PHÂN TÍCH ẢNH HƯỞNG CÁC SIÊU THAM SỐ CỦA MÔ HÌNH LIGHTGBM ĐẾN DỰ BÁO CÔNG SUẤT ĐIỆN MẶT TRỜI

ANALYSIS OF THE IMPACT OF HYPERPARAMETERS OF LIGHTGBM MODEL ON SOLAR POWER FORECASTING

Phạm Mạnh Hải[1*], Nguyễn Tuấn Anh[1], Vũ Minh Pháp[1,2], Nguyễn Ngọc Trung[1], Vũ Thị Anh Thơ[1], Nguyễn Hữu Nguyện[3], Đỗ Quang Hiệp[4], Nguyễn Đức Quang[1]

[1]Trường Đại học Điện lực, [2]Viện Khoa học Công nghệ Năng lượng và Môi trường - Viện Hàn lâm Khoa học và Công nghệ Việt Nam, [3]Trường Đại học Công nghệ Đông Á, [4]Trường Đại học Kinh tế - Kỹ thuật Công nghiệp

[*]Tác giả liên hệ

Ngày nhận bài: 07/02/2025, ngày chấp nhận đăng: 20/04/2025. Phản biện: PGS.TS. Đỗ Như Ý

Tóm tắt:

Bài báo trình bày nghiên cứu về ảnh hưởng của một số siêu tham số trong mô hình LightGBM đến độ chính xác dự báo công suất phát điện mặt trời. Các siêu tham số được xem xét bao gồm số lá tối đa của cây quyết định (num_leaves), tốc độ học (learning_rate) và số lượng cây học (n_estimators). Mười kịch bản với các tổ hợp siêu tham số khác nhau đã được thực hiện và so sánh dựa trên các chỉ số sai số: RMSE, MAPE, NMAPE, cũng như thời gian huấn luyện và dự báo. Kết quả cho thấy, việc điều chỉnh các tham số này có cải thiện hiệu suất dự báo của mô hình, thể hiện qua giảm nhẹ các sai số dự báo, ví dụ MAPE giảm từ 90,67% xuống còn 82,94% khi tăng num_leaves từ 30 lên 60. Tuy nhiên, mức cải thiện không đáng kể, các chỉ số sai số chỉ thay đổi trong biên độ nhỏ giữa các kịch bản. Điều này cho thấy, mô hình LightGBM khá bền vững với các siêu tham số trong phạm vi thử nghiệm, và việc tinh chỉnh vừa phải các giá trị num_leaves, learning_rate, n_estimators không đem lại thay đổi đột biến về độ chính xác dự báo.

Từ khóa: LightGBM, dự báo năng lượng mặt trời, siêu tham số, num_leaves, learning_rate, n_estimators, hiệu suất mô hình.

Abstract:

This paper presents a study on the impact of certain hyperparameters in the LightGBM model on solar power generation forecasting accuracy. The considered hyperparameters include the maximum number of leaves in decision trees (num_leaves), learning rate (learning_rate), and the number of boosting rounds (n_estimators). Ten scenarios with different combinations of these hyperparameters were implemented and compared based on error metrics: RMSE, MAPE, and NMAPE, as well as training and inference time. The results show that adjusting these parameters could improve the forecasting performance of the model, as reflected in a slight reduction in forecasting errors for instance, the MAPE decreased from 90.67% to 82.94% when increasing num_leaves from 30 to 60. However, the improvements are insignificant, the error metrics only vary within a narrow range across scenarios. This indicates that the LightGBM model is relatively robust to changes in hyperparameters within the tested range, and moderate tuning of num_leaves, learning_rate, and n_estimators does not lead to dramatic changes in forecasting accuracy.

Keywords: LightGBM algorithm, photovoltaic power prediction, model hyperparameters, num_leaves, learning_rate, n_estimators, predictive performance.

KÝ HIỆU:

