JOURNAL OF SCIENCE
OF LAC HONG UNIVERSITY
ISSN: 2525 - 2186
Tạp chí Khoa học Lạc Hồng, **2025**, 22, 081-087

XÂY DỰNG MÔ HÌNH DỰ BÁO NHU CẦU PHỤ TẢI ĐIỆN BẰNG KỸ THUẬT ENSEMBLE LEARNING

Nguyễn Trung Phương[1], Dương Thị Kim Chi[1*]
[1]*Viện đào tạo công nghệ thông tin, chuyển đổi số, Trường Đại học Thủ Dầu Một*
* Tác giả liên hệ: *chidtk@tdmu.edu.vn*

THÔNG TIN BÀI BÁO

Ngày nhận: 10/7/2025
Ngày hoàn thiện: 8/8/2025
Ngày chấp nhận: 4/9/2025
Ngày đăng: 15/9/2025

TỪ KHÓA

Random Forest;
Dự báo tiêu thụ điện;
Dự báo chuỗi thời gian;
XGBoost;
Ensemble learning.

TÓM TẮT

Dự báo chính xác mức tiêu thụ năng lượng của doanh nghiệp đóng vai trò quan trọng trong việc tối ưu hóa sản xuất và phân phối điện. Nghiên cứu này đề xuất một mô hình dự báo chuỗi thời gian ứng dụng các thuật toán thuộc nhóm học tập tổng hợp mạnh mẽ như XGBoost, Random Forest nhằm dự đoán nhu cầu tiêu thụ điện hàng ngày dựa trên dữ liệu lịch sử tiêu thụ điện. Nghiên cứu sử dụng bộ dữ liệu tiêu thụ điện được thu thập tại một khu công nghiệp để huấn luyện, kiểm tra và xây dựng mô hình. Ngoài ra, Chúng tôi còn sử dụng và so sánh hiệu suất các mô hình sử dụng các phương pháp chuẩn hóa dữ liệu thông dụng nhất hiện nay. Mô hình đề xuất cho thấy khả năng thích ứng cao cho cả dữ liệu có tính biến động cao do nhiễu và mất dữ liệu quan sát. Mô hình của chúng tôi vượt trội hơn các phương pháp hiện có về độ đo MAE và R^2 với thuật toán XGBoost lần lượt là 0.027 và 98%. Khung mô hình đề xuất mở ra triển vọng ứng dụng mạnh mẽ trong việc quản lý nhu cầu năng lượng và tối ưu hóa chi phí đầu tư, vận hành lưới điện, đặc biệt hỗ trợ ra quyết định hiệu quả trong môi trường vận hành phức tạp, góp phần nâng cao độ tin cậy và hiệu suất của lưới điện thông minh.

DEVELOPING A MODEL FOR POWER LOAD DEMAND FORECASTING USING ENSEMBLE LEARNING TECHNIQUES

Nguyen Trung Phuong[1], Duong Thi Kim Chi[1*]
[1]*Institute of Information Technology and Digital Transformation Training, Thu Dau Mott University, Ho Chi Minh City, Vietnam*
*Author contact: *chidtk@tdmu.edu.vn*

ARTICLE INFO

Received: Jul 10th, 2025
Revised: Aug 8th, 2025
Accepted: Sep 4th, 2025
Published: Sep 15th, 2025

KEYWORDS

Random forest;
Electricity consumption;
Forecasting;
Time series forecasting;
Xgboost;
Ensemble learning.

ABSTRACT

Accurate forecasting of enterprise energy consumption plays a vital role in optimizing electricity production and distribution. This study proposes a time series forecasting model that applies powerful ensemble learning algorithms such as XGBoost and Random Forest to predict daily electricity demand based on historical consumption data. The dataset used for training, testing, and model development was collected from an industrial zone. In addition, we evaluate and compare the performance of models using several commonly adopted data normalization techniques. The proposed model demonstrates high adaptability to volatile data caused by noise and missing observations. Our model outperforms existing approaches in terms of MAE and R^2 metrics, achieving 0.027 and 98%, respectively, with the XGBoost algorithm. This modeling framework shows strong potential for energy demand management and investment cost optimization, as well as for the operation of smart grids. It is particularly effective in supporting decision-making under complex operating conditions, contributing to improved reliability and efficiency of modern power systems.

Doi: https://doi.org/10.61591/jslhu.22.957

Available online at: https://js.lhu.edu.vn/index.php/lachong

1. GIỚI THIỆU

Hiện nay dự báo nhu cầu phụ tải điện ngày càng nhận được nhiều sự quan tâm do tính chất và tầm quan trọng của nó trong công tác điều độ [1], quyết định đầu tư phát triển cơ sở hạ tầng điện cũng như phát triển kinh tế, đảm bảo an ninh năng lượng và đời sống vì vậy đã thu hút nhiều nghiên cứu liên quan trong lĩnh vực dự báo này. Trong nghiên cứu này chúng tôi đề xuất mô hình dự báo phụ tải dựa trên thuật toán hồi quy học tập hợp (Ensemble learning) sử dụng XGBoost và Random Forest. Quy trình đề xuất sử dụng các phương pháp trực quan hóa dữ liệu để hỗ phân tích giúp xác định xu hướng, yếu tố mùa vụ và nhiễu ngẫu nhiên qua đó nhận diện xu hướng dài hạn và các biến động định kỳ của dữ liệu, hiểu được biến động của các biến đặc trưng từ đó giúp hoạch định hướng tiếp cận xây dựng mô hình dự báo. Để giải quyết vấn đề quá khớp chúng tôi sử dụng phương pháp grid search cho mô hình XGBoost và Random Forest để tìm ra bộ siêu tham số cho hai mô hình tương ứng. Nghiên cứu sử dụng bộ dữ liệu tại một khu công nghiệp để huấn luyện, kiểm tra và xây dựng mô hình, chúng tôi còn sử dụng và so sánh hiệu xuất với các phương pháp chuẩn hóa dữ liệu thông dụng nhất hiện nay thể ứng dụng cho cả dữ liệu có tính biến động cao do nhiễu và mất dữ liệu quan sát.

