XÂY DỰNG MÔ HÌNH DỰ BÁO DOANH THU DỰA TRÊN MÔ HÌNH HỒI QUY TUYẾN TÍNH TRONG LĨNH VỰC DỊCH VỤ NHÀ HÀNG

DEVELOPMENT OF A REVENUE FORECASTING MODEL BASED ON LINEAR REGRESSION IN THE RESTAURANT SERVICE INDUSTRY

Nguyễn Thị Thu Thuỷ[1], Nguyễn Thị Thanh Loan[1,*], Trần Hùng Cường[2]

DOI: https://doi.org/10.57001/huih5804.2025.372

TÓM TẮT

Trong thời đại dữ liệu lớn, việc dự báo doanh thu đóng vai trò quan trọng trong tối ưu hóa chiến lược kinh doanh và ra quyết định của doanh nghiệp. Bài báo này trình bày quy trình xây dựng mô hình dự báo doanh thu trong lĩnh vực dịch vụ nhà hàng bằng phương pháp hồi quy tuyến tính. Dữ liệu được thu thập từ bộ dữ liệu với 1.000 nhà hàng, với các biến như số lượng khách hàng, giá thực đơn, chi phí marketing, loại hình ẩm thực và đánh giá khách hàng. Bằng phương pháp nghiên cứu định lượng, kết quả cho thấy số lượng khách hàng có tương quan cao nhất với doanh thu (r = 0,75), mô hình hồi quy tuyến tính đạt độ chính xác đáng kể trên tập dữ liệu thực hiện kiểm tra. Kết quả cũng được so sánh với các mô hình học máy khác như cây quyết định và rừng ngẫu nhiên để đánh giá hiệu quả của các mô hình. Bài báo đưa ra những khuyến nghị giúp doanh nghiệp tối ưu hóa chỉ tiêu doanh thu của các doanh nghiệp kinh doanh trong lĩnh vực này.

Từ khóa: *Mô hình dự báo, doanh thu, mô hình hồi quy, mô hình học máy.*

ABSTRACT

In the era of big data, revenue forecasting plays an important role in optimizing business strategies and decision making of enterprises. This paper presents the process of building a revenue forecasting model in the restaurant service sector using the linear regression method. Data were collected from a dataset of 1,000 restaurants, with variables such as number of customers, menu prices, marketing costs, cuisine types and customer reviews. Using quantitative research methods, the results showed that the number of customers had the highest correlation with revenue (r = 0.75), the linear regression model achieved significant accuracy on the test dataset. The results were also compared with other machine learning models such as decision trees and random forests. The article provides recommendations to help businesses optimize revenue targets of businesses in this field.

Keywords**:** *Forecasting models, revenue models, regression models, machine learning models.*

[1]Trường Kinh tế, Trường Đại học Công nghiệp Hà Nội

[2]Trường Công nghệ thông tin và truyền thông, Trường Đại học Công nghiệp Hà Nội

[*]Email: nguyenthithanhloan@haui.edu.vn

Ngày nhận bài: 10/7/2025

Ngày nhận bài sau phản biện: 15/9/2015

Ngày chấp nhận đăng: 28/10/2025

1. GIỚI THIỆU

Trong bối cảnh môi trường kinh doanh cạnh tranh khốc liệt như hiện nay, việc dự báo doanh thu là một yếu tố vô cùng quan trọng giúp các doanh nghiệp đưa ra các quyết định chiến lược hiệu quả, tối ưu hóa nguồn lực và tăng trưởng bền vững. Dự báo doanh thu chính xác

không chỉ giúp doanh nghiệp chủ động trong việc quản lý tài chính mà còn giúp họ xây dựng kế hoạch kinh doanh phù hợp với thực tế và xu hướng thị trường [2, 5].

Mô hình hồi quy tuyến tính, một trong những phương pháp phổ biến đơn giản và hiệu quả trong phân tích dữ liệu, đã được ứng dụng rộng rãi trong việc dự báo doanh thu. Mô hình này có khả năng tìm ra mối quan hệ giữa các yếu tố tác động đến doanh thu, từ đó giúp dự báo một cách khoa học và chính xác [2, 6]. Tuy nhiên, việc áp dụng mô hình hồi quy tuyến tính vào thực tế kinh doanh còn gặp nhiều thách thức, nhất là trong việc xác định các yếu tố đầu vào và xử lý dữ liệu sao cho phù hợp.