RMSE: sai số trung bình bình phương

MAPE: sai số phần trăm tuyệt đối trung bình

NMAPE: sai số phần trăm đã chuẩn hóa

LightGBM: thuật toán học máy Light Gradient Boosting Machine

CatBoost: thuật toán Categorical Boosting

KNN: thuật toán K-Nearest Neighbors

SHAP: phương pháp SHapley Additive exPlanations

SVR: Hồi quy vector hỗ trợ

1. GIỚI THIỆU CHUNG

Năng lượng mặt trời là nguồn năng lượng tái tạo quan trọng, nhưng công suất phát điện mặt trời biến động mạnh do phụ thuộc thời tiết [1]. Do đó, dự báo công suất điện mặt trời chính xác đóng vai trò then chốt trong vận hành hệ thống điện thông minh và ổn định lưới điện [2]. Những năm gần đây, các phương pháp học máy (machine learning) đã được áp dụng rộng rãi cho bài toán dự báo năng lượng mặt trời nhờ khả năng mô hình hóa các quan hệ phi tuyến giữa các biến đầu vào và sản lượng điện. Đặc biệt, các mô hình ensemble như rừng ngẫu nhiên (Random Forest) và thuật toán gradient boosting cho thấy hiệu quả cao trong dự báo năng lượng tái tạo [3]precise forecasting of Solar Irradiance (SI. LightGBM là một thuật toán gradient boosting trên cây quyết định do Microsoft phát triển [4], nổi bật nhờ tốc độ huấn luyện nhanh và hiệu quả cao so với các thư viện boosting trước đó. LightGBM sử dụng chiến lược tăng trưởng cây theo lá (leaf-wise) thay vì theo độ sâu, giúp giảm thời gian huấn luyện nhưng có nguy cơ quá khớp (overfitting) nếu không điều chỉnh tham số phù hợp. Nhiều nghiên cứu đã áp dụng LightGBM trong dự báo phụ tải và năng lượng, cho kết quả khả quan [5]. Ví dụ Hanif và cộng sự cho thấy LightGBM là mô hình mạnh trong đánh giá ảnh hưởng các yếu tố môi trường đến bức xạ mặt trời, vượt trội hơn mô hình SVR trong thí nghiệm của họ [3]. Tại Việt Nam nhóm nghiên cứu của Nguyễn Hữu Nam đã so sánh hiệu suất của các thuật toán như LightGBM, CatBoost, và KNN, đồng thời sử dụng SHAP để xác định độ quan trọng của các yếu tố đầu vào, cho thấy, nhiệt độ và độ ẩm có vai trò quyết định trong dự báo công suất [5]. Bên cạnh đó, Nguyễn Khánh Toàn cũng chỉ ra việc sử dụng giá trị mặc định của các siêu tham số có thể gây sai lệch lớn trong dự báo phụ tải, do đó, cần thiết phải phân tích ảnh hưởng của chúng đến hiệu suất mô hình [6]. Tuy nhiên, hiệu năng của LightGBM phụ thuộc vào việc lựa chọn bộ tham số siêu (hyperparameters) thích hợp. Các tham số quan trọng nhất trong LightGBM bao gồm: số lá cây quyết định (num_leaves), tốc độ học (learning_rate) và số lượng cây (vòng lặp boosting- n_estimators). Việc tinh chỉnh các tham số này có thể ảnh hưởng lớn đến độ chính xác của mô hình; sử dụng các giá trị mặc định có thể dẫn đến sai số dự báo lớn trong một số trường hợp. Do đó, nghiên cứu ảnh hưởng của các siêu tham số tới kết quả dự báo là cần thiết nhằm tối ưu hóa mô hình [6].

Trong bài báo này, nhóm tác giả thực hiện phân tích định lượng tác động của số lá cây

quyết định, tốc độ học và số lượng vòng lặp boosting đến chất lượng dự báo công suất của mô hình LightGBM. Mục tiêu là đánh giá mức độ cải thiện hiệu suất khi thay đổi các tham số này trong một phạm vi nhất định, qua đó xác định liệu việc tinh chỉnh có thực sự đem lại hiệu quả đáng kể hay không. Tuy đã có một số nghiên cứu phân tích ảnh hưởng siêu tham số [5] [6], nhưng đa phần tập trung vào mô hình dự báo phụ tải hoặc đánh giá từng tham số riêng lẻ. Ít nghiên cứu thực hiện đánh giá có hệ thống tác động phối hợp của nhóm siêu tham số chính lên bài toán dự báo công suất phát điện mặt trời tại Việt Nam. Đây là điểm mới mà bài báo hướng đến. Nội dung bài viết được cấu trúc như sau: Phần 1 giới thiệu về ảnh hưởng của các siêu tham số trong mô hình LightGBM trong dự báo công suất phát điện mặt trời. Phần 2 mô tả phương pháp nghiên cứu, bao gồm mô hình LightGBM, các tham số siêu và bộ chỉ số đánh giá. Phần 3 trình bày thiết kế thực nghiệm và kết quả thu được từ 10 kịch bản tham số khác nhau, kèm theo phân tích chi tiết. Phần 4 đưa ra kết luận về ảnh hưởng của các tham số siêu đối với mô hình LightGBM trong bài toán dự báo công suất điện mặt trời.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Mô hình LightGBM và các siêu tham số chính

LightGBM là mô hình học máy thuộc nhóm gradient boosting, kết hợp nhiều cây quyết định để cải thiện dần độ chính xác dự báo [4]. Mỗi cây mới được xây dựng trên phần sai số còn lại của mô hình hiện tại, với trọng số học được điều chỉnh bởi tốc độ học. Nhờ chiến lược xây dựng cây theo lá, LightGBM đạt tốc độ huấn luyện và dự báo nhanh, đặc biệt trên các tập dữ liệu lớn, đồng thời duy trì được độ chính xác cao. Trong mô hình LightGBM, có ba tham số siêu quan trọng ảnh hưởng trực tiếp đến cấu trúc mô hình và khả năng học của thuật toán:

