IE Sustainability Datathon October 2024 - EDP Challenge

Saviors of the World : Roberto Cedano Delan, Yupeng Chen, Joseph Clerc, Sofia Depoortere, Alessio Mastropietro, Sravan Sridhar

Table of contents

1. Introduction

This project focuses on developing predictive models to enhance the operation of Battery Energy Storage Systems (BESS). BESS plays a pivotal role in bridging the gap between renewable energy generation and consumption. By storing excess energy during periods of low demand or high generation and discharging it when demand spikes, BESS contributes to grid stability and energy reliability. Moreover, their ability to engage in energy trading—buying energy at low prices and selling at peak times—creates additional economic value.

The objective of this project is to forecast two critical metrics: electricity prices and system imbalances. Accurate predictions of these metrics enable BESS to optimize their charging and discharging cycles, enhancing their efficiency and profitability. The analysis incorporates extensive datasets, including historical energy prices, generation metrics, demand loads, and weather conditions. These data sources are processed and integrated using advanced machine learning techniques to ensure precise and actionable predictions.

Through this work, we aim to demonstrate how data-driven solutions can address the complexities of the renewable energy transition. By optimizing the operation of BESS, this project contributes to building a more resilient and sustainable energy infrastructure.

The datasets provided for the IE Sustainability Datathon are as follows:

1. Price Data
2. Generation Data
3. Demand Load Data
4. Balancing Data

Each dataset includes half-hourly data spanning from January 1, 2018, to September 30, 2024, providing 48 records per day.

2. Exploratory Data Analysis

To better understand the datasets and prepare them for modeling, we conducted an extensive Exploratory Data Analysis (EDA). This process aimed to uncover anomalies in the data while ensuring it was clean and ready for predictive modeling.

The Balancing Data revealed critical insights into energy market dynamics. Variables such as System Price, Bid Acceptances, and Offer Acceptances showed notable volatility, reflecting the fluctuations in energy demand and supply. Seasonal and temporal trends were identified in the Net Imbalance Volume (NIV), which plays a pivotal role in balancing energy markets.

The Demand Load Data provided a detailed view of electricity consumption patterns. Metrics like Loss of Load Probability and Actual Total Load highlighted peak demand periods, often corresponding to specific times of the day. These insights are crucial for optimizing energy storage and discharge operations, particularly for Battery Energy Storage Systems (BESS).

The Generation Data emphasized the dependency of renewable energy outputs—such as solar and wind power—on weather conditions. This variability underscored the need to incorporate weather data into predictive models to improve their accuracy. Seasonal trends

in renewable energy generation also offered valuable context for understanding grid dynamics.

The Price Data highlighted significant price fluctuations in both day-ahead and intraday markets. The observed price spreads revealed opportunities for energy arbitrage, where BESS can store energy when prices are low and sell it during peak price periods, maximizing profitability.

Visualizations such as heatmaps, time-series plots, and boxplots were employed to deepen the analysis. Heatmaps helped identify correlations between variables, while time-series plots illuminated temporal trends and seasonality. Boxplots and distribution charts revealed the presence of outliers and provided an understanding of the spread of key metrics, such as system price and energy generation.

We identified several challenges. High variability in intraday prices and renewable energy outputs presented a challenge for model robustness. Additionally, missing data in certain time periods required imputation to avoid introducing bias into the model predictions. Addressing these challenges ensures the development of accurate and reliable predictive models.

3. Preprocessing & Feature Engineering

Data cleaning was an essential first step. Numerical columns with non-standard characters were standardized and converted to proper numeric formats. This ensured consistency across all datasets. Regex (Regular Expressions) played a crucial role in this process, particularly in normalizing column names. This technique was used to replace non-alphanumeric characters with underscores, convert sequences of whitespace into single underscores, and collapse multiple underscores into one. These operations were vital for standardizing dataset headers, making them compatible with downstream machine learning workflows. Missing values were hand in led using a combination of imputation and record removal, depending on their significance and potential impact on the analysis. Additionally, all time-related variables were converted into a uniform datetime format to facilitate accurate temporal analyses.

Temporal features were created to leverage the half-hourly granularity of the data. Variables such as the Hour and Day of the Week were extracted from timestamps to highlight intraday and weekly patterns in energy demand and prices. Seasonal indicators, such as summer and winter flags, were also introduced to capture the influence of weather conditions on renewable energy generation and consumption trends.

To smooth out noise and reveal underlying trends, aggregated metrics such as rolling averages were computed for key variables like system price and energy demand. Lagged features, which reflect past values of variables, were incorporated to model dependencies over time, particularly for price and load forecasting. Additionally, price spreads were calculated as the difference between day-ahead and intraday prices, providing insights into potential arbitrage opportunities for Battery Energy Storage Systems (BESS).

To address nonlinear relationships and enhance model performance, interaction features were created by combining variables such as load and price, reflecting their combined impact on energy market behaviors. Transformations, including logarithmic and polynomial adjustments, were applied to variables with skewed distributions or nonlinear patterns.

The features engineered during this stage provided the foundation for robust predictive modeling. By leveraging temporal patterns, aggregated metrics, and cross-dataset insights,

the models were equipped with the necessary inputs to produce precise and actionable forecasts.

4. Model Selection & Training

The model training process was designed to develop robust predictive models capable of accurately forecasting key energy market metrics. This section outlines the steps taken to prepare, train, and evaluate multiple models, leveraging advanced machine learning techniques to optimize performance.

Data Preparation for Training

The dataset was preprocessed to ensure consistency and compatibility with machine learning algorithms. Features were scaled using StandardScaler to normalize numerical variables, enhancing the models' ability to converge during training. The dataset was then split into training and testing subsets using an 80:20 ratio to evaluate model generalizability on unseen data.