Trong phần này chúng tôi cũng cố gắng chỉ trình bày những nghiên cứu mới nhất có liên quan đến các thuật toán học tập hợp trong mô hình chúng tôi đề xuất nhằm giúp có cái nhìn tổng quan nhất về tình hình và hiệu suất các nghiên cứu dự báo phụ tải sử dụng các thuật toán này. Trong đó kể đến như bài báo của nhóm tác giả Mabrook Al-Rakhami, Abdu Gumaei, Ahmad Alsanad, Atif Alamri, Mohammad Mehedi Hassan[2] năm 2019 hay Farah Mohammad, Kashif Saleem, và Jalal Al-Muhtadi [3] năm 2023; và mới nhất là Harshit Rathore, Hemant Kumar Meena, và Prerna Jain [4] năm 2025. Các tác giả đã xây dựng mô hình dự đoán tải năng lượng bằng cách sử dụng phương pháp học máy Ensemble Learning như thuật toán Extreme Gradient Boosting (XGBoost) hay Random Forest để giải quyết vấn đề quá khớp (overfitting), hay tăng cường các đặc trưng sử dụng các phương pháp tổng hợp như trung bình số học, trung bình điều hòa, trung vị, và trung bình có trọng số. Kết quả cho thấy phương pháp này đạt hiệu suất dự đoán cao nhất và vượt trội hơn so với các phương pháp khác như hồi quy tuyến tính và mạng nơ-ron nhân tạo (ANN).

2. PHƯƠNG PHÁP NGHIÊN CỨU

Trong phần này chúng tôi trình bày chi tiết các lý thuyết của phương pháp nghiên cứu, ứng dụng các kỹ thuật vào trong từng bước trong quá trình thực nghiệm.

2.1 Thành phần chuỗi thời gian

Một chuỗi thời gian [5] có nhiều thành phần chính gồm xu hướng, yếu tố mùa vụ, chu kỳ và thành phần dư. Về mặt toán học sự tương tác của các thành phần của chuỗi thời gian (xu hướng, yếu tố mùa vụ và thành phần sai số) tạo ra hai loại mô hình gồm mô hình cộng và mô hình nhân [6].

Mô hình cộng: Trong mô hình này, chuỗi thời gian có thể biểu diễn toán học bởi công thức sau:

$$x_i = g_i + s_i + w_i \quad (1)$$

Với $x_i|_{i=1\ldots n}$ là chuỗi quan sát, g_i là xu hướng, s_i là yếu tố mùa vụ và w_i là thành phần dao động còn lại. Nếu hiện tượng quan sát này được quan sát vào thời gian thường xuyên, thứ tự giá trị i được quan sát trong thời gian $i \times t_e$, với t_e là chu kỳ lấy mẫu hoặc chu kỳ quan xu hướng. Theo quy ước yêu cầu s_i phải có giá trị trung bình bằng 0 để tách biệt các thành phần xu hướng và yếu tố mùa vụ:

$$\sum_{k=1}^{p} s_{i+k} = 0 \quad (2)$$

Với p là chu kỳ của dao động yếu tố mùa vụ, tương tự cho trung bình của x_i gồm trong xu hướng g_i, chúng tôi cũng thừa nhận rằng w_i có trung bình bằng 0. Trong mô hình này, bộ khuếch đại của s_i và w_i tại một thời điểm nhất định không phụ thuộc vào giá trị của g_i tại cùng một lúc.

Mô hình nhân: Trong mô hình này, chuỗi thời gian có thể biểu diễn toán học bởi công thức (3).

$$\begin{aligned} x_i &= g_i(1+s_i)(1+w_i) \\ x_i &= g_i + g_i s_i + g_i(1+s_i)w_i = g_i + s_i + w_i \end{aligned} \quad (3)$$

Với $x_i|_{i=1\ldots n}$ là chuỗi quan sát, g_i là xu hướng, $1+s_i$ là yếu tố mùa vụ và $1+w_i$ là thành phần dao động còn lại. Cuối cùng hai mô hình có thể kết hợp với nhau tạo mô hình lai hoặc mô hình hỗn hợp. Trong nghiên cứu liên quan có các mô hình lai khác nhau. Công thức 4 biểu diễn mô hình này:

$$x_i = (g_i + s_i)(1+w_i) \quad (4)$$

Mô hình cộng là hữu ích khi biến mùa vụ ổn định theo thời gian, trong khi mô hình nhân bội là hữu ích khi biến mùa vụ là tăng theo thời gian. Các trường hợp khác có thể dùng mô hình lai.