• **num_leaves**: Trong LightGBM, một trong những siêu tham số quan trọng nhất là num_leaves, đại diện cho số lượng lá tối đa mà mỗi cây quyết định trong mô hình có thể đạt được. Tham số này ảnh hưởng trực tiếp đến độ phức tạp của cây: số lá càng lớn, cây càng có khả năng biểu diễn các mối quan hệ phi tuyến phức tạp hơn trong dữ liệu. Tuy nhiên, nếu num_leaves được đặt quá cao so với quy mô và tính đa dạng của tập dữ liệu, mô hình có thể ghi nhớ quá chi tiết đặc điểm của dữ liệu huấn luyện, dẫn đến hiện tượng quá khớp và giảm hiệu quả tổng quát hóa trên dữ liệu mới [6]. Do đó, lựa chọn giá trị num_leaves phù hợp là yếu tố then chốt giúp cân bằng giữa độ chính xác và độ đơn giản của mô hình. Trong thực tiễn, người dùng thường xác định num_leaves dựa trên kinh nghiệm, thử nghiệm lặp lại hoặc sử dụng kỹ thuật tối ưu hóa siêu tham số để tìm được giá trị tốt nhất trong phạm vi cho phép.

• **learning_rate**: Tốc độ học (hệ số bước của thuật toán boosting). Learning rate quyết định mức độ điều chỉnh mô hình ở mỗi vòng boosting: learning rate nhỏ giúp mô hình học dần dần và có thể đạt độ chính xác cao hơn, nhưng cần số vòng lặp (cây) nhiều hơn; ngược lại, learning rate lớn giúp hội tụ nhanh nhưng dễ bỏ qua các mẫu phức tạp, có thể dẫn đến sai số lớn hơn. Thông thường có quan hệ bù

trừ: giảm learning rate đồng thời phải tăng số lượng cây để duy trì khả năng học [7].

• **n_estimators**: Số lượng cây quyết định (số vòng lặp boosting) trong mô hình. Tham số này quy định mô hình gồm bao nhiêu cây được huấn luyện nối tiếp. Số cây quá ít có thể khiến mô hình chưa học đủ (underfitting), trong khi quá nhiều cây có thể gây quá khớp nếu learning rate không được giảm đủ thấp. Thông thường, người ta kết hợp điều chỉnh n_estimators và learning_rate đồng thời để đạt độ chính xác cao trong thời gian huấn luyện hợp lý.

Ngoài ra, LightGBM còn nhiều tham số siêu khác (ví dụ: min_data_in_leaf, max_depth, feature_fraction…) cũng ảnh hưởng đến quá trình huấn luyện. Tuy nhiên trong phạm vi nghiên cứu này, để đảm bảo tính tập trung và giới hạn phạm vi thử nghiệm, chúng tôi chỉ tập trung vào ba siêu tham số được đánh giá là có ảnh hưởng lớn nhất đến hiệu năng mô hình. Các tham số phụ được cố định ở giá trị mặc định. Việc mở rộng phân tích các tham số này sẽ là định hướng trong các nghiên cứu tiếp theo.

2.2. Thiết kế thực nghiệm và bộ dữ liệu

Để phân tích ảnh hưởng của các siêu tham số, chúng tôi sử dụng bộ dữ liệu thực tế từ một nhà máy điện mặt trời tại tỉnh Thanh Hóa với công suất lắp đặt 30MW. Dữ liệu huấn luyện được thu thập trong khoảng thời gian từ ngày 01/01/2024 đến 30/12/2024, bao gồm công suất phát điện thực tế theo thời gian cùng các thông tin thời tiết như bức xạ mặt trời, nhiệt độ không khí và tháng trong năm. Bộ dữ liệu này được chia thành hai phần: tập huấn luyện (80%) dùng để xây dựng mô hình và tập kiểm tra (20%) dùng để đánh giá hiệu suất mô hình sau huấn luyện. Ngoài ra, mô hình còn được áp dụng để dự báo trên một tập dữ liệu được tách biệt khỏi tập huấn luyện và tập kiểm tra nhằm đảm bảo tính khách quan, tập dữ liệu dự báo bao gồm 24 ngày được chọn ngẫu nhiên (mỗi tháng lấy 2 ngày liên tiếp) trong năm 2024, mục đích của việc này là nhằm đánh giá khả năng tổng quát hóa của mô hình LightGBM trên dữ liệu chưa được huấn luyện.

Mô hình LightGBM được huấn luyện trên tập huấn luyện với một tổ hợp tham số siêu nhất định. Sau đó, ta ghi nhận các chỉ số sai số trên tập kiểm tra và trên giai đoạn dự báo tương lai. Trong nghiên cứu này, chúng tôi xác định 10 kịch bản siêu tham số khác nhau như sau:

• Kịch bản 1 (S1) đóng vai trò mốc tham chiếu, sử dụng các giá trị tương đối cơ bản: num_leaves= 30, learning_rate= 0,05; n_estimators=100. Từ đó, các kịch bản tiếp theo thay đổi lần lượt từng tham số hoặc kết hợp để quan sát xu hướng kết quả.

