare anomalies without excessive false positives.

The results highlight the potential of machine learning to enhance the reliability, planning, and operational management of renewable energy systems. By improving forecast accuracy and early event detection, the models developed here can contribute to more stable grid operations and better support the global transition toward sustainable energy.

**Project Methodology**

This project follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, a widely adopted framework for organizing data science and machine learning projects. CRISP-DM provides a structured approach, ensuring that each phase of the project from business understanding to model evaluation is methodically addressed and aligned with the project objectives.

The process began with a clear business understanding, identifying the key goals of forecasting next-day wind energy generation, predicting potential shortfalls, and detecting anomalies in Germany’s renewable energy market. During the data understanding phase, the dataset was thoroughly explored to assess its structure, quality, and statistical properties, focusing on patterns relevant to wind generation. Data preparation involved handling missing values using forward and backward fill techniques, and engineering critical features such as lag variables, rolling averages, and time-based attributes to enrich the dataset for modeling. A binary label was also created to classify shortfall events.

In the modeling phase, machine learning models were developed for each business question. A Random Forest Regressor for wind generation forecasting, Random Forest and Logistic Regression classifiers for shortfall detection, and the DBSCAN clustering algorithm for unsupervised anomaly detection. Each model was selected based on its ability to handle the specific challenges of the data, such as non-linearity, temporal dependency, and class imbalance. The evaluation phase involved assessing the models using appropriate metrics MAE and RMSE for regression, Precision, Recall, F1-Score for classification, and clustering visualization for anomaly detection along with visual inspections to interpret the results. Although model deployment was not carried out in this study, recommendations for practical integration and future enhancements, such as incorporating real-time weather data and periodic retraining, were discussed to further strengthen the system's operational value.

**Problem Statement**

Over the past several years, global energy consumption has increased rapidly across both industrial and residential sectors. This surge in demand has prompted a major shift toward exploring and investing in renewable energy sources, particularly wind and solar. Unlike fossil fuels, however, renewable energy production is heavily dependent on environmental conditions, making it much more variable and difficult to predict. This variability creates a critical need for reliable forecasting, early shortfall detection, and anomaly monitoring to maintain grid stability and optimize energy distribution systems. Accurate forecasting also supports better investment decisions in renewable infrastructure by reducing operational risks.

This project focuses specifically on the German wind energy market, using high-quality 15-minute interval data collected from the Open Power System Data (OPSD) platform. The dataset spans from 2015 to 2018 and provides information on electricity prices, wind and solar generation actuals, load demand values, and various time-based attributes. While the data includes several European countries — Austria, Belgium, Germany, Hungary, Luxembourg, and the Netherlands — the primary emphasis of this project is on Germany, which offers a rich and dynamic environment for wind energy analysis.

The analysis is driven by three core business questions:

1. Can we accurately predict next-day wind energy generation in Germany using historical and time-based features?
2. Can we forecast when wind energy output falls below expected levels to issue early warnings?
3. Can we detect unusual energy behavior that may signal system faults or irregular usage patterns?

The main data mining goals supporting these questions are:

* Applying regression methods to predict next-day wind generation in Germany.
* Using classification techniques to detect potential shortfalls before they occur.
* Employing unsupervised anomaly detection methods to uncover rare and irregular patterns in energy behavior.

By addressing these questions, the project aims to assist energy planners and system operators in improving grid reliability, reducing operational risks, and enhancing the efficient use of renewable energy resources. The project follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, ensuring a structured, clear, and outcome-driven approach from business understanding through evaluation.

**Data Understanding**

Following the CRISP-DM methodology, the initial phase involved developing a comprehensive understanding of the dataset and preparing it for modeling.

The dataset contains 105,815 time-stamped records at 15-minute intervals, covering the period from 2015 to 2018. To explore the dataset, basic descriptive statistics were generated using .info() and .describe(), providing insights into column structures, data types, and key statistical measures such as mean, standard deviation, minimum, maximum, and quartile values. The presence of zero values was identified using the (data == 0).sum() method, an important step particularly for wind and solar generation columns where zeros often represent natural conditions like nighttime periods with no solar output. Missing values (NaNs) were detected using .isnull().sum(), and the dataset's datetime index was properly set and verified for generating time-based features.

**Data Preparation and Feature Engineering**

The preparation of the dataset focused exclusively on Germany’s wind energy data, aiming to build a strong foundation for accurate modeling. By combining time series analysis techniques with a deep understanding of renewable energy behavior, a comprehensive set of features was developed to capture both temporal patterns and operational dynamics.

