Effective Prediction of Electric Power Consumption using Random Forest and XGBoost Regressor

Sasikala M[1], Krithik U[2], Kishore Kumar B P[2], Pandi Selvam P[2]

[1]Professor and Faculty Mentor, Dept. of Computer Science and Engineering, K L N College of Engineering, Pottapalayam, Sivagangai, Tamil Nadu, India.
[2]B. E Student, Dept. of Computer Science and Engineering, K L N College of Engineering, Pottapalayam, Sivagangai, Tamil Nadu, India

---***---

Abstract - *Accurate prediction of electric power consumption is very important for efficient energy management and cost savings. Traditional methods are often known to not handle intricate power consumption patterns. In this work, a machine learning-based approach has been followed using the Random Forest and XGBoost Regressor algorithms to predict electric power consumption. The system applies hourly power consumption data combined with weather conditions and seasonal factors like temperature, rainfall, and humidity for predictions. If the accuracy is very low, the system adds 10%, if it is moderate then 5% on the predicted value. Such adjusted predictions it issues as its output; thus, ensuring that energy supplied equals demand. The management of energy, reduction in imbalances of supply, and healthy productive usage of energy will thus be improved.*

Key Words: **Random Forest, XGBoost Regressor, Electric Power Consumption Prediction, Machine Learning, Energy Forecasting.**

1. INTRODUCTION

Accurate electric power usage prediction is crucial for efficient energy management, cost-reducing savings, and promotion of sustainability. However, against the backdrop of the growth in world energy demand, traditional methods for forecasting often cannot cope with the complex factors influencing power consumption as advanced by various meteorological and seasonal and regional factors. Traditional methods for such forecasting are usually less accurate in non-linear usage data of power. These are the challenges that this project addresses by using Random Forest and XGBoost Regressor, two machine learning algorithms for the power consumption forecast. The system considers past records of the power usage together with weather data, comprising elements of temperature and rainfall, which give it predictive accuracy. Powerful pre-processing techniques would overcome outlier and missing data complications during the prediction in order to present reliable outcomes resulting from imperfect data from the real world. It also autonomously detects dominant factors like peak usage hours and extreme weather events to make its results easy to interpret. The system adjusts the calculated values by 10% in case of poor accuracy predictions and by 5% in case of moderate accuracy predictions for optimum accuracy of its predictions so that the produced energy is always kept up with its demand. This adaptive and scalable solution promises a more efficient way of handling power consumption, where energy sustains themselves across different regions and networks.

2. LITERATURE SURVEY

[1] Modeling and Forecasting Electricity Consumption Amid the COVID-19 Pandemic: Machine Learning vs. Nonlinear Econometric Time Series Models-Lanouar Charfeddine*, Esmat Zaidan*, Ahmad Qadeib*, Alban Hamdi Bennasr*, Ammar Abulibdeh. The paper compares machine learning and nonlinear econometric models for electricity consumption forecasting using data from the COVID-19 pandemic. It emphasizes how weather factors, like temperature and rainfall, affect energy usage and shows that the best accuracy is provided by machine learning models. However, accuracy can decrease when applying to years with changed weather conditions because it depends on past seasonal data.

[2] Spatio-temporal Granularity Co-optimization Based Monthly Electricity Consumption Forecasting-Kangping Li*, Yuqing Wang*, Ning Zhang*, Fei Wang*, Chunyi Huang. The paper recommends an increase in spatio-temporal granularity for enhanced monthly electricity consumption forecasting. When high-resolution load data from meters is used to optimize forecasting accuracy, the integration of information from space and time benefits results. The method developed for short-term forecasting did not discuss or test it on monthly forecasting. While assuming no noise in the load data, it may limit the application into real-world scenarios.

[3] Power Source Flexibility Margin Quantification Method for Multi-energy Power Systems Based on Blind Number Theory-Bai Xiao*, Jialiang Wang*, Zhiwen Xiao*, Gangui

Yan*, Ling Dong*, Maochun Wang*, Hongzhi Yang. The paper develops a model to quantify flexibility margins for each source of power within a multi-energy power system based on the blind number theory under uncertainty in power generation from the sources, such as thermal, hydro, gas, wind, and solar. The methodology is comprehensive but complex with the use of blind number theory, and hence could be problematic in practical application.

