Realtime Short-term Electricity Demand Forecasting with XGBoost: A Conceptual Data Driven Decision Support System for Grid Stability and Power Plant Management

Boylan Pardosi

Abstract
Accurate short-term electricity demand forecasting is essential for optimizing power generation and ensuring grid stability. This study presents the development of a real-time day-ahead electricity demand forecasting model using advanced machine learning techniques, specifically XGBoost. The model utilizes historical load data and weather information, incorporating significant temporal features identified through an in-depth exploratory analysis. The CRISP-DM methodology guided the systematic approach to data collection, processing, modeling, and evaluation. Key findings revealed distinct daily and weekly consumption patterns, which were critical in guiding feature selection and model development. The XGBoost model was selected for its superior performance in handling mixed data types and capturing complex patterns. Parameter tuning significantly improved the model's accuracy, and the inclusion of lag features allowed the model to effectively predict daily and weekly cycles. Comparative analysis with Random Forest and LSTM models confirmed the robustness of XGBoost. The forecasting model was integrated into a user-friendly dashboard, built using Apache Spark for data processing and Apache Superset for visualization, providing stakeholders with real-time updates and actionable insights, enabling optimized power plant operations and proactive grid management.

Keywords: forecasting, electricity demand, power plant, machine learning, decision support system

I. Introduction

On 12 June 2006, a major electrical blackout plunged Auckland into chaos, starting at 08:30 local time and affecting approximately 230,000 customers and at least 700,000 people. The outage disrupted daily life and essential services: suburban commuter railways halted, over 300 traffic lights ceased functioning, and several hospitals operated on emergency services only. Communication channels were compromised with radio transmitters offline and mobile phone services failing. The blackout left people stranded in office lifts, postponed university exams, and caused widespread business disruptions, as companies sent employees home amidst the central city's powerlessness (Johnston, 2018). A report by the Electric Power Research Institute (EPRI) estimated that power interruptions cost the U.S. economy about $150 billion annually (Lacommare, 2006). In developing countries, these costs can be even higher relative to their GDP.

One of the major factors contributing to power interruptions globally is grid instability. Grid instability, often caused by imbalances between electricity supply and demand, can lead to frequent and widespread power outages. Studies have shown that grid instability is a significant concern, particularly in regions with aging infrastructure or high integration of intermittent renewable energy sources (International Energy Agency, 2021). One effective way to address grid instability is through accurate forecasting of electricity consumption. By predicting demand patterns in advance, grid operators can better balance supply and demand, reducing the risk of overloads and power outages. Accurate forecasting allows deployment of demand-response strategies, ultimately enhancing grid stability and reliability.

In addition to grid stability, forecasting electricity consumption provides numerous other benefits. It optimizes operational efficiency by ensuring power plants run efficiently, reducing fuel costs and emissions. It also enhances the integration of renewable energy sources by predicting their variable output and adjusting other power sources, supporting environmental sustainability and regulatory. In electricity markets, they enable better trading decisions, enhancing market efficiency and reducing price volatility (Uddin, et al., 2021).

Electricity consumption forecasting can be categorized into short-term, medium-term, and long-term forecasts, each serving distinct objectives (Panapakidis, 2020). Short-term forecasting, typically ranging from hours to a few days ahead, aims to ensure the real-time balance between supply and demand, optimize daily operations, and prevent immediate grid instability. Medium-term forecasting, spanning from

weeks to months, focuses on planning maintenance schedules, managing fuel procurement, and preparing for seasonal demand variations. Long-term forecasting, covering years ahead, is crucial for strategic planning, infrastructure development, and policy-making, as it helps in anticipating future demand trends, guiding investments in new generation capacities, and shaping energy policies to meet long-term sustainability goals. This paper is focused on short-term electricity consumption forecasting, emphasizing its critical role in maintaining grid stability and optimizing daily energy management.

II. Research Problem

The main challenge in electricity demand forecasting lies in accurately predicting future load values, which are influenced by various factors including weather conditions, time of day, and historical demand patterns. Traditional statistical methods often find it difficult in capturing these complex relationships, thus the use of advanced machine learning models. In this study, we will analyze three years of electricity load data from a substation in Auckland to develop a model capable of accurately forecasting day-ahead electricity consumption at a half-hourly resolution. We will extract features from the time and date information to as the primary predictors. To enhance the accuracy of our predictions, we will include additional information, including weather data.

III. Research Objective

The primary aim of this study is to develop a model capable of accurately forecasting electricity demand for the next 24 hours, with predictions made at half-hourly intervals which is crucial optimizing electricity generation and distribution. Therefore, our research objectives are:

1. Analize electricity consumption patterns in Auckland.
2. Develop a robust prediction model for day-ahead electricity demand.
3. Evaluate model's performance and do comparison with other models.
4. Plan the implementation in Real-time prediction dashboard.