Algorithm Selection

Several regression algorithms were explored to identify the best-performing model. These included:

- Linear Regression, Ridge, and Lasso for baseline comparisons.
- Random Forest Regressor and Gradient Boosting Regressor for capturing nonlinear relationships.
- LightGBM and AdaBoost Regressor for efficient handling of large datasets and fine-grained performance optimization.
- Stacking Regressor and Voting Regressor for combining the strengths of multiple models to improve overall accuracy.

Hyperparameter Tuning

Hyperparameter tuning was performed using GridSearchCV and RandomizedSearchCV. These methods iteratively evaluated combinations of parameters to identify configurations that minimized the root mean squared error (RMSE). Key hyperparameters, such as learning rates, tree depths, and regularization terms, were optimized to enhance model performance.

Model Evaluation

Models were evaluated using metrics such as Root Mean Squared Error (RMSE), ensuring alignment with the datathon’s objectives. Cross-validation with K-Fold (K=5) was employed to mitigate overfitting and provide reliable performance estimates. The LightGBM Regressor demonstrated superior accuracy and computational efficiency, making it the final model choice.

Challenges Addressed

Model training addressed challenges such as overfitting and high variability in the data. We applied techniques like feature selection, regularization, and early stopping to ensure the models were both accurate and generalizable. Computational efficiency was also a priority as well as a challenge, given the large size of the dataset and the complexity of the models. Some of our models actually never finished running.

5. Model testing

The predictive models underwent rigorous evaluation to ensure accuracy and robustness. The testing framework relied on temporal cross-validation, which accounted for the chronological dependencies within the data. This method helped assess how well the models performed when applied to future, unseen periods, a critical aspect for ensuring real-world applicability. Root Mean Squared Error (RMSE) was evaluated as a performance metric to evaluate the models. Additionally, residual analysis was conducted, which revealed areas where the models underperformed, especially during periods of high price volatility or extreme weather conditions.

The Stacking Regressor consistently outperformed other algorithms in both predictive accuracy and computational efficiency. This model was chosen as the final implementation due to its superior performance. Its capabilities were further enhanced through hyperparameter optimization using GridSearchCV, which fine-tuned critical aspects such as learning rate and tree depth to minimize errors.

6. Conclusions

This project successfully demonstrated the feasibility of leveraging machine learning to optimize the operations of BESS. The results confirmed that it is possible to produce highly accurate forecasts of electricity prices and system imbalances, enabling BESS to capitalize on opportunities for profitable arbitrage while simultaneously enhancing grid stability. The inclusion of temporal and seasonal trends in the modeling process provided insights into the impact of seasonal patterns on energy market dynamics.

7. Further Recommendations

While the current project delivers robust forecasts, there are several enhancements that could significantly improve the system's scalability and user interaction in the future. These enhancements were not implemented due to time constraints and limited access to data but offer substantial potential for refining the solution.

One promising avenue is the deployment of the forecasting model on a cloud-based infrastructure. A cloud pipeline would enable continuous time series forecasting, allowing the model to adapt in real time to evolving market conditions. This approach would automate data ingestion, preprocessing, and prediction generation, ensuring the system operates with minimal manual intervention. Additionally, it would enhance scalability, handling larger data volumes seamlessly, while ensuring the model remains accurate by consistently training on the latest data. Such a deployment would provide stakeholders with always-updated insights, aligning predictions closely with real-world dynamics.

Another critical enhancement would involve incorporating real-time weather data into the forecasting process. Since weather fluctuations significantly impact renewable energy generation, integrating real-time weather updates could improve the model's accuracy and responsiveness. This capability would allow the system to adapt dynamically to sudden changes in weather conditions, such as unexpected storms or sunny periods, enhancing the reliability of predictions for energy generation and imbalances. Incorporating this data would

improve decision-making for storage and discharge operations, ensuring a more stable and efficient grid.

Finally, an innovative recommendation is the development of a large language model (LLM)-based assistant to improve how internal employees interact with and understand the data. This tool would enable users to query and analyze complex datasets using natural language, eliminating the need for technical expertise or manual examination of individual data points. Such a system would democratize access to insights across departments, empowering employees at all levels to leverage the data effectively. It would also enhance productivity by streamlining the process of deriving actionable insights, fostering a more data-driven organizational culture.

By incorporating cloud deployment, real-time weather integration, and an LLM assistant, the organization could solidify its position as a leader in data-driven decision-making and sustainable energy solutions.

References

ENGIE (2020) 'Using AI and Weather Forecast To Optimize Renewable Energy Output', *ENGIE Innovation*. Available at: https://innovation.engie.com/en/news/news/new-energies/AI-weather-forecast-optimize-renewable-energy/18235 (Accessed: 27 November 2024).

IBM (2024) 'New IBM Study Data Reveals 74% of Energy & Utility Companies Surveyed Embracing AI', *IBM Newsroom*, 26 February. Available at: https://newsroom.ibm.com/2024-02-26-New-IBM-Study-Data-Reveals-74-of-Energy-Utility-Companies-Surveyed-Embracing-AI (Accessed: 27 November 2024).

Cegal (2024) 'How AI and language models can change the energy industry', *Cegal*. Available at: https://www.cegal.com/en/resources/how-ai-and-language-models-can-change-the-energy-industry (Accessed: 27 November 2024).

Zhang, Y., Wang, J., and Li, X. (2024) 'Data-Driven Real-Time Congestion Forecasting and Relief With High Renewable Energy Penetration', *IEEE Transactions on Power Systems*, 39(5), pp. 4001-4012. Available at: https://ieeexplore.ieee.org/document/10683985 (Accessed: 5 December 2024).