Theo báo cáo thị trường kinh doanh ẩm thực tại Việt Nam, tính đến hết năm 2024, số lượng cửa hàng F&B tại Việt Nam ước tính đạt 323.010 cửa hàng, tăng 1,8% so với cùng kỳ năm trước. Bất chấp những khó khăn về tiêu dùng, doanh thu ngành F&B tại Việt Nam năm 2024 vẫn đạt khoảng 688,8 nghìn tỷ đồng, tăng 16,6% so với năm 2023. Mặc dù doanh thu chung của ngành tăng, nhưng theo khảo sát từ 4.005 đơn vị F&B trên cả nước, chỉ 25,5% doanh nghiệp ghi nhận doanh thu ổn định so với cùng kỳ năm 2023 và 14,7% có mức tăng trưởng. Dưới sức ép của giá nguyên vật liệu tăng cao, có tới 49,2% doanh nghiệp F&B dự kiến tăng giá trong năm 2025 để đối phó áp lực chi phí. Việc áp dụng mô hình tuyến tính trong việc dự báo doanh thu các doanh nghiệp kinh doanh trong lĩnh vực nhà hàng là cần thiết để giúp doanh nghiệp cải thiện được hoạt động quản lý về hiệu quả kinh doanh [3].

Nghiên cứu đã thực hiện phát triển và xây dựng mô hình hồi quy tuyến tính phù hợp để dự báo doanh thu, xác định các yếu tố ảnh hưởng đến doanh thu và tìm hiểu mối quan hệ giữa các yếu tố này trong các doanh nghiệp thuộc mẫu nghiên cứu kinh doanh trong lĩnh vực cung cấp dịch vụ nhà hàng. Phân tích và đánh giá mức độ chính xác của mô hình dự báo doanh thu, so sánh kết quả dự báo với dữ liệu thực tế để xác định khả năng ứng dụng trong các doanh nghiệp. Từ đó, nghiên cứu cũng đề xuất một số khuyến nghị giúp các doanh nghiệp trong việc đưa ra các quyết định chiến lược kịp thời, chính xác để tối ưu hóa chiến lược phát triển, cải thiện quá trình dự báo doanh thu và quản lý nguồn lực.

2. MÔ HÌNH VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Dữ liệu nghiên cứu

Dữ liệu "Restaurant_revenue.csv" được thu thập từ trang web: *kaggle.com* bao gồm thông tin chi tiết về 1.000 nhà hàng, với 8 biến quan trọng (bảng 1 và 2).

Bảng 1. Mô tả biến trong mô hình nghiên cứu

STT	Các biến	Mô tả biến	Ý nghĩa
1	Number_of_ Customers	Số lượng khách hàng ghé thăm nhà hàng trong mỗi tháng	Mức độ phổ biến và khả năng thu hút khách hàng của nhà hàng.
2	Menu_Price	Giá trung bình của các món ăn trên thực đơn.	Là một yếu tố quan trọng thể hiện chiến lược định giá sản phẩm và ảnh hưởng trực tiếp đến doanh thu.
3	Marketing_Spend	Tổng chi phí tiếp thị hàng tháng của nhà hàng	Đo lường mức đầu tư vào các hoạt động quảng bá và chiến dịch tiếp thị nhằm tăng lượng khách hàng.
4	Cuisine_Type	Loại ẩm thực mà nhà hàng phục vụ	Phản ánh đặc điểm định vị của nhà hàng trên thị trường và thị hiếu khách hàng mục tiêu.
5	Average_ Customer_Spending	Mức chi tiêu trung bình của một khách hàng tại nhà hàng	Là thước đo quan trọng về mức độ tiêu dùng của khách hàng, thể hiện giá trị trung bình mà một khách hàng đóng góp vào doanh thu.
6	Promotions	Biến nhị phân, thể hiện nhà hàng có thực hiện chương trình khuyến mãi hay không (0: Không; 1: Có)	Đo lường hiệu quả của các chương trình khuyến mãi trong việc thúc đẩy doanh thu và thu hút khách hàng.
7	Reviews	Tổng số lượng đánh giá mà nhà hàng nhận được trên các nền tảng trực tuyến.	chỉ số phản ánh mức độ hài lòng và mức độ tương tác của khách hàng.
8	Monthly_Revenue	Doanh thu hàng tháng của nhà hàng.	Là biến mục tiêu cần dự đoán, đại diện cho hiệu quả kinh doanh của nhà hàng.

Bảng 2. Bảng tổng hợp dữ liệu

No.	Number_of _Customers	Menu_Price	Maketing_Spend	Cuisine_Type	Average_ Customer_Spending	Promotions	Reviews	Monthly_Revenue
0	61	43,117635	12,663793	Japanese	36,236133	0	45	350,912040
1	24	40,020077	4,577892	Italian	17,952562	0	36	221,319091
2	81	41,981485	4,652911	Japanese	22,600420	1	91	326,529763
3	70	43,005307	4,416053	Italian	18,984098	1	59	348,190573
4	30	14,456199	3,475052	Italian	12,766143	1	30	185,009121

2.2. Phương pháp nghiên cứu

Nghiên cứu thực hiện xây dựng mô hình hồi quy tuyến tính để xác định mối quan hệ giữa doanh thu (biến phụ thuộc) và các yếu tố tác động (biến độc lập).