IV. Literature Review

The prediction of electricity consumption has been extensively studied, employing a variety of approaches. Generally, these approaches can be categorized based on how the forecasting problem is framed, falling into two main categories: time series and regression/tabular methods (Zhe, Tianzhen, Han, & Piette, 2021). The time series method emphasize the temporal order of data to predict future values based on past observations. This method is particularly effective for capturing trends, seasonality, and autocorrelations inherent in electricity consumption data. Common algorithms in this method are Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA). ARIMA models are widely used for time series forecasting due to their simplicity and effectiveness in handling different types of data patterns (Box, Reinsel, Jenkins, & Ljung, 2015). They combine autoregression (AR), differencing (I), and moving average (MA) components to model the data. SARIMA extends ARIMA by incorporating seasonal components, making it suitable for data with seasonal patterns, such as electricity consumption that varies with seasons. Chodakowska et. al. (2021) studied the use of ARIMA in electrical load forecasting and its robustness to noise. They concluded that although the predictions were not perfectly accurate, ARIMA models are quite robust to random noise if the data preprocessing stage in data mining and learning is properly conducted. Bilgili and Pinar (2023) compare SARIMA and a machine learning model (LSTM) to forecast long-term gross electricity consumption in Türkiye. It was found that despite the similarity, LSTM model outperform SARIMA in terms of mean absolute percentage error (MAPE). Another commonly used method in electricity demand forecasting is the exponential smoothing method (ETS). ETS models apply exponential smoothing of the electricity load time series curve to capture trend and seasonality. In his study, Pełka (2023) showed the high accuracy of ETS model (among other statistical methods) in predicting monthly electricity demand. In addition to classic time series methods previously mentioned, machine learning algorithms such as LSTM gained its popularity in time series forecasting. It's a type of recurrent neural network (RNN), that is capable of learning long-term dependencies in time series data. Chung and Jang (2022) use LSTM model for monthly electricity consumption forecasting for provincial scale, concluding it's near perfect accuracy by incorporating economic insights in the model.

In contrast to the time series method, regression-based methods make use of features rather than the time-

ordered data. Popular techniques for this approach are mainly machine learning algorithms random forest (RF), gradient boosting machine (GBM), support vector (SVR). Classic statistical techniques such as linear regression can be used, harnessing advanced computation power in determining combination of features that produce higher accuracy. RF is an ensemble learning method that constructs multiple decision trees during training and outputs the mean prediction of the individual trees, able to handle both categorical and continuous variables effectively. In his study, Dudek (2022) concludes that RF performs better than both statistical and machine learning models in short-term horizon, using nation-wide dataset. GBM, including is powerful ensemble methods that build models in a stage-wise fashion, correcting errors of the previous models. XGBoost is an algorithm based on GBM, equipped with parameters to control the learning environment, and avoid overfitting. Bae et. al. (2022) compares XGBoost and LSTM models to forecast day-ahead electricity load. They conclude that using XGBoost creates 21-29% improvement in terms of MAPE, leveraging the hyper-parameter tuning. SVR looks for optimal hyperplane to maximize the margin between different feature spaces, using kernel space to handle non-linear separations. Duan et. al. (2019) improved the traditional SVR model with mixture maximum correntropy criterion as the cost function, achieving low MAPE in predicting monthly electricity consumption. Table 1 summarize various algorithms that have been applied in electricity consumption forecasting.

Table 1 Algorithms in electricity consumption forecasting

Reference	Forecasting Horizon	Scale	Algorithms
Chodakowska et.al.(2021)	Short-term	Nation-wide	ARIMA
Bilgili and Pinar (2023)	Long-term	Nation-wide	SARIMA
Pełka (2023)	Medium-term	Nation-wide	ETS
Chung and Jang (2022)	Medium-term	Province	LSTM
Dudek (2022)	Short-term	Nation-wide	RF
Bae et. al. (2022)	Short-term	Nation-wide	XGBoost
Duan et. al. (2019)	Medium-term	Building	SVR

V. Research Methodology

This study follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which is a comprehensive framework for data mining projects. The CRISP-DM process consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. This structured approach ensures a systematic and thorough analysis for developing a real-time day-ahead electricity demand prediction model.

The first phase, Business Understanding, involves defining the primary objectives of the study, which are to develop an accurate predictive model for day-ahead electricity demand with half-hourly resolution, ensure the model meets company needs for accuracy and reliability, and implement the model in a real-time prediction dashboard. In the Data Understanding phase, historical load data and relevant weather data are collected. This phase involves describing, exploring, and assessing the quality of the data to identify trends, patterns, and anomalies that will help in building prediction model. The Data Preparation phase involves cleaning the data, handling missing values, and engineering new features to enhance the model's predictive capabilities. The integrated dataset must be prepared in a format suitable for modeling, ensuring it is clean, consistent, and ready for analysis.

During the Modeling phase, various machine learning algorithms are explored and selected based on their suitability for the forecasting objectives. Both time series models and tabular models are considered to leverage their respective strengths in capturing temporal patterns and utilizing various predictive features. The Evaluation phase focuses on assessing the performance of the developed models using appropriate evaluation metrics. Models are validated to ensure they meet the predefined accuracy criteria. Iterative refinement and tuning are conducted to improve model performance and robustness.