[4] Reliability of Linear Losses-to-Power Scaling Method of Electric Drive Systems-Ayoub Aroua*, Luis Ramirez*, Walter Lhomme*, Florian Verbelen*, Philippe Delarue*, Alain Bouscayrol*, Peter Sergeant*, Kurt Stockman. This paper used the linear scaling method as one applicable to predict the electric drive system efficiency, besides demonstrating a value of less than 1% deviation in such prediction of efficiency. It exhibited effectiveness in scaling up electric systems but downscaling is practically not efficient due to its associated high inaccuracies in predicting losses in copper.

3.METHODOLOGY

3.1 DATASET

The dataset used here is of 8,759 hourly observations from January 2020, capturing both power consumption and weather-related variables. The three target variables are Power Consumption_Zone1, Power Consumption_Zone2, and PowerConsumption_Zone3 that correspond to electricity consumption in kWh, by the three different zones. Although each record is related to certain weather parameters, including Temperature (°C),Humidity (%), and WindSpeed(m/s), among which are quite crucial demand factors for heating, cooling, and ventilation, it's interesting that the dataset also presents measurements of solar radiation through GeneralDiffuseFlows and DiffuseFlows (W/m^2) values, thus giving an impression of what the impact of solar energy might be on some power consumption patterns - mainly during daylight hours. All the other variables are recorded as floating-point numbers; the column for Datetime formatted as a datetime object, and is in good time series analysis. No missing values in this dataset: thus, all data quality for modeling is ensured to be high. This potentially comprehensive dataset will allow one to go deep into the relationship between environmental conditions and electricity consumption, thereby supporting predictive modeling to forecast energy demand and optimize grid efficiency across different zones.

3.2 DATA ACQUISITION

The data set captures the hour-by-hour energy consumption from residences, industries, and commerce sectors of one year, along with other influencing factors such as weather conditions-including temperature, humidity, wind speed; seasonal factors; and location-to represent varied patterns in energy use. Sources of energy include electricity, solar, wind, and hydropower, giving information regarding consumption patterns and trends in each region. The pre-processing phase then deals with cleaning the data to resolve any missing values, outliers, and inconsistencies, thereby making the dataset suitable for model training. Next, it splits the cleaned data into training and testing sets in order to apply the machine learning algorithms used, namely Random Forest and XGBoost Regressors. This more structured approach to data collection and analysis increases the predictive model's accuracy and leads to better forecasts about consumption patterns of power.

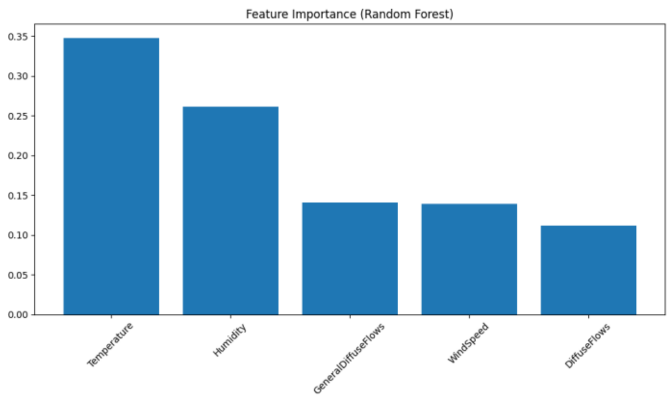

Fig -1: Factors Influencing Electric Power Transmission

This is to rank the importance of factors contributing to a given outcome using feature importance. Analyzing the importance of features can make it clear which variables are most important and need more attention in making decisions or refinement of the model. For instance, in this system, Temperature and Humidity are the key drivers, while others such as WindSpeed and DiffuseFlows contribute moderately. This allows better predictions or processes in the future to be designed around the most influential variables.

3.3 DATA DETOXIFICATION

The data detoxification stage serves to clean the dataset to ensure more accuracy and reliability of the power consumption predictions. After data has been acquired, this process involves cleaning the dataset with a total elimination of inconsistencies and noise - first of all, through interpolation or imputation on missing values, identification and correction of outliers, and normalization of the data to guarantee consistency in all variables. Anomalies are unexpected spikes in consumption and sensor errors that are caught and corrected so as not to skew the analysis. Other filtrations include irrelevant, redundant, or sensitive information so that the dataset remains relevant and clean. The detoxification effectiveness can be validated through the performance metrics of data completeness, error rate, anomaly detection rate, and consistency. It will continuously improve its cleansing methods based on those metrics such that the dataset remains of high quality and valid for machine learning. The more it streamlines the detoxification process from the data, the better the chances of utilizing the most reliable and relevant information that will directly influence the accuracy of power consumption forecasts.