Nghiên cứu cũng sử dụng các phương pháp phân tích thống kê để kiểm tra độ phù hợp của mô hình [4], bao gồm: Phân tích tương quan thông qua kiểm tra tính phù hợp của mô hình (Sử dụng hệ số xác định R^2, p-value, F-test) và kiểm tra các giả thuyết về mối quan hệ giữa các yếu tố và doanh thu.

Ngoài ra, để kiểm tra độ chính xác của mô hình, nghiên cứu sử dụng các chỉ số Mean Squared Error (MSE), Mean Absolute Error (MAE), kiểm tra độ chính xác dự báo (Forecast accuracy) để đánh giá khả năng dự báo của mô hình. Sau đó, nghiên cứu thực hiện so sánh với các mô hình khác kết quả thực hiện của các mô hình để kiểm tra độ chính xác và xem xét tính khả thi của mô hình dự báo.

2.3. Quy trình thực hiện nghiên cứu

Nghiên cứu được thực hiện theo các bước như sau [4]:

Bước 1: Tiền xử lý, nhóm nghiên cứu kiểm tra dữ liệu trước khi phân tích như: Kiểm tra dữ liệu còn thiếu để đếm giá trị thiếu trong từng cột và sắp xếp chúng theo giá trị giảm dần. Kiểm tra giá trị trùng lặp trong toàn bộ dataset để loại bỏ (bằng hàm duplicate()). Loại bỏ ngoại lệ bằng IQR và Z-score, mã hóa biến phân loại (Cuisine_Type) thành nhị phân.

Bước 2: Thực hiện khám phá dữ liệu bằng cách phân tích dữ liệu đơn biến để hiểu rõ đặc điểm của từng biến trong tập dữ liệu, có thể xác định hình dạng phân phối và nhận diện giá trị bất thường.

Bước 3: Phân tích tương quan bằng cách sử dụng biểu đồ để xác định các mối quan hệ tuyến tính mạnh.

Bước 4: Xây dựng mô hình hồi quy tuyến tính đa biến trên Python (sklearn), chia dữ liệu thành tập huấn luyện và kiểm tra.

Bước 5: Xây dựng mô hình cây quyết định, mô hình rừng ngẫu nhiên và so sánh kết quả thực hiện với mô hình hồi quy.

3. KẾT QUẢ NGHIÊN CỨU

3.1. Xây dựng mô hình hồi quy tuyến tính

Nghiên cứu tiến hành chia tập dữ liệu thành tập huấn luyện (Training set) chiếm 80% và tập kiểm tra (Test set) chiếm 20%. Việc chia tập dữ liệu như vậy đảm bảo rằng mô hình được đánh giá một cách khách quan và tránh tình trạng overfitting (mô hình học quá khớp với dữ liệu huấn luyện mà không khả năng tổng quát hóa tốt) [9].

Đồng thời, nghiên cứu cũng thực hiện tạo một đối tượng lr đại diện cho một mô hình hồi quy tuyến tính. Đối tượng này sẽ được sử dụng để tìm ra mối quan hệ tuyến tính giữa các biến độc lập (trong x_train) và biến phụ thuộc (trong y_train).

lr.fit(x_train, y_train)

Khi gọi phương thức fit(), mô hình sẽ tìm ra các hệ số của phương trình hồi quy tuyến tính sao cho đường thẳng biểu diễn mối quan hệ giữa các biến độc lập và biến phụ thuộc đi qua các điểm dữ liệu trong tập huấn luyện một cách gần nhất.

y_pred = lr.predict(x_test)

Phương thức này dùng để dự đoán giá trị của biến phụ thuộc cho các mẫu dữ liệu mới.

Thực hiện dự đoán các giá trị của biến phụ thuộc trên chính tập dữ liệu huấn luyện (training set) mà mô hình đã được học.

Kiểm tra xem mô hình đã học được tốt đến đâu: Bằng cách so sánh các giá trị dự đoán y_pred_train_lr với các giá trị thực tế trong y_train, chúng ta có thể đánh giá xem mô hình đã nắm bắt được mối quan hệ giữa các biến độc lập và biến phụ thuộc tốt đến mức nào.

Phát hiện các vấn đề tiềm ẩn: Nếu mô hình không thể dự đoán chính xác trên chính dữ liệu mà nó đã được huấn luyện, điều đó có thể cho thấy các vấn đề như:

Overfitting: Mô hình quá phức tạp, học quá kỹ các chi tiết nhiễu trong dữ liệu huấn luyện, dẫn đến khả năng tổng quát hóa kém.