Finally, in the Deployment phase, the selected model is implemented in a real-time prediction dashboard. This phase ensures that the model integrates seamlessly with existing systems and provides actionable insights to stakeholders. Continuous monitoring and updating of the model are essential to maintain its accuracy and reliability over time. By following the CRISP-DM methodology, this study ensures a systematic and comprehensive approach to developing a robust and accurate real-time day-ahead electricity demand prediction model.

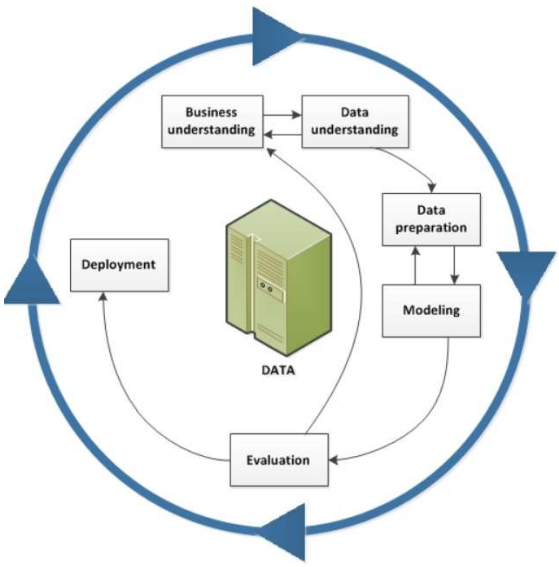


Figure 1 Steps in CRISP-DM Cycle (IBM, 2018)

VI. Design Process

Data Gathering. Historical load data was obtained from the Electricity Authority of New Zealand, covering the period from 2021 to 2023. This dataset included half-hourly electricity consumption values. Additionally, weather data, such as temperature and humidity, was sourced from reliable weather APIs. This data was collected on an hourly basis and provided crucial external variables that could influence electricity demand.

Exploratory Analysis. Various visualization techniques were employed to examine the temporal patterns of electricity consumption, including daily cycles and weekly trends. For instance, it was observed that electricity demand typically peaked during certain hours of the day and varied significantly between weekdays and weekends. The analysis also explored how external features like temperature and humidity influenced electricity consumption (see appendix). Correlation analysis helped identify the strength and nature of these relationships, providing insights into which features would be most valuable for the predictive models.

Data Processing. Initially, data cleaning was performed to handle missing values and remove outliers. Imputation techniques were used to fill in missing values, ensuring a complete dataset. Outliers were identified and removed to prevent them from skewing the analysis and model predictions. Data reduction was carried out both vertically and horizontally. Vertically, irrelevant rows that did not contribute to the analysis, such as those with null values that couldn't be imputed, were removed. Horizontally, unnecessary columns that did not add value to the prediction, such as identifiers and other non-informative attributes, were dropped. This streamlined the dataset and reduced computational complexity. Feature extraction was another important aspect of data processing. New features were created to enhance the predictive power of the models. These included temporal features like the day of the week, holiday indicators, and trading periods. Additionally, data transformation techniques, such as normalization and scaling, were applied to ensure that all features were on a comparable scale, improving model performance. The cleaned and transformed load data was then merged with the weather data based on their timestamps, resulting in a comprehensive dataset ready for modeling.

Algorithm Selection. The selection process considered both time series models and tabular models, each chosen for their specific strengths in handling the data and meeting the study's objectives Time series models, such as ARIMA and LSTM, are particularly adept at capturing temporal dependencies and seasonality inherent in electricity consumption data. ARIMA models are known for their proficiency in modeling linear relationships and handling seasonal patterns, making them effective for short-term forecasting tasks where the temporal order of data is crucial. LSTM networks, a type of recurrent neural network, excel in learning long-term dependencies and non-linear patterns, providing robust performance in capturing complex temporal dynamics. On the other hand, tabular models, such as XGBoost (Extreme Gradient Boosting), offer significant advantages in handling a wide range of features, including both temporal and exogenous variables. XGBoost is renowned for its robustness, high accuracy, and efficiency, especially when dealing with large datasets (Grinsztajn.Léo, Oyallon, & Gaël, 2022). The flexibility of XGBoost in incorporating various types of features and its ability to manage non-linear relationships and interactions between features make it an ideal choice for scenarios where multiple factors influence the predictions (Wang, Shi, Lyu, & Deng, 2017). Given these considerations, XGBoost was selected as the primary algorithm for developing the day-ahead electricity demand prediction model. The decision to use XGBoost was driven by its superior performance in

handling mixed data types and its efficiency in processing large datasets. Additionally, XGBoost's capability to incorporate lag features allows it to accommodate the seasonal patterns identified during the exploratory analysis. By using lag features, the model can effectively capture the temporal dependencies and seasonal variations in electricity consumption, enhancing its predictive accuracy. Despite choosing XGBoost as the primary algorithm, the study also includes a comparative analysis with LSTM, an algorithm that is also popular in forcasting. This comparison ensures a comprehensive evaluation of model performance and validates the robustness of the XGBoost model.