2.2 Phương pháp phát hiện điểm ngoại lai và giá trị bất thường

Chúng tôi sử dụng thuật toán Rừng cô lập (Isolation Forest), một thuật toán học máy không giám sát (unsupervised learning) dùng để phát hiện điểm bất thường (outlier detection) trong dữ liệu. Isolation Forest cô lập các điểm dữ liệu bất thường dựa trên cây quyết định. Thuật toán đánh giá độ bất thường dựa trên độ dài đường đi trung bình để cô lập một điểm. Điểm càng bất thường thì độ dài đường đi càng ngắn.

Chúng tôi xây dựng mô hình sử dụng thuật toán Rừng cô lập để phát hiện điểm ngoại lai với giả định ban đầu rằng 1% dữ liệu là điểm ngoại lai và đảm bảo kết quả có thể tái lập, sau đó áp dụng mô hình Rừng cô lập lên biến mục tiêu (Consumption), quá trình huấn luyện thuật toán sinh

ra một biến mới (outlier) có hai giá trị gồm 1 đối với điểm bình thường và -1 là điểm ngoại lai, sau đó lọc ra các quan sát không phải là điểm ngoại lai và xóa cột outlier.

2.3 Tăng cường đặc trưng

Tạo đặc trưng là một bước cơ bản trong lĩnh vực phân tích dữ liệu bao gồm lựa chọn, thao tác biến đổi dữ liệu thô vào các chức năng có thể được sử dụng trong kỹ thuật học máy.

Đặc trưng thời gian: Gồm các thứ trong tuần(day_of_week), ngày trong năm (day_of_year), tháng (month), quý (quarter) và năm (year) giúp mô hình nắm bắt các mẫu tiêu thụ năng lượng theo mùa vụ và xu hướng các ngày trong tuần, trong tháng do các ngày làm việc thường có mức tiêu thụ năng lượng khác so với các ngày cuối tuần và các yếu tố mùa vụ trong năm và giúp mô hình học được chu kỳ tuần hoặc tháng.

Đặc trưng Trung bình động (MA): Chúng tôi tăng cường đặc trưng Trung bình động 2 ngày, 4 ngày, 7 ngày và 14 ngày giúp làm mượt dữ liệu tăng cường khả năng dự báo của mô hình giúp mô hình nắm bắt các mẫu tiêu thụ điện năng theo thời gian, xu hướng và sự biến động từ đó cải thiện độ chính xác của dự báo thời gian trong đó, đặc trưng trung bình động (moving_avg) để loại bỏ các biến động ngẫu nhiên và nắm bắt xu hướng. Xu hướng được ước lượng bằng trung bình động đối xứng vượt qua chiều dài p : Nếu chu kỳ là lẻ $\left(p = 2q+1\right)$ thì:

$$(g_i)_e = \frac{1}{2q+1}\sum_{j=-q}^{q} x_{i+j} = \frac{1}{p}\sum_{j=-q}^{q} x_{i+j} \qquad (5)$$

nếu chu kỳ là chẵn $\left(p = 2q\right)$ thì

$$(g_i)_e = \frac{1}{p}\left(0.5x_{i-q} + \sum_{j=q+1}^{-q} x_{i+j} + 0.5x_{i+q}\right) \qquad (6)$$

Với $i = q+1, \ldots, n-q$

Đặc trưng Độ trễ (Lag): Ngoài phương pháp làm mượt dữ liệu bằng MA để khai thác tốt các mối liên hệ trong bài toán chuỗi thời gian, tăng cường nhận biết tương quan dữ liệu giữa các khung thời gian khác nhau giúp mô hình nắm bắt các mối quan hệ phụ thuộc giữa giá trị hiện tại và giá trị quá khứ dựa trên các đặc tính phụ thuộc tuần tự những giá trị này được gọi là đặc trưng độ trễ (lag feature), do nắm bắt các mối liên hệ giá trị giữa các khung thời gian quá khứ và hiện tại nên đặc trưng trễ rất hữu ích để nắm bắt các chu kỳ trong chuỗi thời gian.

$$lag_k\left(t\right) = x_{t-k} \text{ , với } k = 1, 2, ..., n \qquad (7)$$

Trong đó $lag_k\left(t\right)$ là giá trị độ trễ bậc k tại thời điểm t với x_t giá trị gốc tại thời điểm t và k là số bước trễ từ 1 đến n.

Đặc trưng Trung bình động theo hàm mũ (EMA): Phương pháp thống kê này sẽ giúp mô hình dự báo phản ứng tốt hơn với những thay đổi ngắn hạn. Dựa trên giá trị trung bình lăn 7 ngày nhưng gán trọng số cao hơn cho các quan sát gần. Trái ngược với tính năng trễ EMA cho phép mô hình hiểu ảnh hưởng trực tiếp của các giá trị tiêu thụ trước đó lên giá trị hiện tại và những hành vi biến đổi chung của chuỗi thời gian giúp xác định xu hướng và biến động mà các tính năng trễ có thể bỏ qua. Ngoài ra đặc trưng trung bình động theo hàm mũ còn giảm thiểu tác động của nhiễu ngẫu nhiên do hiệu ứng làm mịn của chúng. EMA (Exponential Moving Average) được xác định bởi công thức sau:

$$EMA_t = \alpha.X_t + 1 - \alpha.EMA_{t-1} \qquad (8)$$

Trong đó EMA_t là giá trị gốc tại thời điểm *t* với X_t là giá trị độ trễ bậc *k* tại thời điểm *t* và EMA_{t-1} là số bước trễ từ 1 đến *n*. α là hệ số làm mịn, xác định trọng số dành cho quan sát gần nhất. Hệ số này có thể lấy các giá trị trong phạm vi [0, 1].