Underfitting: Mô hình quá đơn giản, không thể nắm bắt được mối quan hệ phức tạp giữa các biến.

Đánh giá hiệu suất của một mô hình hồi quy tuyến tính (linear regression) bằng cách sử dụng chỉ số R-squared. Có hai dòng code, mỗi dòng đánh giá trên một tập dữ liệu khác nhau (test set và train set) [5]:

r2_score_lr_test = r2_score(y_test,y_pred

r2_score_lr_train = r2_score(y_train,y_pred_train_lr)

Kết quả dự báo thực hiện được như sau:

```
print(f"The Train r2 score is: {r2_score(y_train,y_pred_train_lr)}")
print(f"The RMSE score for Train data is: {np.sqrt(mean_squared_error(y_train,y_pred_train_lr ))}")
sns.kdeplot(y_train, label='Actual Value')
sns.kdeplot(y_pred_train_lr, label='Predicted Value')
plt.grid()
plt.title('Plot of Y_Train Vs Y_Train_Predicted')
plt.legend()
plt.show()
print("--"*50)
print(f"The Test r2 score is: {r2_score(y_test,y_pred)}")
print(f"The RMSE score for Test data is: {np.sqrt(mean_squared_error(y_test,y_pred))}")
sns.kdeplot(y_test, label='Actual Value')
sns.kdeplot(y_pred, label='Predicted Value')
plt.grid()
plt.title('Plot of Y_Test Vs Y_Test_Predicted')
plt.legend()
plt.show()
```

Nghiên cứu thực hiện đánh giá hiệu suất mô hình bằng cách tính toán R^2 và RMSE cho cả tập dữ liệu huấn luyện và dữ liệu thử nghiệm, có thể đánh giá mức độ mô hình nắm bắt mối quan hệ cơ bản trong dữ liệu huấn luyện và mức độ khái quát hóa dữ liệu không nhìn thấy trong tập thử nghiệm [4].

Hơn nữa, nghiên cứu tiến hành trực quan hóa phân phối bằng biểu đồ KDE để cho phép so sánh phân phối của giá trị thực tế và giá trị dự đoán, cung cấp sự hiểu biết trực quan về mức độ phù hợp của các dự đoán với giá trị thực tế.

Đánh giá mô hình dựa trên tập dữ liệu Train

R-squared (R^2) = 0,6901 cho biết mô hình giải thích được khoảng 69,01% sự biến động của biến phụ thuộc (doanh thu). Điều này có nghĩa là mô hình có khả năng dự đoán khá tốt. Tuy nhiên, vẫn còn khoảng 31% sự biến động chưa được giải thích, có thể do các yếu tố khác không được đưa vào mô hình hoặc do sai số ngẫu nhiên.

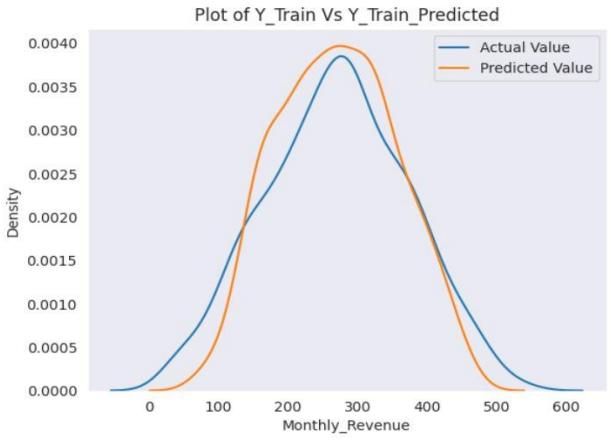
RMSE (Root Mean Squared Error) = 57,31: RMSE là một thước đo sai số trung bình căn bậc hai. Giá trị RMSE càng nhỏ thì mô hình càng chính xác. Giá trị RMSE ở đây khá lớn, cho thấy sai số dự đoán của mô hình còn cao.

Qua biểu đồ phân phối mật độ (hình 1) cho thấy, đường biểu diễn giá trị thực tế và giá trị dự đoán đều có hình dạng tương tự nhau, điều này chỉ ra rằng mô hình đã bắt được xu hướng chung của dữ liệu.

Về độ lệch: có thể thấy một số điểm khác biệt giữa hai đường biểu diễn, đặc biệt là ở các giá trị lớn của doanh thu. Điều này cho thấy mô hình chưa thể dự đoán chính xác các giá trị cực đại và cực tiểu.