VII. Implementation Process

To proceed building the model with XGBoost, we first do the feature selection. It began with an in-depth analysis of the correlation matrix (see appendix) to identify the most relevant lag features. By examining the correlation matrix, we identified the lag periods that had the highest correlation with the target variable. Lag features, such as the previous day's load at the same half-hour interval, were included to capture daily cycles and seasonal variations observed during the exploratory analysis. Non lag features include month, day_of_week, is_holiday, temperature, and humidity. The training and test data splitting is set to 2:1 as we have 3 years of observations, and we want to have complete and balanced information related to time and seasonal features. Hence the first 2 (2021 and 2022) years are for training and the year 2023 data for testing.

The initial model development involved training the XGBoost model without any parameter tuning. This baseline model provided a reference point to understand the default performance of the algorithm. The initial training was conducted using the selected features, including the lag features identified from the correlation matrix. Following the initial model development, parameter tuning was performed to optimize the model's performance. Grid search and cross-validation techniques were employed to identify the best combination of hyperparameters, including the learning rate, maximum depth, and number of trees. The parameter tuning process involved systematically adjusting these parameters to minimize the prediction error and prevent overfitting. Since there are 48 half hourly prediction targets with different correlation magnitude with the lag features, we built 48 set of parameters.

There are 3 main parameters to be fine-tuned: *eta*, *max_depth*, and *lambda*. *Eta* is the learning rate, a parameter that controls the step size at each iteration while moving toward a minimum of the loss function. It scales the contribution of each new tree added to the model, with the goal of making the boosting process more conservative and reducing the risk of overfitting. *Max_depth* specifies the maximum depth of each tree in the model. It controls the complexity of the model by limiting how deep the individual trees can grow, which in turn influences the model's ability to capture patterns in the data. *Lambda* represents L2 regularization on the leaf weights that helps to prevent overfitting by adding a penalty proportional to the sum of the squared values of the leaf weights to the loss function. We only provide 3 values for each parameter to ensure reasonable computation runtime.

Upon finishing the parameter tuning process, we observe that selected parameters tend to be conservative (low *eta*, low *max_depth*, and high *lambda*). A scenario was considered where the weather data was excluded from the feature set to evaluate its impact on the model's performance. This consideration was driven by two main reasons. First, weather data used in actual predictions would itself be a product of forecasting, which could further increase the prediction error. Second, excluding weather data helps avoid multicollinearity, as weather data and the month feature can be correlated. The model was retrained using only the historical load data and temporal features, such as the day of the week and lag features.

A scenario was considered where the weather data was excluded from the feature set to evaluate its impact on the model's performance. This consideration was driven by two main reasons. First, weather data used in actual predictions would itself be a product of forecasting, which could further increase the prediction error. Second, excluding weather data helps avoid multicollinearity, as weather data and the month feature can be correlated. The model was retrained using only the historical load data and temporal features, such as the day of the week and lag features.

In addition to the XGBoost model, LSTM (Long Short-Term Memory) and Random Forest models were developed for comparison. LSTM networks are well-suited for time series forecasting due to their ability to capture long-term dependencies and non-linear patterns in sequential data. The architecture of the LSTM model included layers designed to capture the

temporal dynamics of electricity consumption. Random forest models extract information from features by creating multiple decision trees to capture complex patterns that can improve prediction accuracy for future electricity demand. These models were trained on the same dataset, with the same train-test split as the XGBoost model. The performance of the LSTM and Random Forest models were evaluated using the same metrics, allowing for a direct comparison with the XGBoost model.

VIII. Result and Interpretation

The analysis of electricity consumption data revealed distinct patterns in daily and weekly cycles. Daily patterns showed significant peaks in demand during certain hours, typically corresponding to morning and evening activities, with a noticeable dip during late night hours. Weekly patterns indicated higher consumption on weekdays compared to weekends (Figure 3), reflecting industrial and commercial activities. Seasonal variations were also evident, with higher demand during colder and warmer months due to heating and cooling needs, respectively.

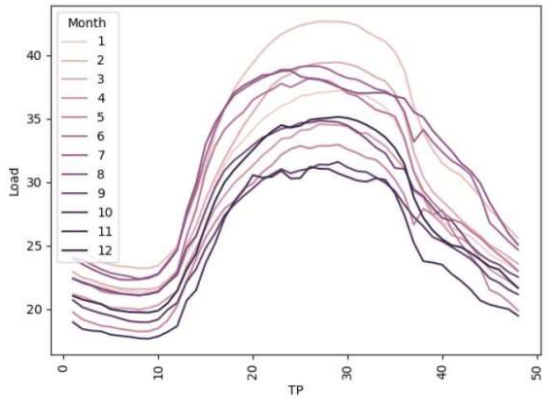


Figure 2 Monthly average of electricity demand against TP.