• S2, S3 tăng dần num_leaves (60 và 90) so với S1 (giữ nguyên learning_rate= 0,05; n_estimators=100).

• S4 tăng nhẹ n_estimators lên 150 (và num_leaves=120 trung bình).

• S5, S6, S7 tiếp tục tăng num_leaves (150, 180, 210) và cố định n_estimators= 200 (cao hơn S1) nhằm đánh giá ảnh hưởng khi mô hình phức tạp dần.

• S8 thử giảm mạnh learning_rate xuống 0,01 đồng thời tăng nhiều n_estimators (500)

và num_leaves (240), để xem khả năng cải thiện khi mô hình học chậm hơn nhưng lâu hơn.

• S9 thử tăng learning_rate lên 0,1 (cao hơn mặc định) và giảm số cây (100) với num_leaves khá cao (270), để kiểm tra trường hợp học nhanh.

• S10 sử dụng learning_rate=0,07, num_leaves=300, n_estimators=300 như một cấu hình kết hợp tương đối lớn của cả ba tham số.

Mười kịch bản này được thiết kế nhằm bao quát các tổ hợp đại diện cho từng xu hướng: tăng độ phức tạp dần, học nhanh, học chậm, và kết hợp nhiều yếu tố. Số lượng kịch bản được chọn dựa trên giới hạn tính toán thực tế và mức độ đại diện cần thiết để đánh giá xu hướng.

Tất cả các mô hình đều được huấn luyện trên cùng một tập dữ liệu và được đánh giá trên cùng tập dự báo để đảm bảo tính công bằng khi so sánh.

2.3. Các chỉ số đánh giá

Hiệu suất mô hình được đánh giá bằng các chỉ số sai số phổ biến trong dự báo thời gian thực: RMSE, MAPE và NMAPE. Cụ thể:

RMSE (Root Mean Square Error) là sai số trung bình bình phương, nhấn mạnh các sai số lớn do lấy bình phương trước khi trung bình. Công thức tính RMSE như sau [8]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (1)$$

Trong đó: $\hat{y}_i$ là công suất dự báo (kW), y_i là công suất thực tế (kW), n là số lượng điểm dữ liệu.

MAPE (Mean Absolute Percentage Error): sai số tuyệt đối trung bình phần trăm, thể hiện sai số trung bình tương đối so với giá trị thực (%). Công thức tính MAPE như sau [9]:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \qquad (2)$$

Trong đó: MAPE là sai số tuyệt đối phần trăm trung bình %, giá trị dự báo của công suất phát dự báo thứ i (kW), là giá trị công suất trong thực tế thứ i (kW), n là số lượng điểm dữ liệu.

MAPE cho biết dự báo sai lệch bao nhiêu phần trăm so với thực tế, nhưng có nhược điểm là không xác định khi và dễ bị ảnh hưởng lớn khi rất nhỏ.

NMAPE (Normalized MAPE – MAPE được chuẩn hóa): để khắc phục hạn chế của MAPE tại điểm dữ liệu gần 0, ta chuẩn hóa sai số tuyệt đối so với một giá trị đặc trưng (thường là công suất định mức hoặc giá trị lớn nhất của công suất thực tế). Trong bài báo, NMAPE được tính bằng cách chia cho công suất định mức của hệ thống rồi nhân 100%. Công thức tính NMAPE như sau [10]:

$$\mathrm{NMAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{P_{đm}}\right| \qquad (3)$$

NMAPE là sai số tuyệt đối phần trăm trung bình đã được chuẩn hóa, giá trị dự báo của công suất phát dự báo thứ i (kW), là giá trị công suất định mức lắp đặt của nhà máy (kW), n là số lượng điểm dữ liệu.

Chỉ số NMAPE cho biết sai số trung bình chiếm bao nhiêu phần trăm công suất định mức, giúp đánh giá trực quan mức độ sai số độc lập với quy mô hệ thống.

Các chỉ số trên được tính cho giai đoạn dự báo của từng kịch bản mô hình. Ngoài ra, thời gian thực thi được đo gồm thời gian huấn luyện mô hình và thời gian dự báo cho mỗi kịch bản, nhằm xem xét khía cạnh chi phí tính toán.

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Kết quả

Sau khi huấn luyện và đánh giá 10 kịch bản mô hình LightGBM với các tham số siêu khác nhau, chúng tôi thu được kết quả tổng hợp như trong Bảng 1. Bảng này liệt kê các siêu tham số (num_leaves, learning_rate, n_estimators) của từng kịch bản (S1 đến S10), kèm theo thời gian huấn luyện, thời gian dự báo và các chỉ số RMSE, MAPE, NMAPE tương ứng dự báo.

Kết quả thực nghiệm tại bảng 1 cho thấy mối quan hệ rõ rệt giữa số lá tối đa (num_leaves) và các sai số dự báo như RMSE, MAPE và NMAPE. Khi num_leaves tăng từ 30 đến khoảng 150, sai số có xu hướng giảm, đặc biệt là MAPE giảm từ 90,67% (S1) xuống còn 82,94% (S2), cho thấy mô hình được cải thiện đáng kể khi tăng độ phức tạp ở mức vừa phải.