Về độ phân tán: Các giá trị dự đoán có độ phân tán tương đối lớn so với giá trị thực tế, đặc biệt ở phần đuôi của phân phối. Điều này cũng góp phần làm tăng giá trị RMSE.

```
The Train r2 score is: 0.6901128223461557
The RMSE score for Train data is: 57.30589579168436
```



Hình 1. Biểu đồ phân phối mật độ

Mô hình đã thể hiện khả năng dự đoán tương đối tốt về xu hướng chung của doanh thu hàng tháng. Tuy nhiên, vẫn còn một số hạn chế của mô hình được nhận thấy như sau:

Sai số dự đoán còn cao: Giá trị RMSE khá lớn cho thấy mô hình chưa đủ chính xác để đưa ra các quyết định kinh doanh quan trọng.

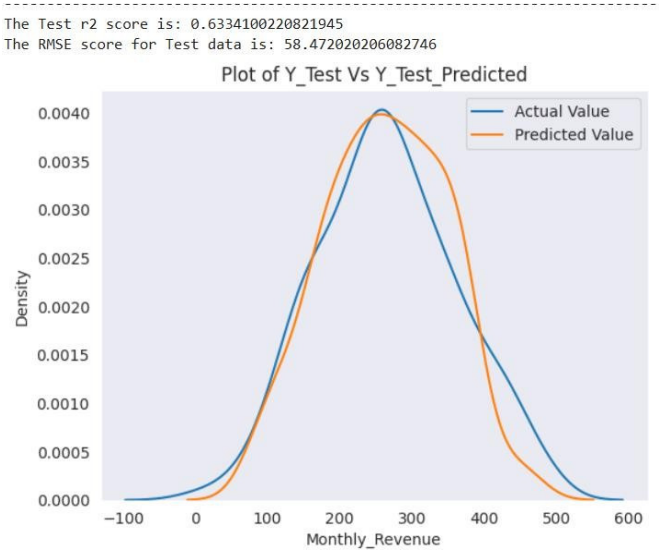
Khả năng dự đoán các giá trị cực trị còn hạn chế: Mô hình chưa thể bắt được các biến động đột ngột của doanh thu.

Đánh giá mô hình trên tập dữ liệu Test

R-squared (R^2) = 0,6334: Tương tự như trên tập huấn luyện, R^2 trên tập kiểm thử cho biết mô hình giải thích được khoảng 63,34% sự biến động của doanh thu. Điều này có nghĩa là mô hình cũng có khả năng dự đoán khá tốt trên dữ liệu mới. Tuy nhiên, vẫn còn một phần đáng kể sự biến động chưa được giải thích.

RMSE (Root Mean Squared Error) = 58,472: Giá trị RMSE trên tập kiểm thử cho thấy sai số dự đoán trung bình của mô hình là khoảng 58,47. Giá trị này gần tương đương với giá trị RMSE trên tập huấn luyện, cho thấy mô hình không bị overfitting (tức là quá khớp với dữ liệu huấn luyện mà không khả năng tổng quát hóa tốt) [5].

Qua biểu đồ phân phối mật độ (hình 2) cho thấy cả đường biểu diễn giá trị thực tế và giá trị dự đoán trên tập kiểm thử đều có hình dạng tương tự nhau, cho thấy mô hình đã bắt được xu hướng chung của dữ liệu mới.



Hình 2. Biểu đồ phân phối mật độ (KDE plot)

Về độ lệch: có một số điểm khác biệt giữa hai đường biểu diễn, đặc biệt là ở các giá trị lớn của doanh thu. Điều này cho thấy mô hình vẫn chưa thể dự đoán chính xác các giá trị cực đại và cực tiểu.

Về độ phân tán: Các giá trị dự đoán trên tập kiểm thử có độ phân tán tương đối lớn so với giá trị thực tế, tương tự như trên tập huấn luyện.

Tóm lại, nhìn chung mô hình đã thể hiện khả năng tổng quát hóa tương đối tốt từ tập huấn luyện sang tập kiểm thử, cho thấy mô hình không bị overfitting quá nhiều. Hơn nữa, mặc dù mô hình đã giải thích được một phần đáng kể sự biến động của doanh thu, nhưng sai số dự đoán vẫn còn khá cao. Điều này có nghĩa là các dự đoán của mô hình vẫn chưa đủ chính xác để đưa ra các quyết định kinh doanh quan trọng. Tuy vậy, mô hình hồi quy tuyến tính cũng cho thấy tiềm năng trong việc dự đoán doanh thu hàng tháng của tập dữ liệu về kinh doanh trong lĩnh vực dịch vụ nhà hàng.