3.4 RANDOM FOREST

This model bases its power consumption predictions on complex, nonlinear relationships that connect historical usage, the weather conditions, and seasonality variables, all through Random Forest. Using its ensemble approach to build multiple decision trees from different random data subsets, it avoids overfitting and, hence, generalizes well to unseen data. Random Forest can handle data in numeric and categorical, and it is very versatile to be applied to different types of data, such as appliance usage, temperature, and humidity. The model further detects the leading root causes through feature importance, thus allowing for more focused energy management strategies. It also properly handles missing values and outliers, thus being reliable even with imperfect data, and, consequently, providing accurate power consumption forecasts necessary for optimizing the use of energy resources.

3.5 XGBOOST REGRESSOR

This model focuses on using XGBoost in the prediction of power consumption by taking full advantage of its gradient boosting algorithm, with every decision tree built sequentially so as to correct the errors made by the previous one. High accuracy and scalability are the attributes of XGBoost, and it is one of the best choices to apply on big complex datasets where non-linear relationships and interactions occur among several features like weather conditions, time of the day, and appliance usage, among others. The algorithm uses regularization, L1 and L2, to counter overfitting and enhance model generalization. Feature importance is also provided by XGBoost, enabling one to identify the most influential variables on the consumption prediction. It also has the ability to accommodate missing data as well as efficiently compute sparse datasets, hence robust even with imperfect data. Hence, benefits of using XGBoost include faster training, improved predictive performance, and more accurate power consumption forecasts, all necessary to an effective energy management strategy.

4. PREDICTIVE ANALYSIS

The model, therefore, forecasts the energy consumption for a future point in time through the use of historical data merged with prominent factors like weather and seasonal, as well as appliance use patterns. Complex, non-linear relationships between variables informing demand energy are reflected using machine learning algorithms like Random Forest and XGBoost Regressor. The dataset was preprocessed, involving corrections to missing values and outliers so as not to bias the prediction. Performance measures rely on Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) to estimate the accuracy of the forecasts from the results of their comparison with actual consumption. The mechanism adjusts dynamically: in the case of low accuracy, it augments the predicted consumption levels by 10%; for moderate accuracy, it is increased by 5%, and does not change the forecasted values in high-confidence cases. This approach enhances the reliability of the forecasts, allowing better distribution of resources and cut costs, that means sustainable energy management.

5. POWER CONSUMPTION PREDICTION

There is a dataset regarding historical energy usage, which the relevant factors associated with it are weather conditions, time of day, and seasonal changes. These data are preprocessed by removing outliers and handling missing values to make a clean and reliable dataset. The machine learning algorithms, namely Random Forest and XGBoost Regressor, have been used to find complex interlink relationships between these variables for predicting power consumption for the future. These models are trained against a training dataset and the testing dataset only to validate the prediction accuracy. The system measures its performance using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) to ensure the most accurate possible predictions, thus enabling the energy demands to be confidently predicted.

The system has used an incremented dynamic methodology that is predictive in nature regarding the precision of the prediction. If the accuracy of prediction is low, the consumption value from the predicted value is incremented by 10% to avoid low-valued underestimation of demand. At moderate levels of accuracy, a 5% adjustment is used; no adjustment at all in the case of high-confidence predictions. The adjustment serves as an adaptation method to build the system toward any confidence level in its predictions, thereby reducing the chance of imbalance in supply. The adjusted power consumption forecast becomes the final output and will be delivered to the energy management system. It keeps on monitoring and refining the truth of the system to enable real-time adjustment in energy production, optimize resource usage, and improve energy efficiency. This approach will provide reliable, actionable insights toward making decisions that make better energy management and sustainability.