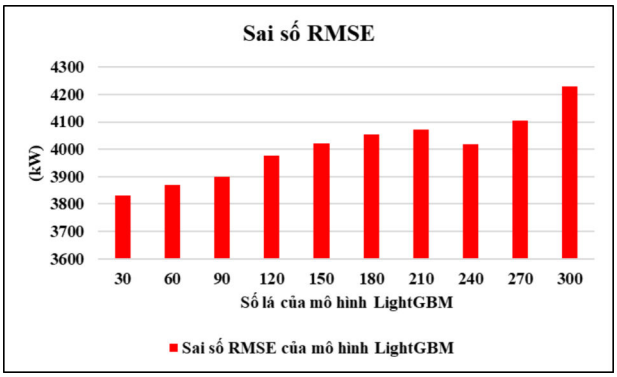
Tuy nhiên, từ mức num_leaves lớn hơn 180, cả RMSE và MAPE bắt đầu dao động và tăng nhẹ trở lại, trong khi NMAPE giữ mức dao động nhỏ từ 9,57% đến 10,36%. Điều này cho thấy mô hình bắt đầu có dấu hiệu quá khớp, và việc tăng số lá quá cao không còn đem lại hiệu quả mà còn làm sai số tăng lên. Trong khi đó, việc điều chỉnh learning_rate cho thấy hiệu quả không rõ rệt: giảm quá thấp (0,01) khiến thời gian huấn luyện tăng đáng kể mà sai số không được cải thiện tương ứng, trong khi tăng quá cao (0,1) có thể khiến sai số lớn do học quá nhanh.

Bảng 1. Kết quả dự báo của mô hình LightGBM với 10 kịch bản tham số siêu khác nhau

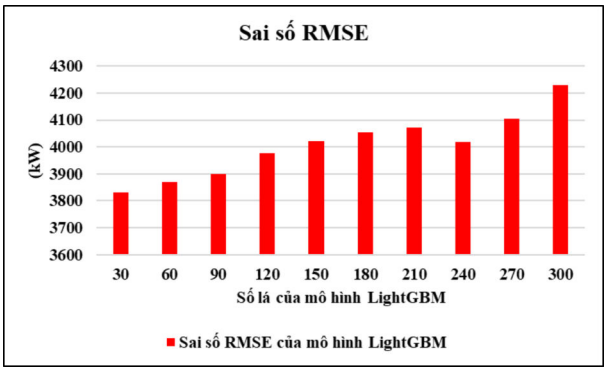
Kịch bản	Số lá tối đa (num-leaves)	Tốc độ học (learning-Rate	Số lượng cây học (n_estimators)	Thời gian huấn luyện (s)	Thời gian dự báo (s)	RMSE (kW)	MAPE (%)	NMAPE (%)
S1	30	0,05	100	0,23	0,011	3830,47	90,67	9,57
S2	60	0,05	100	0,46	0,006	3869,26	82,94	9,61
S3	90	0,05	100	0,57	0,007	3900,765	85,48	9,68
S4	120	0,05	150	0,98	0,014	3975,42	85,04	9,83
S5	150	0,05	200	1,53	0,041	4020,084	85,17	9,91
S6	180	0,05	200	1,38	0,023	4054,824	87,62	9,99
S7	210	0,05	200	1,56	0,024	4071,2	85,68	10,01
S8	240	0,01	500	4,65	0,051	4018,344	86,46	9,90
S9	270	0,1	100	1,01	0,02	4105,243	87,03	10,10
S10	300	0,07	300	3,14	0,05	4230,925	84,82	10,36

Tương tự, tăng số lượng cây học (n_estimators) từ 100 lên 300 không mang lại cải thiện đáng kể về độ chính xác, nhưng làm tăng chi phí tính toán. Vì vậy, có thể kết luận, việc tăng num_leaves giúp cải thiện hiệu suất dự báo trong một giới hạn nhất định, nhưng sau một ngưỡng (~150–180 lá), hiệu quả bắt đầu bão hòa hoặc suy giảm. Do đó, nên lựa chọn tổ hợp tham số vừa phải, đặc biệt là giới hạn num_leaves, để đảm bảo sự cân bằng giữa độ chính xác và chi phí huấn luyện của mô hình.