2.4 Phương pháp chuẩn hóa dữ liệu trong mô hình chuỗi thời gian

Nhằm xử lý các biến đổi đáng kể của độ lớn, giá trị hoặc đơn vị trong bộ dữ liệu chúng tôi kỹ thuật chuẩn hóa normalization [6], trước khi chuẩn hóa dữ liệu tôi thực hiện loại bỏ các giá trị bất thường (outliers), điều này sẽ giúp cho việc chuẩn hóa giúp thang đo ổn định và đúng với dữ liệu thực tế hơn. Normalization Phương pháp chuẩn hóa Min-Max được sử dụng theo công thức sau:

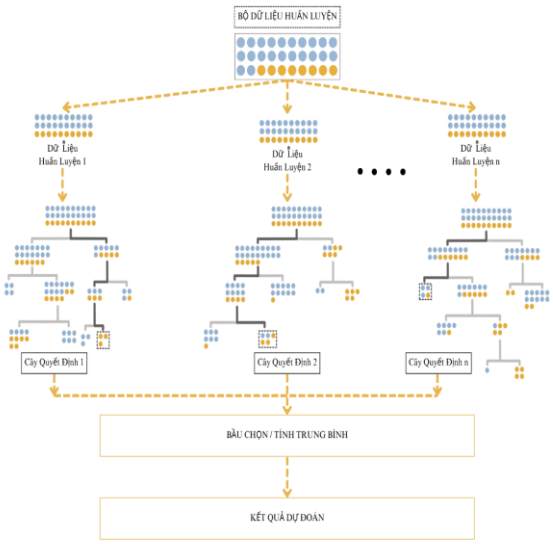
$$X = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \qquad (9)$$

Với $X_{\min}$ và $X_{\max}$ là giá trị quan sát nhỏ nhất và lớn nhất của đặc trưng X và X là giá trị của đặc trưng chúng ta đang cố gắng chuẩn hóa.

2.5 Thuật toán Ensemble Learning sử dụng xây dựng mô hình dự báo

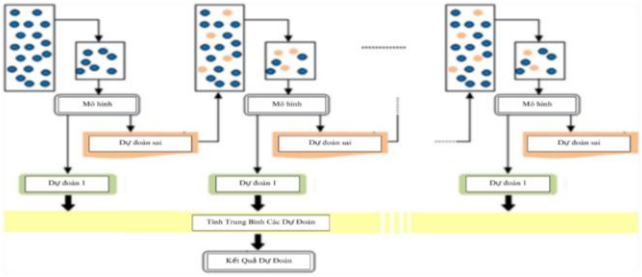
Ensemble Learning là kỹ thuật kết hợp nhiều mô hình học máy để cải thiện độ chính xác và khả năng tổng quát hóa. Trong đó, **Random Forest** là thuật toán thuộc nhóm Bagging, sử dụng nhiều cây quyết định huấn luyện song song, còn **XGBoost** thuộc nhóm Boosting, xây dựng mô hình tuần tự để tối ưu sai số.

Random Forest: Là một thuật toán học máy thuộc kiểu "ensemble learning" dựa trên xây dựng nhiều cây quyết định (decision trees) độc lập [7], kết hợp đầu ra của nhiều cây quyết định để đưa ra một kết quả duy nhất. Mỗi cây đưa ra dự đoán độc lập dựa trên dữ liệu mình đã học, kết quả cuối cùng được xác định qua việc tổng hợp các dự đoán của tất cả các cây. Thuật toán này sử dụng hai yếu tố chính gồm Bagging (Bootstrap Aggregating) tạo ra các mẫu dữ liệu huấn luyện (bootstrap samples) khác nhau từ tập dữ liệu gốc và chọn ngẫu nhiên tập con các đặc trưng để xây dựng từng cây, từ đó giảm thiểu sự tương quan giữa các cây và tăng tính đa dạng. Đối với bài toán phân loại Random Forest sẽ lấy kết quả xuất hiện nhiều nhất dựa trên cơ chế "voting" (bầu chọn), đối với bài toán hồi quy kết quả cuối cùng sẽ được tính bằng trung bình các dự đoán trên cơ chế "averaging" (trung bình) kết quả của nhiều cây (hình). Nhờ vậy Random Forest cho khả năng tổng quát cao và chống overfitting tốt.



Hình 1 *Mô hình Random Forest*

XGBoost : XGBoost là một thành phần cây quyết định tăng cường gradient với cây quyết định được tạo theo tuần tự [8]. Trọng số đóng vai trò quan trọng trong XGBoost, tất cả các biến được đánh trọng số và sau đó đưa vào vây quyết định để dự đoán kết quả. Trọng số của biến dự đoán sai bởi cây sẽ được tăng và các biến này sau đó được đưa vào cây quyết định thứ hai. Kết hợp các phân loại riêng lẻ tạo ra một mô hình mạnh mẽ và chính xác (hình 3.10).