3.2. Xây dựng mô hình Cây quyết định (Decision Tree), mô hình rừng ngẫu nhiên (Random Forest) và so sánh kết quả giữa các mô hình

3.2.1. Mô hình Cây quyết định

Trong nghiên cứu cũng sử dụng mô hình cây quyết định để thực hiện dự báo doanh thu theo các bước như sau [1, 7]:

Thứ nhất, khởi tạo mô hình và huấn luyện mô hình với dữ liệu huấn luyện (training data), và cuối cùng là dự đoán giá trị cho dữ liệu kiểm tra (testing data).

Thứ hai, đánh giá hiệu suất của một mô hình cây quyết định. Bằng cách tính toán R^2 trên cả tập dữ liệu huấn luyện và tập kiểm tra.

Thứ ba, thực hiện so sánh với các mô hình khác qua R^2 và r^2_score.

R^2 có thể được sử dụng để so sánh hiệu suất của mô hình cây quyết định với các mô hình khác. Nó đo lường mức độ phù hợp của mô hình với dữ liệu huấn luyện. Giá trị càng gần 1, mô hình càng tốt.

Tính toán sai số bình phương trung bình (mean_squared_error), sau đó tính căn bậc hai để được RMSE. RMSE cho biết độ lớn trung bình của sai số dự đoán.

Kết quả có được từ tập huấn luyện (train) cho thấy:

R-squared (R^2) = 1,0: Giá trị R^2 bằng 1 cho thấy mô hình giải thích hoàn toàn sự biến động của dữ liệu huấn luyện. Nói cách khác, mô hình đã "học thuộc lòng" dữ liệu huấn luyện.

RMSE (Root Mean Squared Error) = 0,0: Giá trị RMSE bằng 0 cho thấy không có sai số giữa giá trị dự đoán và giá trị thực tế trên tập huấn luyện.

Mô hình có R^2 bằng 1 và RMSE bằng 0 trên tập huấn luyện thường là dấu hiệu của overfitting (quá khớp). Điều

này có nghĩa là mô hình đã quá phức tạp, bắt quá nhiều nhiễu trong dữ liệu huấn luyện, dẫn đến khả năng tổng quát hóa kém trên dữ liệu mới.

Kết quả có được tập kiểm thử (test) cho thấy:

R-squared (R^2) = 0,0860: Giá trị R^2 rất thấp, cho thấy mô hình chỉ giải thích được khoảng 8.6% sự biến động của doanh thu trên tập dữ liệu kiểm thử. Điều này có nghĩa là mô hình có khả năng dự đoán rất kém.

RMSE (Root Mean Squared Error) = 92,325: Giá trị RMSE rất cao, cho thấy sai số dự đoán trung bình của mô hình là rất lớn. Điều này cũng khẳng định mô hình không thể dự đoán chính xác doanh thu.

3.2.2. Mô hình rừng ngẫu nhiên

Nghiên cứu sử dụng mô hình rừng ngẫu nhiên trong Python để xây dựng một mô hình dự đoán giá trị liên tục (regression) [8] bằng cách thực hiện các bước sau:

Thứ nhất, tạo mô hình.

rfr = RandomForestRegressor(): Tạo một đối tượng Random Forest Regressor.

model = rfr.fit(x_train, y_train): Huấn luyện mô hình trên tập dữ liệu huấn luyện (x_train, y_train).

Thứ hai, dự đoán.

y_pred_test_rfr = rfr.predict(x_test): Dự đoán giá trị cho tập dữ liệu kiểm thử (x_test).

y_pred_train_rfr = rfr.predict(x_train): Dự đoán giá trị cho tập dữ liệu huấn luyện (để đánh giá độ khớp).

Thứ ba, đánh giá mô hình.

Nghiên cứu thực hiện đánh giá mô hình thông qua R^2 và RMSE, trong đó:

R-squared: Đo lường mức độ phù hợp của mô hình với dữ liệu. Giá trị càng gần 1, mô hình càng tốt.

RMSE (Root Mean Squared Error): Đo lường sai số trung bình bình phương căn bậc hai, càng nhỏ càng tốt.

Quá trình đánh giá sẽ được thực hiện trên tập huấn luyện (train) và trên tập kiểm tra (test) với kết quả như sau:

Giá trị R^2 trên tập huấn luyện: 0,9501. Giá trị R^2 rất cao, gần bằng 1, cho thấy mô hình phù hợp rất tốt với dữ liệu huấn luyện. Điều này có nghĩa là mô hình đã học được các quy luật trong dữ liệu huấn luyện một cách hiệu quả.

Giá trị RMSE trên tập huấn luyện: 22,9906. Giá trị RMSE tương đối lớn so với R^2. Điều này có thể cho thấy mô hình đang bị overfitting (học quá kỹ dữ liệu huấn luyện) hoặc đơn vị đo của biến mục tiêu (Monthly_Revenue) lớn.