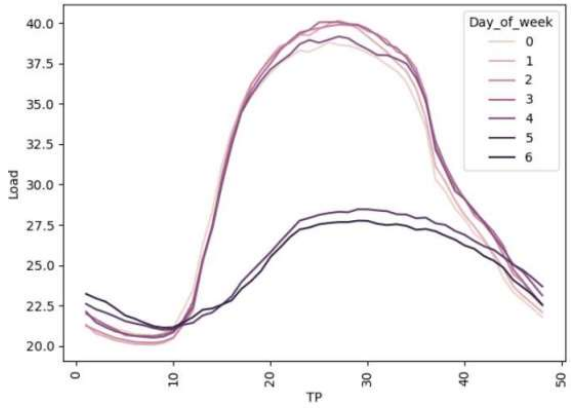


Figure 3 Day of week average of electricity demand against TP.

The developed models successfully recognized and captured these daily and weekly patterns (Figure 4). The use of lag features allowed the models to account for temporal dependencies, enabling accurate predictions of peak and off-peak hours. The XGBoost model, in particular, effectively made use of these lag features to mirror the cyclical nature of electricity demand. This capability was further enhanced through parameter tuning, which optimized the model's sensitivity to these temporal patterns.

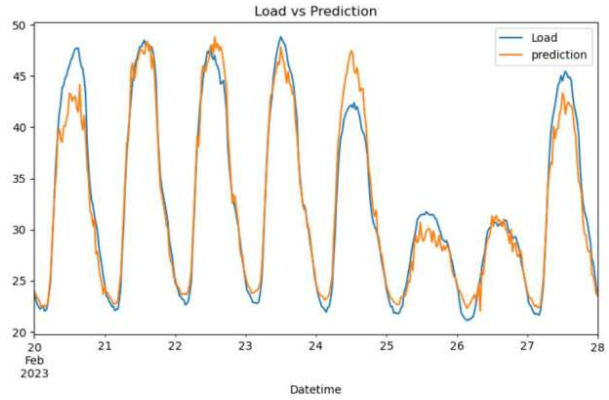


Figure 4 Prediction vs actual demand in daily scale.

The error measurement distribution of the prediction model reveals that most predictions exhibit very low error, with the distribution closely following a gamma distribution, having a mode near zero. However, there are a few predictions with very high error, which raises the overall mean error, resulting in a significant gap between the median and mean error. This indicates that while the model performs well in general, a small number of high-error predictions skew the average error metrics.

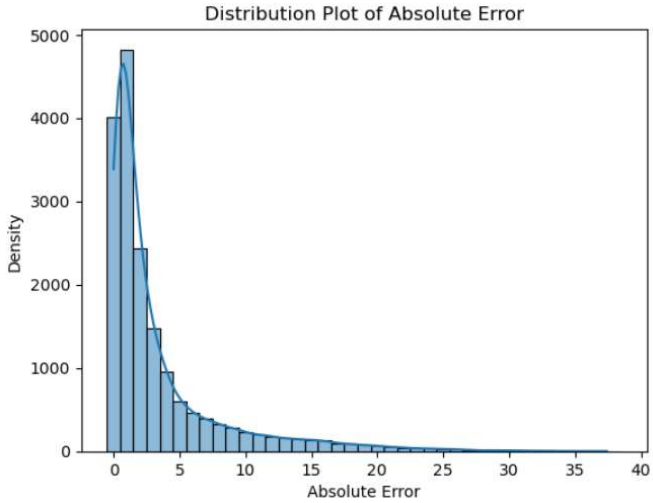


Figure 5 Absolute error distribution.

When plotting the error against trading periods (TP) shown in Figure 6, it became evident that certain TPs, particularly TP35-37 (which represent 5-7 PM), experienced very high errors. These trading periods coincide with the transition from office to domestic activities, a time of day characterized by high variability in electricity consumption. The shift in activities during this period contributes to the increased prediction error, highlighting a challenge for the model in accurately forecasting demand during these times of significant change.

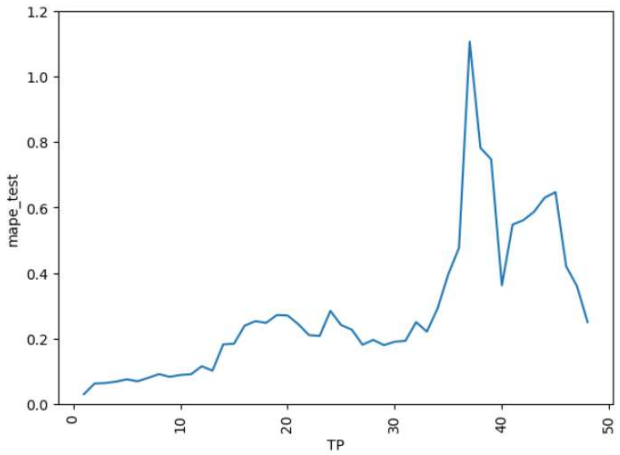


Figure 6 Error distribution across TP.

Parameter tuning significantly improved the model's performance. The initial model without parameter tuning served as a baseline, showing good predictive capabilities but with some limitations in handling complex patterns and temporal dependencies. After parameter tuning, the XGBoost model demonstrated a marked reduction in error metrics such as MAE and MAPE. The tuning process optimized key parameters, enhancing the model's ability to capture nuanced patterns and dependencies in the data.