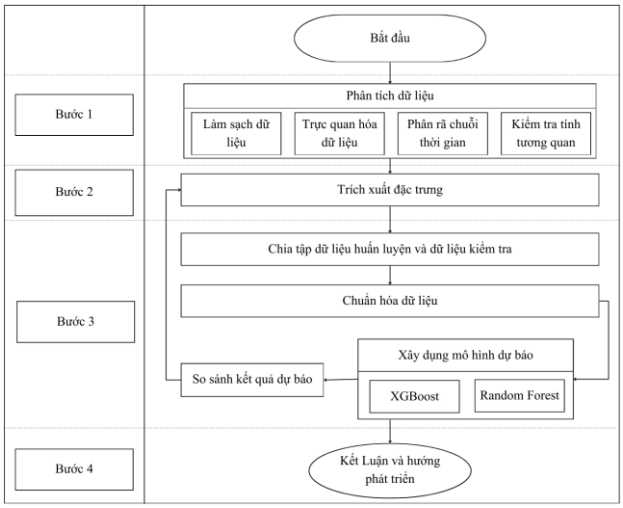


Hình 2 *Mô hình chuỗi thời gian XGBoost*

3. GIẢI PHÁP XÂY DỰNG MÔ HÌNH DỰ BÁO BẰNG PHƯƠNG PHÁP HỌC TẬP HỢP ENSEMBLE LEARNING

Trong chương này chúng tôi đề xuất một quy trình xây dựng mô hình dự báo chuỗi thời gian dự báo phụ tải sử dụng các kỹ thuật, quy trình xây dựng gồm 4 bước: Phân tích dữ liệu, Trích xuất đặc trưng và xây dựng mô hình sử dụng thuật toán XGBoost hoặc Random Forest và phương pháp chuẩn hóa dữ liệu, kết luận và hướng phát triển (hình 3).

Bước 1 Phân tích dữ liệu là bước quan trọng trong nghiên cứu chuỗi thời gian và nó hữu ích để phân tích và tổng hợp các chức năng chính sử dụng trực quan hóa dữ liệu và phương pháp thống kê, các quan sát có thể sử dụng bằng thuật toán học máy. Công việc phân tích dữ liệu gồm các bước làm sạch dữ liệu, phân tích dữ liệu, kiểm tra tự tương quan dữ liệu. **Bước 2** thực hiện tăng cường các đặc trưng thời gian, đặc trưng trung bình động, đặc trưng trung bình động theo hàm mũ và đặc trưng trễ như đã trình bày phần trước. **Bước 3** Xử lý các biến đổi đáng kể của độ lớn, giá trị hoặc đơn vị trong bộ dữ liệu, kỹ thuật chuẩn hóa phổ biến nhất là normalization (Min-Max) được sử dụng trong nghiên cứu trước khi chuẩn hóa dữ liệu chúng tôi sử dụng thuật toán Rừng cô lập để phát hiện điểm ngoại lai (Outlier detection) và thực hiện loại bỏ các giá trị bất thường (outliers), điều này sẽ giúp cho việc chuẩn hóa giúp thang đo ổn định và đúng với dữ liệu thực tế hơn. **Bước 4** đánh giá mô hình kết luận và hướng phát triển trong tương lai.



Hình 3 *Quy trình xây dựng mô hình dự báo phụ tải gồm 4 bước phân tích dữ liệu, trích xuất đặc trưng và chuẩn hóa dữ liệu, so sánh kết quả dự báo và kết luận.*

4. ĐỘ ĐO HIỆU NĂNG ĐÁNH GIÁ MÔ HÌNH

Các chỉ số thống kê hoặc hàm chi phí, biểu diễn các chỉ số các độ đo chính của các mô hình khác nhau chúng cho phép đánh giá hiệu suất của một mô hình thực tế. Trong các chỉ số sử dụng để đánh giá mô hình chuỗi thời gian, RMSE, MAPE và MAE [8] được sử dụng thường xuyên nhất bởi vì chúng không cùng thang đo do đó các giá trị của chúng là không giống nhau. Ngoài ra độ đo hệ số xác định R^2 [9] được sử dụng để đánh giá mức độ phù hợp của mô hình hồi quy.

RMSE là độ lệch chuẩn của phần dư (lỗi dự đoán) cho chúng ta khoảng cách giữa điểm dữ liệu và đường hồi quy. Về mặt toán học công thức RMSE có thể được viết như sau:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i\right)^2} \qquad (10)$$

Với y_i là các giá trị dự đoán, y_i là các quan sát và n là giá trị của các quan sát có sẵn cho phân tích. RMSE là rất thường được sử dụng và nó được xem là một chỉ số lỗi tổng quát xuất sắc cho dự đoán.

MAPE chỉ số thứ hai được sử dụng nhiều nhất để đánh giá mô hình chuỗi thời gian và được biểu diễn toán học bằng công thức sau;

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\left(y_i - y_i\right)}{y_i}\right| \qquad (11)$$

Chỉ số đo này đo lường sự khác biệt của các lỗi dự đoán và sau đó thực hiện phép chia bằng giá trị quan sát thực tế và vì vậy nó không phụ thuộc vào quy mô, tức là

nó có thể được sử dụng để so sánh trên các bộ dữ liệu khác nhau.