To address missing values without disrupting the natural flow of the data, forward fill (ffill()) and backward fill (bfill()) methods were applied. This approach helped maintain the continuity of the time series, ensuring that models would not be misled by artificial breaks or sudden gaps.

Feature engineering played a crucial role in strengthening model performance. To help the models understand temporal dependencies, lag features were created by shifting the wind generation values by 15 minutes (wind\_15min\_ago) and by 24 hours (wind\_24hr\_ago). These features enabled the models to learn how recent and previous-day production levels influence future outputs. In addition to lagged information, a rolling average feature, expected\_wind, was introduced by calculating the 24-hour moving mean of wind generation. This smoothed out short-term fluctuations and emphasized broader production trends that are essential for both forecasting and shortfall detection.

Recognizing the importance of time-of-day and seasonal effects on wind generation, additional time-based features were extracted from the timestamp, including hour, day\_of\_week, and month. These attributes allowed the models to capture repeating daily and seasonal patterns inherent in wind energy behavior.

To support the shortfall prediction task, a binary classification label named Germany\_shortfall was created. This label flagged periods when actual wind generation fell below 80% of the expected rolling average, signaling significant underperformance relative to recent trends.

After all feature engineering steps were completed, any rows containing missing values introduced by shifting and rolling operations were systematically removed using .dropna(). This final cleaning step ensured that the dataset used for training and evaluation was clean, consistent, and reliable.

**Data Visualizations**:

1. Line Plot: Wind Energy Generation Over Time (Germany vs Netherlands)

A graph with blue and orange lines

AI-generated content may be incorrect.

The line graph provides a wind energy generation comparison between Germany and Netherlands over a three-year period. The x-axis represents utc\_timestamp and y-axis represents wind energy generation in megawatts (MW). The blue line represents Germany wind generation and orange represents Netherlands wind generation.

The graph provides clear picture that Germany provides high wind energy production crossing more than 30000 megawatts MW. The Netherland shows very low and stable wind generation with less than 3000 megawatts MW. This comparison reveals regional difference in wind generation, making Germany is a better choice for modeling shortage and anomaly identification.

1. Line Plot: Wind Renewable Energy Generation Over Time (Germany)

A blue line graph with black text

AI-generated content may be incorrect.

The line graph provides information about the wind energy generation in Germany from January 2015 to January 2018, based on 15 minutes of interval. The x axis represents the utc\_timestamp and y axis represents the German\_wind\_generation\_actual, measured in megawatts (MW).

The line graph reveals key insights:

There is a clear increase in spikes at the end of the month every year. Suggest seasonal wind patterns. At the same time, the production of wind power generation is unpredictable, with short term increase at all times. According to the plot, wind generation capacity has generally increased over time, whether as a result of improved weather or the growth of infrastructure.

This above line chart provides a clear view of time-based behavior of the wind data, supporting the decision to include **lag features and time attributes** in predictive modeling.

1. Line Plot: Solar Energy Generation Over Time (Austria vs Netherlands)

A graph with numbers and lines

AI-generated content may be incorrect.

This line graph visualizes the solar energy generation in Austria and Netherlands over a three-year period. The x-axis represents utc\_timestamp and y-axis represents solar energy generation in megawatts (MW). The blue line represents Austria solar generation and orange represents Netherlands solar generation.

Netherlands constantly produces higher solar generation output compared to Austria throughout the timeline, suggesting better infrastructure or larger solar panel coverage. Also, we can observe the solar production peaks in summer and decrease in winter. These trends give a clear view that month-based features are required for predictive modelling and validate the inclusion of time-based lag variables.

1. Histogram – Distribution of Wind Generation (Germany)

A graph of a distribution of wind generation

AI-generated content may be incorrect.

This histogram visualizes Distribution of wind generation in Germany throughout the full data collection from 2015 to 2018. The X-axis describes the wind generation in megawatts MW, at interval of 1000 MW. The Y-axis represents frequency, which is 15-minute intervals for each wind generation.

The distribution is right-skewed, with most values ranging between 2,000 to 10,000 MW.

1. Line Plot – Germany Load vs Renewable Energy Generation

A screenshot of a graph

AI-generated content may be incorrect.

This multi-line chart compares Germany total energy demand with actual solar and wind renewable energy generation from 2015 – 2018. The X-axis represents utc\_timestamp and y-axis represents power generation in megawatts (MW). The blue line Germany\_load\_actual\_entsoe\_transparency represents Germany energy demand. The orange line Germany\_wind\_generation\_actual represents actual wind energy genaration. The green line Germany\_solar\_generation\_actual represents actual solar energy genaration in Germany.