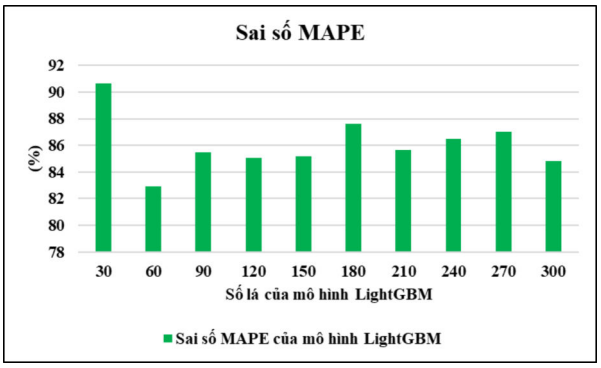
Để trực quan hóa xu hướng biến đổi của các sai số theo số lá tối đa, Hình 1 trình bày ba biểu đồ mô tả mối quan hệ giữa số lá của mô hình LightGBM với các chỉ số sai số RMSE, MAPE và NMAPE:



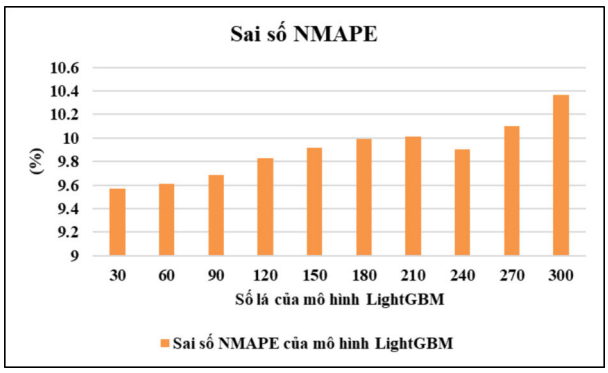
Hình 1a



Hình 1b



Hình 1c



Hình 1d

Hình 1. Các sai số RMSE, MAPE và NMAPE của mô hình LightGBM

Ba biểu đồ trong Hình 1 thể hiện rõ xu hướng biến động của các chỉ số sai số theo sự thay đổi của số lá (num_leaves) trong mô hình LightGBM. Biểu đồ RMSE cho thấy sai số tăng dần khi số lá tăng từ 30 đến 300. Mặc dù RMSE ban đầu khá thấp tại num_leaves = 30 (≈3830 kW), giá trị này liên tục tăng và đạt đỉnh tại num_leaves = 300 (≈4230 kW). Điều này cho thấy, khi mô hình trở nên quá phức tạp, sai số tổng thể tăng do hiện tượng quá khớp. Biểu đồ MAPE có xu hướng biến động không đều. Sau khi giảm mạnh từ 90,67% (30 lá) xuống 82,94% (60 lá), MAPE dao động quanh mức 84–87% cho các kịch bản còn lại. Điều này phản ánh rằng, mức cải

thiện chính diễn ra ở giai đoạn đầu, và việc tăng thêm độ phức tạp mô hình sau đó không mang lại lợi ích rõ ràng. Biểu đồ NMAPE thể hiện độ ổn định cao hơn. Chỉ số này dao động nhẹ từ 9,57% đến 10,36%, và cũng đạt giá trị thấp nhất tại num_leaves = 30, sau đó tăng dần, đặc biệt từ num_leaves ≥ 180. Mặc dù biên dao động nhỏ, xu hướng chung là sai số chuẩn hóa tăng theo độ phức tạp mô hình. Tổng hợp ba biểu đồ cho thấy, mô hình đạt hiệu suất tối ưu trong khoảng số lá từ 60 đến 150. Việc tăng số lá vượt mức này không chỉ không cải thiện đáng kể sai số, mà còn làm tăng rủi ro quá khớp (overfitting) và chi phí tính toán. Các biểu đồ này giúp củng cố nhận định, việc tăng số lá ban đầu giúp cải thiện hiệu suất dự báo. Tuy nhiên, hiệu quả này không tiếp tục tăng tuyến tính mà có dấu hiệu bão hòa, thậm chí tăng sai số nhẹ khi mô hình trở nên quá phức tạp.

3.2. Thảo luận

Kết quả thực nghiệm cho thấy, mô hình LightGBM có tính ổn định cao khi các siêu tham số thay đổi trong phạm vi hợp lý. Việc tăng số lá tối đa (num_leaves) từ giá trị nhỏ đến trung bình giúp cải thiện đáng kể độ chính xác dự báo, đặc biệt ở kịch bản S2 (num_leaves = 60), sai số MAPE giảm mạnh xuống 82,94% so với mức 90,67% ở kịch bản S1 (num_leaves = 30). Tuy nhiên, khi num_leaves lớn hơn 180, các chỉ số MAPE và RMSE không tiếp tục giảm mà có xu hướng dao động hoặc tăng nhẹ. Ví dụ, RMSE đạt 4230,92 kW ở kịch bản S10 (num_leaves = 300), là giá trị cao nhất trong các thử nghiệm, trong khi thời gian huấn luyện cũng kéo dài đến hơn 3 giây, gấp hơn 10 lần so với mô hình đơn giản ban đầu. Điều này cho thấy khả năng mô hình bắt đầu quá khớp tại các mức độ phức tạp cao. Khi num_leaves tăng từ 60 đến 180, MAPE tăng nhẹ từ 82,94% lên 87,62%, cho thấy mô hình bắt đầu giảm hiệu quả do quá khớp. Tuy nhiên, từ num_leaves = 210 đến 300, MAPE dao động không rõ xu hướng, vì chỉ số này nhạy với các điểm dữ liệu có giá trị thực gần 0 (ví dụ, lúc bình minh hoặc hoàng hôn). Tại những thời điểm đó, sai số dự báo nhỏ cũng gây tỷ lệ phần trăm cao, làm cho MAPE dao động bất ổn giữa các kịch bản. Trong khi NMAPE lại chỉ tăng nhẹ và khá ổn định (9,61% tại num_leaves = 60 đến 10,36% tại 300), vì được chuẩn hóa theo công suất định mức nên ít bị ảnh hưởng bởi các giá trị nhỏ. Tuy nhiên, từ num_leaves = 180 trở đi, NMAPE gần như không cải thiện, cho thấy mô hình đã đạt đến ngưỡng bão hòa, việc tăng phức tạp không mang lại hiệu quả đáng kể. RMSE liên tục tăng từ num_leaves = 60 đến 300 (từ 3869 kW đến 4230 kW), RMSE là sai số trung bình có trọng số bình phương, do đó nó nhạy với các điểm sai số lớn. Khi num_leaves tăng, mô hình phức tạp hơn có thể gây ra các dự báo sai lệch lớn ở một vài thời điểm, điều này làm RMSE tăng rõ rệt.