MAE đo lường giá trị ước lượng và giá trị dự đoán khác nhau bao nhiêu từ giá trị thực tế. Hầu hết được sử dụng trong chuỗi thời gian, nhưng cũng có thể được áp dụng bất kỳ loại ước lượng thống kê nào. MAE is đơn giản được xác định như sau:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(y_i - y_i)| \quad (12)$$

Với y_i và y_i là giá trị thực tế và giá trị dự đoán.

R-squared (R²) là một thước đo đánh giá mức độ phù hợp của mô hình hồi quy với dữ liệu. Nó được định nghĩa là tỷ lệ phương sai của biến phụ thuộc có thể được giải thích bởi các biến độc lập. Về mặt toán học, R² được biểu diễn như sau:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (13)$$

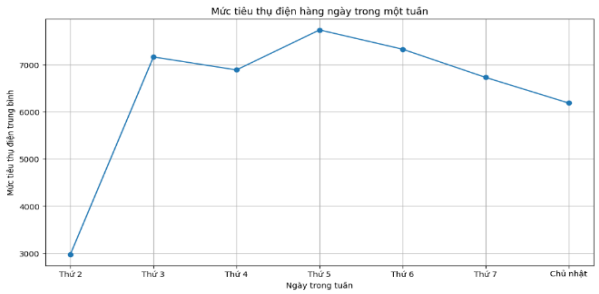
Trong đó SS_{res} là tổng bình phương sai số còn lại (residual sum of squares), SS_{tot} là tổng bình phương tổng thể (total sum of squares). R² càng cao thì mô hình càng phù hợp với dữ liệu.

5. THỰC NGHIỆM

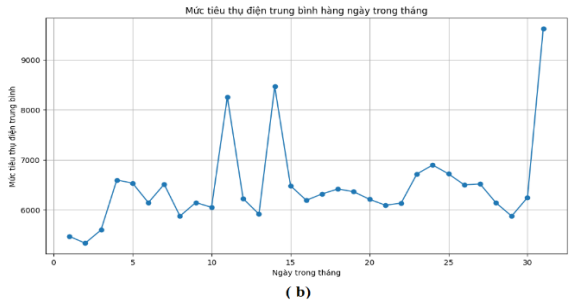
Trong phần này trình quá trình thực nghiệm mô hình dựa trên phương pháp đề xuất cho chuỗi thời gian

5.1 Phân tích chuỗi thời gian

Chúng tôi giả định rằng mô hình có tính cộng và chu kỳ mùa vụ theo năm 365 ngày.

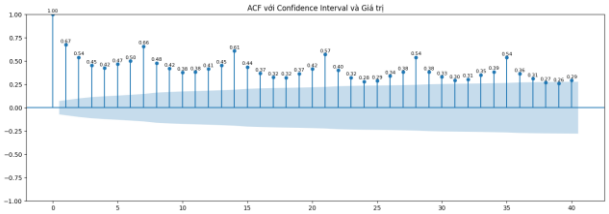


Hình 4 *Biểu đồ mức tiêu thụ điện trung bình theo ngày trong tuần*



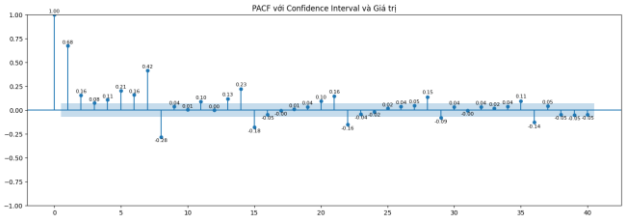
Hình 5 *Biểu đồ mức tiêu thụ điện trung bình theo ngày trong tháng.*

Biểu đồ đặc tính chuỗi thời gian của bộ dữ liệu được biểu diễn ở hình 4 và hình 5 cho thấy dữ liệu còn có xu hướng tăng phi tuyến tính theo thời gian ngắn và tính mùa vụ mạnh mẽ với chu kỳ theo ngày, theo tuần. Phân tích mối quan hệ giữa quan sát hiện tại của chuỗi thời gian và quá khứ là điều quan trọng trong phân tích chuỗi thời gian kỹ thuật này được gọi là lag. Chúng tôi thực hiện điều này bằng cách sử dụng hàm tự tương quan (ACF) của chuỗi tiêu thụ điện.



Hình 6 *Biểu đồ thể hiện mức tương quan trễ 40 ngày của công suất tiêu thụ điện sử dụng hàm ACF*

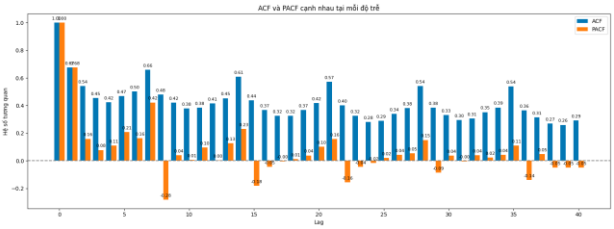
Bộ dữ liệu có tương quan mạnh thời gian ngắn hạn với các độ trễ từ 1 ngày đến 5 ngày. Với khoản dự báo thời gian trung hạn ta thấy lag_7 có chỉ số tương quan rất cao là 0.66 (hình 6). Để hỗ trợ việc xác định sẽ trích xuất đặc trưng lag phù hợp, chúng tôi tiến hành kiểm tra biểu đồ tương quan từng phần (Partial autocorrelation plot) (hình 7) cho thấy mức tương quan cao và liên tục các ngày thứ 1 đến ngày thứ 7 và ngày thứ 14, trong khi các ngày 21 và 28 cũng cho mức tương quan khá tốt nhưng không liên tục.