Giá trị R^2 trên tập kiểm tra: 0,5774. Giá trị R^2 thấp hơn so với giá trị trên tập huấn luyện, cho thấy mô hình không phù hợp với dữ liệu kiểm thử tốt bằng dữ liệu huấn luyện. Điều này là dấu hiệu của overfitting. Mô hình đã học quá kỹ các đặc điểm ngẫu nhiên trong dữ liệu huấn luyện, dẫn đến khả năng dự đoán kém trên dữ liệu mới.

Giá trị RMSE trên tập kiểm tra: 62,7798. Giá trị RMSE khá lớn, cho thấy sai số dự đoán trung bình của mô hình là khá cao. Điều này đồng nghĩa với việc các giá trị dự đoán của mô hình thường cách xa giá trị thực tế.

3.3. So sánh kết quả đánh giá các mô hình học máy

Từ kết quả thực hiện trên các mô hình hồi quy, mô hình rừng ngẫu nhiên và mô hình cây quyết định, tác giả tổng hợp và so sánh kết quả đánh giá giữa các mô hình thông qua (bảng 3).

Bảng 3. So sánh kết quả đánh giá giữa các mô hình

Model	R-squared train	R-squared test	RMSE train	RMSE test
Linear Regression	0.6901128223461557	0.6334100220821945	57.30589579168436	58.472020206082746
Decision Tree	1.0	0.08602993525725888	0.0	92.32595122480981
Random Forest	0.9501223193168491	0.5774045571557661	22.99061705660671	62.77984076074729

Kết quả cuối cùng cho thấy, mô hình rừng ngẫu nhiên có hiệu suất tốt nhất, tất cả kết quả các độ đo của mô hình rừng ngẫu nhiên (Random Forest) đều cao hơn so với mô hình cây quyết định (Decision Tree), với R-squared train là 0,95, R- squared test là 0,577, chỉ số RSME train là 22,99 và RMSE là 62,77. Đặc biệt RMSE test của mô hình rừng ngẫu nhiên đã giảm so với mô hình cây quyết định đáng kể. RMSE train đã tăng 22,99 đơn vị so với mô hình mô hình cây quyết định.

3.4. Hạn chế của nghiên cứu và khuyến nghị

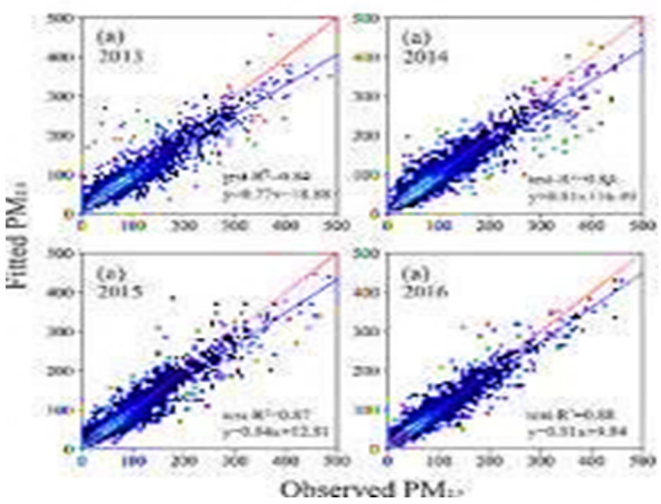
3.4.1. Hạn chế của nghiên cứu

Nghiên cứu thực hiện dự báo doanh thu thông qua mô hình hồi quy tuyến tính và hai mô hình học máy (mô hình rừng ngẫu nhiên và mô hình cây quyết định). Tuy nhiên, trên thực tế còn có nhiều mô hình khác để dự báo doanh thu phù hợp hơn với dữ liệu hiện có. Đồng thời, việc thực hiện điều chỉnh siêu tham số của các mô hình nếu được áp dụng có thể giúp cải thiện hiệu suất. Nghiên cứu cũng cần phân tích sâu hơn về các đặc trưng của dữ liệu để tìm ra những đặc trưng quan trọng ảnh hưởng đến kết quả dự đoán đặc biệt là những đặc thù về lĩnh vực kinh doanh dịch vụ khách sạn, tính chất mùa vụ.

3.4.2. Khuyến nghị

Trong các mô hình được đưa ra trong nghiên cứu, mô hình rừng ngẫu nhiên được đề xuất chọn để thực hiện dự báo. Vì mô hình này cho thấy hiệu suất tốt nhất trên cả hai tập dữ liệu, với khả năng tổng quát hóa cao và lỗi dự đoán thấp. Nhóm tác giả khuyến nghị cần thực hiện thêm

một số nội dung để đạt được kết quả dự báo tốt hơn, cụ thể thực hiện điều chỉnh siêu tham số của mô hình rừng ngẫu nhiên. Mô hình rừng ngẫu nhiên đã cho kết quả tốt (hình 3) nhưng việc điều chỉnh các siêu tham số như số lượng cây, độ sâu của cây,... có thể giúp cải thiện hiệu suất hơn nữa.