As depicted in the feature importance plot (Figure 7), lag features were the most important across all TPs, indicating the strong influence of previous consumption patterns on future demand. Temporal features such as the day of the week and month also contributed significantly, especially during certain TPs. Weather variables like temperature and humidity, while less important than lag features, still played a notable role, particularly during periods of high variability in weather conditions. This analysis highlights the critical factors that drive electricity consumption and underscores the importance of temporal dependencies in the prediction model.

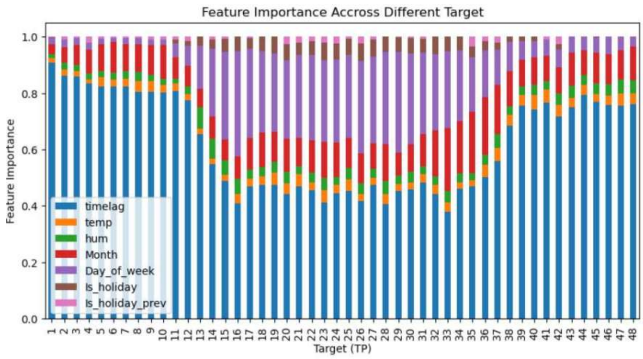


Figure 7 Feature importance plot.

Excluding weather data from the feature set provided valuable insights into its impact on model performance. While the model trained without weather data showed a slight increase in prediction error, this increase was not substantial. Despite the higher error, it might be preferable to exclude weather data because the actual weather data used for prediction is itself a forecast, which could accumulate errors and further impact the prediction accuracy. Therefore, the marginal increase in error when excluding weather data might be an acceptable trade-off to avoid the potential compounded errors from using forecasted weather data.

For comparative analysis, RF and LSTM models were also developed and evaluated. The RF model demonstrated good performance, benefiting from its ensemble approach to handle the variability in the data effectively. However, it did not outperform the tuned XGBoost model. The RF model's error metrics were slightly higher, indicating less precision in capturing complex temporal patterns and dependencies. While the RF model provided a robust baseline, the XGBoost model's advanced gradient boosting techniques and optimized parameters made it the superior choice for this specific forecasting task. The LSTM models perform worse than RF and XGBoost by all error measurements.

Table 2 Error Measurements for developed forecasting models

Model	Error	
	MAE	MAPE
XGBoost - Base	3.169	0.263
XGBoost - With Tuning	2.531	0.196
XGBoost - Without Weather	2.680	0.229
RF Model	3.798	0.276
LSTM Model	7.263	0.268

IX. Data-driven Actions

A. Developing the Forecasting Dashboard

The primary objective of this study is to develop a real-time day-ahead electricity demand forecasting dashboard. This dashboard will serve as a critical tool for stakeholders, providing them with timely and accurate predictions to optimize electricity generation and distribution. The development process involves several key steps, from automatic data ingestion to dashboard visualization, utilizing advanced technologies to ensure scalability and reliability.

The first step in developing the forecasting dashboard is to establish a robust data ingestion pipeline. This pipeline will automate the process of collecting and integrating data from various sources, including historical load data from the Electricity Authority of New Zealand and weather data from reliable APIs. Apache Spark will be used as the unified engine for large-scale data analytics, utilizing its capabilities to process large volumes of data efficiently. Spark's streaming capabilities will allow for real-time data ingestion, ensuring that the latest data is always available for forecasting.

Once the data is ingested, it will be processed to clean, transform, and prepare it for modeling. This involves handling missing values, normalizing data, and engineering features as discussed in the data preparation phase. The preprocessed data will then be fed into the XGBoost model, which has been optimized through parameter tuning to provide accurate predictions. Spark's MLlib (https://spark.apache.org/mllib/) library will be used to manage the model execution.

The next step is to develop the dashboard, which will be the interface for stakeholders to access the predictions. This dashboard will be built using Apache Superset (https://superset.apache.org/), an open-source data exploration and visualization platform. Superset provides a user-friendly interface and powerful visualization tools, making it ideal for displaying complex data in an accessible manner. The dashboard will feature real-time updates, visualizing the day-ahead electricity demand predictions along with historical trends and weather data. It will include various interactive elements, such as time-series graphs, heatmaps, and alert systems, to help users quickly interpret the data and make informed decisions.

The technology stack for the dashboard will include Apache Spark for data processing, Apache Kafka for real-time data streaming, Apache Superset for visualization, and a cloud-based platform like AWS or Azure for infrastructure. One of the main risks in this setup is the potential for data latency or loss during ingestion. To mitigate this, redundant data pipelines will be established, and robust error-handling mechanisms will be implemented. Additionally, regular audits and monitoring will be conducted to ensure data integrity and system performance.