Hình 7 *Biểu đồ thể hiện mức tương quan trễ 40 ngày của công suất tiêu thụ điện sử dụng hàm PACF*

5.2 Trích xuất đặc trưng

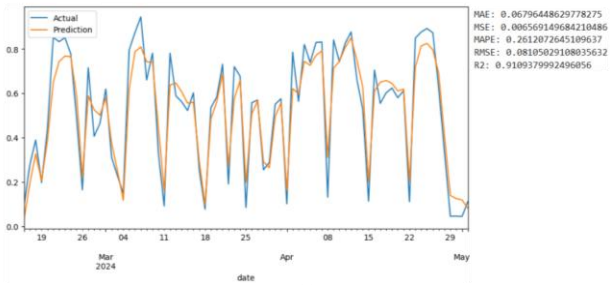
Trích xuất đặc trưng đóng vai trò quan trọng trong việc xây dựng mô hình dữ liệu, nó xác định các đặc trưng thực tế của dữ liệu để tạo thuật toán ML có chính xác cao.



Hình 8 *Biểu đồ kết hợp giá trị tương quan công suất tiêu thụ điện sử dụng hàm ACF và PACF*

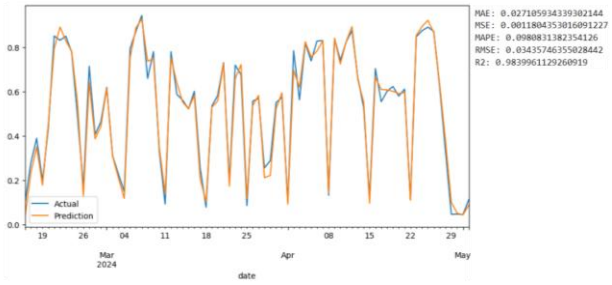
Từ phân tích mục 5.1 và biểu đồ hình 7 cho thấy từ độ trễ 6 ngày bắt đầu có dao động và không còn ổn định do vậy chúng tôi quyết định tăng cường các đặc trưng trễ từ 1 ngày đến 6 ngày, các đặc trưng trung bình động và trung bình động theo hàm mũ 2 ngày, 4 ngày, 7 ngày và 14 ngày và các đặc trưng thời gian đã đề cập ở mục 2.2.

5.3 Chuẩn hóa dữ liệu



Hình 9 *Kết quả mô hình dự báo sử dụng thuật toán Random Forest, chuẩn hóa dữ liệu bằng phương pháp Normalization*

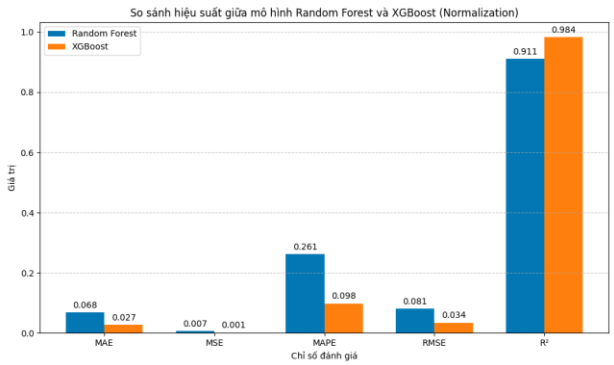
Các tác động của yếu tố ngoại lai tạo ra thành phần dữ trong chuỗi dữ liệu gây nhiễu ảnh hưởng đến hiệu suất dự báo mô hình do đó chúng tôi sử dụng phương pháp chuẩn hóa Normalization (sử dụng Min-Max scaling) đưa dữ liệu về khoảng [0, 1] để đưa các đặc trưng về cùng một thang đo giúp cả hai mô hình sử dụng thuật toán XGBoost (hình 9) và Random Forest (hình 8) xử lý dữ liệu hiệu quả hơn nhờ giữ nguyên mối quan hệ tỷ lệ giữa các giá trị, đồng thời đảm bảo các đặc trưng đầu vào có quy mô đồng đều. Chuẩn hóa dữ liệu còn giúp loại bỏ các yếu tố ngoại lai gây nhiễu, giúp dữ liệu ổn định và mô hình nắm bắt mối quan hệ các biến tốt hơn, dự báo chính xác hơn rất nhiều. Phương pháp chuẩn hóa Normalization và thực hiện tính toán kỹ lưỡng để giúp mô hình có thể dự báo chính xác mà không bị rò rỉ dữ liệu xảy ra hiện tượng overfitting. Chúng tôi chỉ fit scaler trên tập huấn luyện (X_train, y_train) và dùng scaler này để transform tập kiểm tra (X_test, y_test).



Hình 10 *Kết quả mô hình dự báo sử dụng thuật toán XGBoost, chuẩn hóa dữ liệu bằng phương pháp Normalization*

6. KẾT QUẢ MÔ HÌNH DỰ BÁO VÀ ĐÁNH GIÁ

Dựa theo giải pháp xây dựng mô hình dự báo đề xuất từ nghiên cứu, chúng tôi đã xây dựng đồng thời hai mô hình dự báo sử dụng hai thuật toán XGBoost, Random Forest với 20 biến đặc trưng được bổ sung. Nghiên cứu thực hiện so sánh 2 mô hình dự báo này và các nghiên cứu liên quan đã trình bày ở phần trước.