6. RESULT

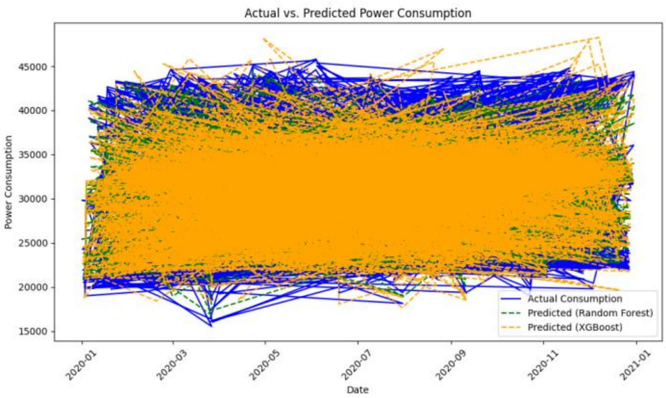

Chart -1: Actual vs Predicted Power Consumption

A very good performance of the power consumption prediction system is obtained from the Random Forest model with an overall accuracy of 89.14%. Such a high accuracy reflects the capability of the model to capture complex patterns and relationships in the data, which makes it very efficient for forecasting future energy demand. The ensemble approach of the Random Forest helps prevent overfitting and enhances capacity for generalization, thus greatly contributing to the reliability of the system.

The overall accuracy the XGBoost model generates, following closely to the one in this study, is 88.32%. Although slightly lower than Random Forest, boosting techniques successively refine the model by correcting errors, so predictions of power consumption become very accurate. This indicates that the two models make a system of predicting power consumption quite reliable for optimization of energy management and resource allocation, with Random Forest showing a slightly better performance.

7. FUTURE ENHANCEMENT

Renewable sources of energy, including solar and wind, may be integrated into the future upgrading of power consumption prediction systems to consider variables of output. Further development can include the integration of energy storage solutions, such as batteries, to optimize usage by predicting when to store excess power and when to release according to peak demand. That, in addition, would allow energy managers the visualization of any consumption trends in real-time analytics dashboard capability. Integration with IoT devices and smart meters might help in more dynamic and up-to-date predictions of energy demand based on how the patterns are changing because new data is flowing into the system. By refining the system also through continuous learning algorithms, the model can learn and change over time, making it also much more precise in tracking shifting patterns of energy consumption.

8. CONCLUSIONS

The power consumption prediction system exhibits high accuracy in energy demand prediction, particularly by machine learning algorithms such as Random Forest and XGBoost Regressor. Analyzing historical data, key influencing factors that are valid, such as weather conditions, and seasonal variations, yield dependable predictions in the generation of the framework to assist in adjusting energy management and resource utilization. Dynamic incrementation techniques will be included in the system so that real-time forecasts are adjusted to the accuracy needed, thus nearly accommodating actual demand on a regular basis. Overall, it would thus ensure significant improvements in operation efficiency, costs reduction, and contribution towards energy sustainability. Improvements might even happen later on - the inclusion of renewable sources for energy and real-time data from smart grids - but this system may prove to represent a promising approach for more advanced energy management practices.

REFERENCES

[1] Mobarak Abu Mohsen, Amani Yousef Owda, Majdi Owda (2023) 'Electrical Load Forecasting Based on Random Forest, XGBoost, and Linear Regression Algorithms', 2023 International Conference on Information Technology (ICIT), ResearchGate

[2] Harshit Rathore, Hemnath Kumar Meena, Prerna Jain (2023) 'Prediction of EV Energy consumption Using Random Forest And XGBoost', 2023 International Conference on Power Electronics and Energy (ICPEE), ResearchGate

[3] Yasminah Alali, Fouzi Harrou, Ying Sun (2023) 'Unlocking the Potential of Wastewater Treatment: Machine Learning Based Energy Consumption Prediction', Article, ResearchGate

[4] Witwisit Kesornsit, Yaowarat Sirisathitkul (2022) 'Hybrid Machine Learning Model for Electricity Consumption Prediction Using Random Forest and Artificial Neural Networks', Article, ResearchGate

[5] Pedro C. Albuquerque, Daniel O. Cajueiro, Maria D. C. Rossi (2021) 'Machine learning models for forecasting power electricity consumption using a high dimensional dataset', Article, ResearchGate

[6] Weizeng Wang, Yuliang Shi, Gaofan Lyu, Wanghua Deng (2017) 'Electricity Consumption Prediction Using XGBoost Based on Discrete Wavelet Transform', Article, ResearchGate

[7] Abulibdeh, A., Zaidan, E., & Jabbar, R. (2022). The impact of covid-19 pandemic on electricity consumption and electricity demand forecasting accuracy: empirical evidence from the state of qatar. Energy Strategy Reviews, 44, 100980. Science Direct

[8] Arce, J. M. M., & Macabebe, E. Q. B. (2019). Real-time power consumption monitoring and forecasting using regression techniques and machine learning algorithms. In 2019 IEEE international conference on internet of things and intelligence system (IoTaIS) (pp. 135–140). IEEE.