B. Managing Power Plant and Grid Stabilization

The insights gained from the exploratory analysis and prediction results can significantly enhance the management of power plants and grid stabilization. With the predictive model's ability to forecast demand with high accuracy, power plant operators can optimize their operations by adjusting generation schedules to match expected demand. For instance, during periods identified with high variability, such as TP35-37 (5-7 PM), plants can be increased to ensure sufficient supply. Conversely, during periods of low demand, generation can be scaled back to save resources and reduce operational costs. This dynamic adjustment helps in maintaining a balance between supply and demand, minimizing waste, and improving overall efficiency.

Accurate demand forecasting also helps in ensuring grid stabilization. By anticipating peak demand periods, grid operators can take proactive measures to prevent overloads and ensure a stable supply. The patterns identified in the exploratory analysis, such as daily and weekly cycles, can be used to develop load-shedding plans and demand response strategies. For example, during peak hours, non-essential loads can be temporarily reduced, or energy storage systems can be utilized to buffer the grid. These measures help in maintaining voltage stability and preventing blackouts.

The forecasting model and dashboard also aid in risk management and contingency planning. By providing advance warnings of potential demand spikes, operators can prepare for contingencies such as equipment failures or sudden increases in demand. This foresight allows for better allocation of resources and ensures that backup systems are ready to be deployed when needed. Additionally, the integration of weather data helps in anticipating weather-related

demand surges, allowing operators to take preventive measures in advance.

X. Conclusion and Future Work

Through the systematic application of the CRISP-DM methodology, we ensured a comprehensive approach to data collection, processing, modeling, and evaluation, ultimately leading to the successful implementation of a forecasting dashboard.

The exploratory analysis revealed distinct patterns in daily and weekly electricity consumption, highlighting peak periods during mornings and evenings, as well as higher demand on weekdays compared to weekends. The XGBoost model was selected as the primary forecasting tool due to its flexibility in handling mixed data types and its superior performance in capturing complex patterns. Parameter tuning further enhanced the model's accuracy, demonstrating the importance of iterative refinement in achieving optimal results. The model's ability to leverage lag features effectively allowed it to recognize and predict daily and weekly cycles with high precision. The comparative analysis with other models, including Random Forest and LSTM, reinforced the superiority of the XGBoost model in this specific context. While the LSTM model showed promise in capturing long-term dependencies, the XGBoost model's efficiency and accuracy in processing large-scale data made it the preferred choice for real-time forecasting.

Future research should focus on several key areas to further enhance the effectiveness and applicability of the forecasting model:

1. Incorporate more detailed weather data, socio-economic factors, and special events that can influence electricity demand. This could improve the model's accuracy by capturing a wider range of influencing factors.
2. Explore the combination of different machine learning models, such as blending XGBoost with LSTM or other deep learning techniques. Hybrid models could utilize the strengths of each algorithm to improve predictive performance.
3. Conduct detailed analysis of prediction errors to identify and mitigate sources of high variance. This could involve refining the feature set, adjusting model parameters, or incorporating additional data preprocessing steps.
4. Implement automated pipelines for real-time data updates, ensuring that the model always has access to the most current information. This includes setting up continuous integration and deployment (CI/CD) practices for the model.

References

Bae, D.-J., Kwon, B.-S., & Song, K.-B. (2022). XGBoost-Based Day-Ahead Load Forecasting Algorithm Considering Behind-the-Meter Solar PV Generation. *Energies*.

Bilgili, M., & Pinar, E. (2023). Gross electricity consumption forecasting using LSTM and SARIMA approaches: A case study of Türkiye. *Energy*.

Box, G. E., Reinsel, G. C., Jenkins, G. M., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control, 5th Edition.* Hoboken, New Jersey: Wiley.

Chodakowska, E., Nazarko, J., & Nazarko, Ł. (2021). ARIMA Models in Electrical Load Forecasting and Their Robustness to Noise. *Energies* .

Dudek, G. (2022). A Comprehensive Study of Random Forest for Short-Term Load Forecasting. *Energies* .

Grinsztajn.Léo, Oyallon, E., & Gaël, V. (2022). *Why Do Tree-based Models Still Outperform Deep Learning on Tabular Data.* arXiv:2207.08815v1.

Johnston, M. (2018, April 16). *NZ Herald*. Retrieved from nzherald.co.nz: https://www.nzherald.co.nz/nz/a-crisis-recalled-the-power-cuts-that-plunged-the-auckland-cbd-in-darkness-for-five-weeks/IZBJMV3I4H4FOQIX3MIG5JBN3Y/

Lacommare, K. H. (2006). Cost of Power Interruptions to Electricity Consumers in the United States (US). *Energy* , 1845-1855.

Panapakidis, I. (2020). Short-Term, Medium-Term and Long-Term Load Forecasting: Methods and Applications (Special Issue). *Forecasting, 2*.