Hình 11 *So sánh hiệu suất hai mô hình dự báo sử dụng thuật toán XGBoost và Random Forest sử dụng phương pháp chuẩn hóa Normalization*

Với phương pháp chuẩn hóa Normalization làm cho quá trình học của mô hình diễn ra ổn định hơn đặc biệt quan trọng với các thuật toán boosting như XGBoost, vốn nhạy cảm với độ lớn của giá trị đầu vào. Cho thấy mô hình với thuật toán XGBoost vượt trội hơn so với mô hình sử dụng Random Forest (hình 10), trong đó chỉ số độ đo MAE ở mô hình XGBoost 0.027, mô hình Random Forest 0.068 trong khi chỉ số R^2 ở mô hình XGBoost là 98% so với 91% ở mô hình Random Forest.

Bên cạnh đó mô hình đề xuất trong nghiên cứu này có thể cho ra kết quả tốt hơn ở nhiều độ đo khác nhau khi so sánh với các mô hình đề xuất bởi nhóm tác giả đến từ đại học King Saud về phát hiện tổn thất điện do trộm cắp năng lượng năm 2023 sử dụng học máy Ensemble Learning đạt độ chính xác 94.75% hay nghiên cứu từ nhóm tác giả đến từ Brazil Dự đoán sản lượng điện gió đăng trên tạp chí (Social Science Research Network) năm 2024 sử dụng Ensemble Learning đạt 87.5% hoặc nhóm tác giả đến từ trường Đại học Leeds Beckett dự đoán nhu cầu năng lượng của các hộ gia đình tại London công bố trên tạp chí Sustainability (MDPI) năm 2025 sử dụng Prophet, eXtreme Gradient Boosting (XGBoost) đạt 81.48%.

Qua quá trình thực nghiệm cho thấy Mặc dù bộ dữ liệu không có tính chu kỳ rõ ràng và biến động phức tạp, thuật toán XGBoost vẫn cho thấy hiệu quả cao nhờ khả năng kết hợp nhiều cây quyết định độc lập. Cơ chế học tăng cường giúp mô hình liên tục cải thiện độ chính xác bằng cách giảm thiểu sai số từ các lần dự báo trước. Nhờ đó, XGBoost có thể xử lý tốt các mối quan hệ phi tuyến giữa biến đầu vào và biến mục tiêu, đồng thời vẫn giải thích được các tương tác tuyến tính tiềm ẩn trong dữ liệu, đặc biệt khi chuỗi thời gian được chuẩn hóa min-max đã giúp mô hình đáp ứng yêu cầu độ sai lệch dự báo phụ tải điện mà còn mang đến một công cụ dự báo lĩnh vực năng lượng khác như nước, gas hày cho các lĩnh vực khác trong đời sống như dự báo tài chính-chứng khoán hay dự báo sản xuất công nghiệp và thương mại điện tử.

7. TÀI LIỆU THAM KHẢO

[1] Bộ Công Thương, Cục Điều Tiết Điện Lực, "Ban hành Quy trình dự báo nhu cầu phụ tải điện hệ thống điện quốc gia" trong Quyết định số 7/ QĐ-ĐTĐL, **2013**.

[2] Mabrook Al-Rakhami, Abdu Gumaei, Ahmed Alsanad, Atif Alamri, and Mohammad Mehedi Hassan, "An Ensemble Learning Approach for Accurate Energy Load Prediction in Residential Buildings," IEEE Access, vol. 7, p. 48328–48338, **2019**.

[3] Harshit Rathore, Hemant Kumar Meena, và Prerna Jain, "Prediction of EV Energy consumption Using Random Forest And XGBoost," 2023 International Conference on Power Electronics and Energy (ICPEE), **2023**.

[4] P. Trebuna, J. Halcinová, M. Filo, and J. Markovic, "The importance of normalization and standardization in the process of clustering" trong IEEE 12th International Symposium on Applied Machine Intelligence and Informatics (SAMI), Herl'any, Slovakia, **2014**.

[5] E. B. Dagum, S. Bianconcini, Seasonal Adjustment Methods and Real Time Trend-cycle Estimation, Cham, Switzerland: Springer, **2016**.

[6] Debojyoti Chakraborty, Jayeeta Mondal, Hrishav Bakul Barua, and Ankur Bhattacharjee, "Computational Solar Energy - Ensemble Learning Methods for Prediction of Solar Power Generation based on Meteorological Parameters in Eastern India" Here's a suggested citation for the paper in IEEE format, tập 44, pp. 277-294, **2023**.

[7] David S. Moore, George P. McCabe, Bruce A. Craig, Introduction to the Practice of Statistics, W. H. Freeman, **2017**.

[8] Niraj Buyo, Akbar Sheikh-Akbari, và Farrukh Saleem, "An Ensemble Approach to Predict a Sustainable Energy Plan for London Households, " Sustainability, vol. 17, no. 2, p. 500, **2025**.

[9] Samara Deon, José Donizetti de Lima, Geremi Gilson Dranka, và Matheus Henrique Dal Molin Ribeiro, "in New Trends in Disruptive Technologies, Tech Ethics, and Artificial Intelligence, " Springer Nature Switzerland, **2024**, p. 15–27.