The energy demand in Germany remains stable over time with minor seasonal drops or rises. Compared to solar, wind energy generation is significantly higher throughout the years, with fewer drops due to climatic changes. Solar energy performs good in summer and lower in winter highlighting a predictable trend. By this visualization we can observe renewable energy wind and solar falls below the expected load, supporting our shortfall classification predictions.

**Data Modelling**

In this project, a structured modeling approach was followed to address the core business objectives: predicting next-day wind energy generation, detecting shortfalls, and identifying anomalies in the German renewable energy market. Statistical and machine learning methods were carefully selected, applied correctly, and supported by a clear rationale rooted in the nature of the data and business goals.

A targeted modeling strategy was adopted to meet the project's forecasting, classification, and anomaly detection goals using Germany’s wind energy data. Methods were selected based on the nature of each task, and all models were built using a consistent set of engineered features. The three primary modeling approaches are summarized below:

1. Regression Model: Predicting Next-Day Wind Generation
2. Classification Model: Predicting Wind Energy Shortfalls
3. Unsupervised Learning: Anomaly Detection

**Regression Model: Predicting Next-Day Wind Generation**

To forecast next-day wind energy generation in Germany, a Random Forest Regressor was implemented. This ensemble method was selected due to its robustness against noise, its ability to model complex nonlinear relationships, and its effectiveness in handling outliers and skewed distributions — characteristics inherent in renewable energy data.

The modeling process began by engineering critical features, including lag variables (wind\_15min\_ago and wind\_24hr\_ago) to capture temporal dependencies, and a rolling average feature (expected\_wind) to model local production trends. Time-based features such as hour, day\_of\_week, and month were also extracted from the timestamp to reflect daily, weekly, and seasonal patterns in wind behavior. After feature construction, rows containing missing values resulting from lagging and rolling operations were removed to ensure data integrity.

The processed dataset was then used to train the Random Forest model, leveraging the engineered features to capture historical trends and time-based variations. The non-linear nature of the wind generation data further validated the choice of a tree-based ensemble over simpler linear approaches.

**Classification Model: Predicting Wind Energy Shortfalls**

To predict shortfall events in Germany’s wind energy production, both Logistic Regression and Random Forest Classifier models were developed. Logistic Regression was initially applied as a baseline due to its simplicity and interpretability, while the Random Forest Classifier was chosen to better capture non-linear relationships and complex interactions among features.

The modeling workflow involved using the engineered features, including lag variables, rolling averages, and time-based attributes (hour, day\_of\_week, and month). A binary target variable (Germany\_shortfall) was created by flagging instances where the actual wind generation fell below 80% of the expected rolling average. Special attention was given to addressing multicollinearity among predictors and verifying class balance to ensure that both shortfall and non-shortfall cases were adequately represented during model training. The prepared dataset was then used to train and benchmark both models against the shortfall prediction task.

**Unsupervised Learning: Anomaly Detection**

To identify irregular patterns in Germany’s wind energy generation data, the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm was applied. DBSCAN was selected for its capability to detect clusters of varying shapes and to isolate noise points without requiring prior labeling of anomalies.

Prior to modeling, key features such as wind generation output, load values, and time-based indicators were selected. The dataset was scaled appropriately to ensure that distance-based clustering was effective. Model hyperparameters, including the epsilon neighborhood radius and the minimum number of samples per cluster, were optimized using k-distance plots to achieve meaningful separation between normal operational points and outliers. The trained DBSCAN model allowed the identification of rare or abnormal production patterns critical for monitoring the reliability of renewable energy systems.

**Model Evaluation and Results**

Once the models were developed and trained, we carefully evaluated their performance using suitable metrics tailored specifically to each task—forecasting wind generation, predicting shortfalls, and identifying anomalies. The evaluations provided clear insights into how effectively these models could perform in real-world scenarios, particularly in supporting energy management decisions for Germany’s renewable energy sector.

**Regression Model Evaluation and Results**

We evaluated the Random Forest Regressor’s performance using two key metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). MAE represents the average error of the predictions, while RMSE gives additional weight to larger prediction errors. Our regression model achieved an MAE of 1.64 and an RMSE of 2.58, reflecting robust predictive accuracy. Given the inherent variability of wind generation due to fluctuating weather conditions, these results are quite strong.

This indicates that our model successfully learned from important patterns in the data, benefiting significantly from the lagged and rolling features. Though no model can predict wind energy perfectly, the low error values confirm that our model is reliable enough for practical forecasting tasks. Additionally, visual comparisons (Figure 1) demonstrate that predicted values align closely with actual wind generation data, with deviations mainly occurring during short periods of sudden change—expected behaviors in renewable energy production.

Regression – Actual vs. Predicted Wind Generation:

A line plot comparing actual versus predicted values further confirms the model’s performance.