Việc điều chỉnh learning_rate và n_estimators đơn lẻ cũng không cho thấy hiệu quả vượt trội nếu không đi kèm với cấu hình num_leaves phù hợp. Chẳng hạn, S8 sử dụng learning rate

thấp (0,01) và số cây cao (500), có thời gian huấn luyện cao nhất (4,65 giây), nhưng sai số MAPE vẫn ở mức 86,46%, không tốt hơn so với S2 hoặc S5. Điều này cho thấy, chỉ khi các siêu tham số phối hợp hợp lý, mô hình mới phát huy hiệu quả tối ưu.

Từ góc độ ứng dụng thực tiễn, kết quả này rất có ý nghĩa. Trong điều kiện tài nguyên tính toán hạn chế và yêu cầu vận hành thời gian thực, việc lựa chọn cấu hình LightGBM đơn giản với num_leaves từ 60–150, learning_rate từ 0,05 đến 0,07 và số cây từ 100–200 đã cho kết quả tương đối tốt. Như vậy, người dùng không cần tối ưu quá sâu từng siêu tham số mà vẫn có thể đạt được hiệu suất dự báo chấp nhận được, giảm thiểu chi phí tính toán và thời gian huấn luyện.

4. KẾT LUẬN

Bài báo đã tiến hành đánh giá ảnh hưởng của các siêu tham số chính trong mô hình LightGBM, bao gồm số lá tối đa (num_leaves), tốc độ học (learning_rate) và số lượng cây học (n_estimators) đến hiệu suất dự báo công suất phát điện mặt trời. Thông qua 10 kịch bản thử nghiệm với các tổ hợp tham số khác nhau, kết quả cho thấy việc tinh chỉnh siêu tham số có thể góp phần cải thiện độ chính xác dự báo, đặc biệt khi tăng num_leaves từ giá trị nhỏ lên khoảng 150. Tuy nhiên, khi tiếp tục tăng độ phức tạp mô hình vượt quá ngưỡng này, sai số không tiếp tục giảm mà còn có xu hướng tăng nhẹ, cho thấy dấu hiệu quá khớp. Cụ thể, mô hình LightGBM đạt sai số MAPE thấp nhất 82,94% khi num_leaves = 60, learning_rate = 0,05 và n_estimators = 100. Tuy nhiên, sai số không giảm thêm khi mô hình phức tạp hơn. Điều này xác nhận rằng, LightGBM có thể đạt hiệu suất tốt với cấu hình vừa phải mà không cần tối ưu hóa sâu.

Các chỉ số như RMSE, MAPE, NMAPE chỉ dao động trong biên độ hẹp giữa các kịch bản, thể hiện tính ổn định của mô hình LightGBM trong phạm vi tham số thử nghiệm. Đồng thời, việc giảm hoặc tăng tốc độ học và số lượng cây học không mang lại cải thiện đáng kể nếu không được điều chỉnh đồng thời một cách phù hợp.

Từ kết quả nghiên cứu, có thể kết luận rằng LightGBM là một mô hình có độ ổn định cao, không quá nhạy cảm với thay đổi vừa phải của các siêu tham số. Do đó, trong thực tế ứng dụng, việc lựa chọn các giá trị siêu tham số vừa phải và hợp lý (ví dụ: num_leaves trong khoảng 60 – 150), có thể đem lại hiệu quả dự báo tốt mà không cần tiêu tốn nhiều thời gian để tinh chỉnh sâu. Hướng nghiên cứu tiếp theo có thể mở rộng sang việc tối ưu hóa tham số tự động hoặc kết hợp thêm các đặc trưng đầu vào mới để nâng cao hơn nữa hiệu suất mô hình.

TÀI LIỆU THAM KHẢO

[1]. K. Reba, J. Bevc, A.- Vásquez, and M. Jankovec, "Photovoltaic Energy Production Forecasting using LightGBM," *55th Int. Conf. Microelectron. Devices Mater. with Work. Laser Syst. Photonics*, pp. 36–39, 2019.