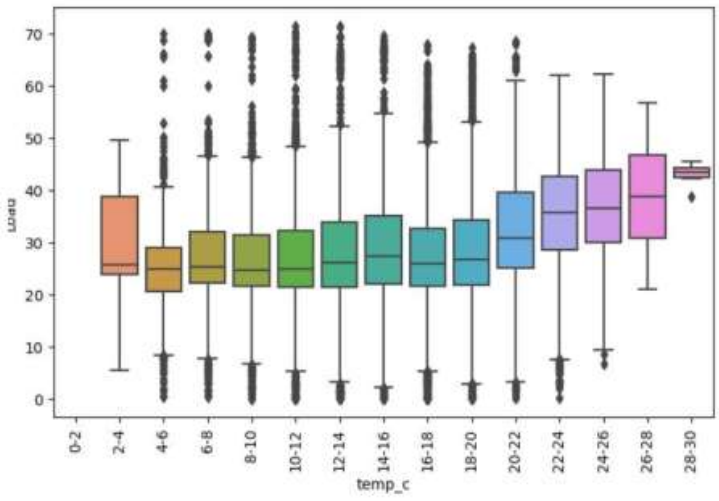
Pełka, P. (2023). Analysis and Forecasting of Monthly Electricity Demand Time Series Using Pattern-Based Statistical Methods. *Energies*.

Uddin, G. S., Tang, O., Sahamkhadam, M., Taghizadeh-Hesary, F., Yahya, M., Cerin, P., & Jakob, R. (2021). Analysis of Forecasting Models in an Electricity Market under Volatility. *ADBI Working Paper Series*.

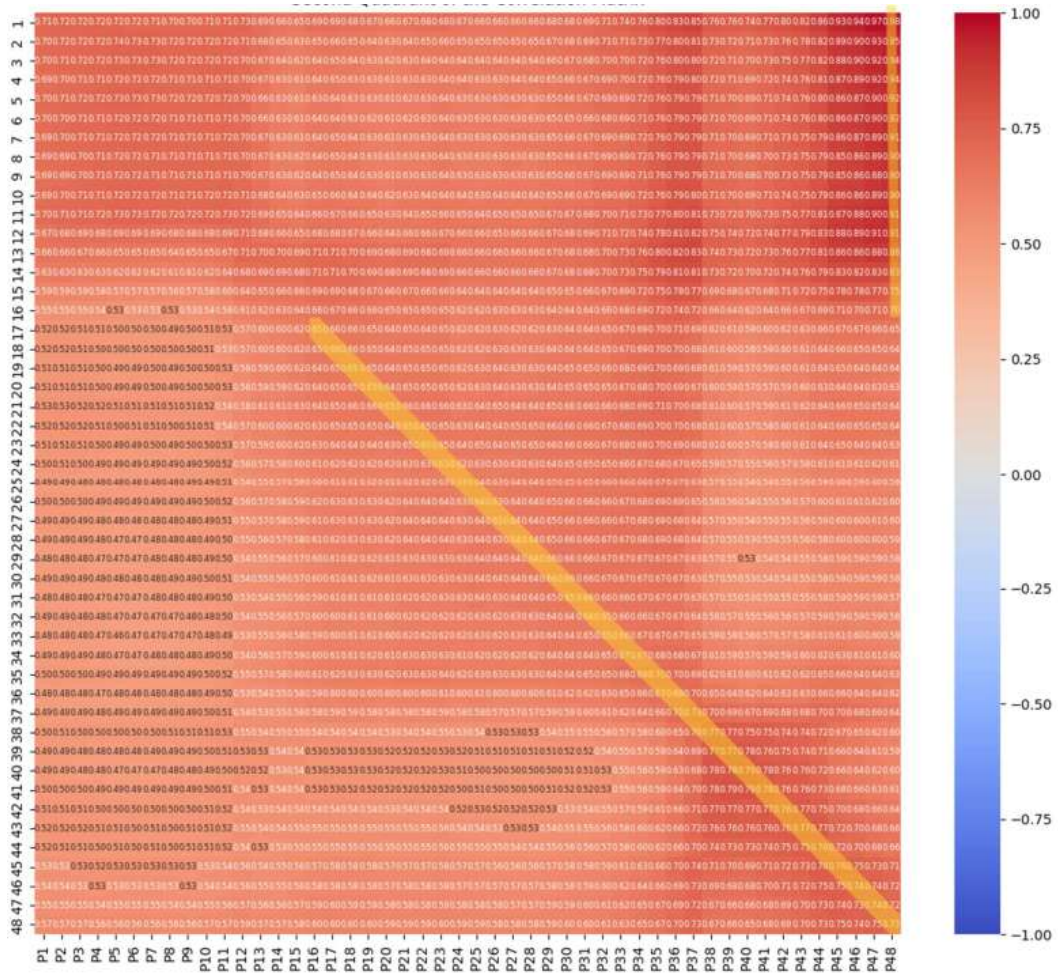
Wang, W., Shi, Y., Lyu, G., & Deng, W. (2017). Electricity Consumption Prediction Using XGBoost Based on Discrete Wavelet Transform. *DEStech Transactions on Computer Science and Engineering*.

Zhe, W., Tianzhen, H., Han, L., & Piette, M. A. (2021). Predicting city-scale daily electricity consumption using data-driven models. *Advances in Applied Energy, 2(100025)*.

Appendix I - Temperature vs Electricity Consumption



Appendix II - Lag Features Correlation Matrix



Appendix III - Error Distribution across Different Features