A graph with blue and orange lines

AI-generated content may be incorrect.

**Explanation:** Figure 1 shows a comparison between the actual wind generation values and the next-day predictions produced by the Random Forest Regressor. The two lines closely track each other across time, demonstrating that the model accurately captures both overall trends and short-term fluctuations in energy production. Minor deviations are observed during periods of rapid change, which is expected given the inherent variability of wind resources. Overall, the plot confirms that the model provides reliable forecasts that could support daily energy planning operations.

**Classification Model Evaluation and Results**

For shortfall prediction, we evaluated our models using Accuracy, Precision, Recall, and F1-Score, metrics that together help measure the model’s capability in distinguishing shortfall events from normal operation. Between Logistic Regression and the Random Forest Classifier, the Random Forest model showed better results and thus was chosen for detailed evaluation.

Our classifier achieved an overall accuracy of 86**%**, precision of 92**%**, recall of 64**%**, and an F1-score of 0.75. Although high precision means that a few non-shortfall instances were incorrectly flagged, the recall rate of 64% indicates that the model missed some actual shortfall cases.

The confusion matrix (Figure 2) provides further insight into this performance, clearly showing that while the classifier accurately identifies most normal conditions. Nonetheless, the current model provides significant operational value in managing grid reliability.

Classification – Confusion Matrix for Shortfall Prediction:

A chart of a number of colored squares

AI-generated content may be incorrect.

**Explanation:** Figure 2 displays the confusion matrix for the Random Forest Classifier used to predict wind energy shortfalls. The matrix shows that the model correctly identified most instances where no shortfall occurred and also captured a substantial number of true shortfall events. However, some shortfalls were missed, which is reflected in the recall score discussed earlier. The visualization highlights the model’s strength in minimizing false alarms while identifying the majority of operational risks, making it valuable for early warning systems.

**Anomaly Detection Evaluation and Results**

We applied DBSCAN, an unsupervised clustering method, to detect unusual patterns in the wind generation data. Since DBSCAN doesn’t require labeled anomalies, we evaluated its effectiveness by examining the number and nature of detected outliers.

After fine-tuning its parameters, DBSCAN flagged 16 anomalies out of over 105,000 datapoints. A very low anomaly rate that suggests the method is precise and not overly sensitive. The detected anomalies typically corresponded to sudden and extreme variations in wind production and load values. Because these types of events, although rare, can significantly impact operational stability, DBSCAN’s ability to identify them is particularly valuable.

The scatter plot of DBSCAN clusters (Figure 3) visually demonstrates clear separation between densely clustered normal operational points and sparsely located outliers. These results confirm DBSCAN as a reliable approach for monitoring unusual energy patterns, potentially alerting operators to equipment issues or unexpected environmental conditions.

DBSCAN Cluster Visualization for Anomaly Detection:

A visual scatter plot of the DBSCAN clusters clearly shows the separation between dense, stable regions and the sparsely located outliers.

A diagram of a wind

AI-generated content may be incorrect.

**Explanation:** Figure 3 presents the clustering results from the DBSCAN algorithm applied to wind energy generation data. Most data points are grouped into well-defined clusters, representing normal energy production behavior. A small number of points are marked as outliers (cluster label -1), indicating rare or unusual patterns in the data. These anomalies often correspond to sudden drops or spikes in wind generation that could signal system faults, forecasting errors, or unexpected weather events. The figure confirms that DBSCAN effectively isolates critical irregularities from normal operational patterns.

**Conclusion**

This project highlighted the potential of machine learning models to support forecasting, monitoring, and decision-making in Germany’s wind energy sector. By combining time series analysis with domain-driven feature engineering, we developed solutions that tackled three important challenges: predicting next-day wind generation, detecting shortfalls before they impact operations, and identifying unusual patterns in energy behavior.

The Random Forest Regressor delivered strong predictive accuracy, reliably estimating next-day wind generation while accounting for natural variability in renewable production. The Random Forest Classifier showed high precision in identifying shortfall events, though improving recall remains a key area for future enhancement to ensure more complete detection. The DBSCAN clustering approach successfully flagged rare anomalies, providing early warnings for potential operational issues without generating excessive false alarms.

Together, these models create a solid foundation for smarter energy management. They offer tools that can help grid operators anticipate supply fluctuations, respond to emerging risks, and plan resources more effectively. As renewable energy systems continue to grow in complexity, integrating predictive modeling like this will be essential for ensuring reliability, efficiency, and sustainability. Future improvements could focus on refining models with real-time environmental inputs, expanding feature sets, and adapting to changing energy dynamics through regular retraining.