[2]. J. Ye, B. Zhao, and H. Deng, "Photovoltaic Power Prediction Model Using Pre-train and Fine-tune Paradigm Based on LightGBM and XGBoost," *Procedia Comput. Sci.*, vol. 224, pp. 407–412, 2023, doi: 10.1016/j.procs.2023.09.056.

[3]. M. F. Hanif, M. S. Naveed, M. Metwaly, J. Si, X. Liu, and J. Mi, "Advancing solar energy forecasting with modified ANN and light GBM learning algorithms," *AIMS Energy*, vol. 12, no. 2, pp. 350-386., 2024, doi: 10.3934/energy.2024017.

[4]. G. Ke and I.-Y. Meng, QiFinely, ThomasWang, TaifengChen, WeiMa, WeidongYe, QiweiTLiu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *31st Conf. Neural Inf. Process. Syst. (NIPS 2017), Long Beach, CA, USA*, no. Nips, 2017.

[5]. H. N. Nguyen, Q. T. Tran, C. T. Ngo, D. D. Nguyen, and V. Q. Tran, "Solar energy prediction through machine learning models: A comparative analysis of regressor algorithms," *PLoS One*, vol. 20, no. 1, pp. 1–23, 2025, doi: 10.1371/journal.pone.0315955.

[6]. Khanh-Toan Nguyen, Thanh-Ngoc Tran, and Huy-Tuan Nguyen, "Research on the Influence of Hyperparameters on the LightGBM Model in Load Forecasting," *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 5, pp. 17005–17010, 2024, doi: 10.48084/etasr.8266.

[7]. E. S. Solano, P. Dehghanian, and C. M. Affonso, "Solar Radiation Forecasting Using Machine Learning and Ensemble Feature Selection," *Energies*, vol. 15, no. 19, 2022, doi: 10.3390/en15197049.

[8]. R. Ahmed, V. Sreeram, Y. Mishra, and M. D. Arif, "A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization," *Renew. Sustain. Energy Rev.*, vol. 124, no. February, p. 109792, 2020, doi: 10.1016/j.rser.2020.109792.

[9]. P. Li, K. Zhou, X. Lu, and S. Yang, "A hybrid deep learning model for short-term PV power forecasting," *Appl. Energy*, vol. 259, no. November, p. 114216, 2020, doi: 10.1016/j.apenergy.2019.114216.

[10]. Q. T. Phan, Y. K. Wu, Q. D. Phan, and H. Y. Lo, "A Novel Forecasting Model for Solar Power Generation by a Deep Learning Framework With Data Preprocessing and Postprocessing," *IEEE Trans. Ind. Appl.*, vol. 59, no. 1, pp. 220–231, 2023, doi: 10.1109/TIA.2022.3212999.

Giới thiệu tác giả

Tác giả **Nguyễn Tuấn Anh** tốt nghiệp Trường Đại học Nông nghiệp Hà Nội (nay là Học viện Nông nghiệp Việt Nam) ngành Kỹ thuật điện năm 2012 và nhận bằng Thạc sĩ Kỹ thuật điện tại Trường Đại học Điện lực vào năm 2017. Hiện, tác giả là giảng viên tại Khoa Kỹ thuật điện, Trường Đại học Điện lực.

Hướng nghiên cứu chính: khí cụ điện, năng lượng tái tạo, dự báo điện mặt trời, DCS và SCADA.

Tác giả **Phạm Mạnh Hải** tốt nghiệp Trường Đại học Bách khoa Hà Nội ngành Hệ thống điện năm 2006; nhận bằng Thạc sĩ ngành Kỹ thuật điện tại Đại học Paul Sabatier, Toulouse, Pháp năm 2008; bảo vệ Luận án Tiến sĩ ngành Hóa hữu cơ ứng dụng - Plasma cho năng lượng tại Đại học Poitiers (ENSIP), Poitiers, Pháp năm 2011. Hiện, tác giả công tác tại Khoa Năng lượng mới, Trường Đại học Điện lực.

Lĩnh vực nghiên cứu: thuật toán điều chỉnh tham số tuabin gió, thuật toán tối ưu, dự báo phụ tải điện, thuật toán trí tuệ nhân tạo, dự báo điện gió và điện mặt trời, năng lượng tái tạo, sản xuất Biomass-Biogas, độ tin cậy trong hệ thống điện.

Tác giả **Vũ Minh Pháp** tốt nghiệp Học viện Kỹ thuật Quân sự năm 2007; nhận bằng Thạc sĩ năm 2012 ngành Kỹ thuật điện tử tại Trường Đại học Giao thông vận tải, bằng Tiến sĩ năm 2018 ngành Kỹ thuật hệ thống điện – điện tử tại Đại học Mie, Nhật Bản. Hiện, tác giả đang là nghiên cứu viên chính tại Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Hướng nghiên cứu chính: năng lượng mặt trời, năng lượng tái tạo, hệ thống năng lượng kết hợp, hệ thống chuyển đổi năng lượng.