Hình 3. Biểu diễn hiệu suất của mô hình rừng ngẫu nhiên so với các mô hình khác

Trong khâu xử lý dữ liệu, dữ liệu ngoại lai cần được kiểm tra do các điểm dữ liệu ngoại lai có thể ảnh hưởng đến hiệu suất của mô hình. Vì vậy, các điểm dữ liệu này cần được loại bỏ hoặc xử lý. Đồng thời, các đặc trưng mới từ dữ liệu hiện có cũng cần được tạo thêm giúp mô hình học được những quy luật phức tạp hơn. Dữ liệu cần thực hiện chuẩn hóa giúp các thuật toán học máy hoạt động hiệu quả hơn. Kỹ thuật xác thực chéo (cross-validation) được sử dụng để đánh giá chính xác hơn hiệu suất của mô hình và tránh overfitting.

4. KẾT LUẬN

Nghiên cứu đã xây dựng mô hình hồi quy tuyến tính dự đoán doanh thu với độ chính xác cao. Mô hình này đã xác định được các yếu tố ảnh hưởng đáng kể đến doanh thu. Kết quả nghiên cứu cho thấy mô hình hồi quy tuyến tính là một công cụ hữu ích trong việc hỗ trợ doanh nghiệp như công tác kế hoạch sản xuất kinh doanh, quản lý kho, kế hoạch bán hàng, đánh giá hiệu quả của các chiến dịch quảng cáo, đưa ra quyết định đầu tư. Đồng thời, nhóm nghiên cứu cũng khuyến nghị nên thường xuyên cập nhật mô hình để đảm bảo tính chính xác và phù hợp với sự thay đổi của thị trường.

TÀI LIỆU THAM KHẢO

[1]. Breiman, L., "Random Forests," *Machine Learning*, 45(1), 5-32, 2001.

[2]. Deepak R., Sathyanarayanan R., Arunnehru J., "Demand and Sales Forecasting Using Random Forest and Linear Regression," in Simic M., Bhateja V., Azar A.T., Lydia E.L. (Eds), *Smart Computing Paradigms: Advanced Data Mining and Analytics, SCI 2024*, Lecture Notes in Networks and Systems, 2025. https://doi.org/10.1007/978-981-96-1981-8_42.

[3]. Fawagreh K., Gaber M.M., Elyan E., "Random Forests: From Early Developments to Recent Advancements," *Systems Science & Control Engineering*, 2(1), 602-609, 2014.

[4]. Pereira L. N., Cerqueira V., "Forecasting hotel demand for revenue management using machine learning regression methods," *Current Issues in Tourism*, 25 (17), 2733-2750, 2022.

[5]. Irfan, M., "Optimizing Publisher Revenue in Digital Marketing Using Decision Trees and Random Forests," *Journal of Digital Market and Digital Currency*, 1(3), 247-266, 2024, https://doi.org/10.47738/jdmdc.v1i3.19.

[6]. Pahmi S., Arianti N. D., Fahrudin I., Amalia A. Z., Ardiansyah E., Rahman S., "Estimated Missing Data on Multiple Linear Regression Using Algorithm EM (Expectation-Maximization) for Prediction Revenue Company," in *2018 International Conference on Computing, Engineering, and Design (ICCED),* IEEE 81-86, 2018. DOI: 10.1109/ICCED.2018.00025.

[7]. Lin Y., Li B., Moiser T. M., Griffel L. M., Mahalik M. R., Kwon J., Alam S. S., "Revenue prediction for integrated renewable energy and energy storage system using machine learning techniques," *Journal of Energy Storage*, 50, 2022.

[8]. Yakin E. A., Kusumawati R., Pagalay U., "Prediction model of revenue restaurant's business using random forest," *Indonesian Journal of Artificial Intelligence and Data Mining*, 6 (2), 252-261, 2023.

[9]. Wang Z., Li J., "Healthcare Revenue Forecasting Based on ARIMA and Linear Regression Models," in *2025 IEEE 3rd International Conference on Image Processing and Computer Applications* (ICIPCA), IEEE, 1320-1325, 2025. DOI: 10.1109/ICIPCA65645.2025.11138220.

AUTHORS INFORMATION

Nguyen Thi Thu Thuy[1], Nguyen Thi Thanh Loan[1], Tran Hung Cuong[2]

[1]School of Economics, Hanoi University of Industry, Vietnam

[2]School of Information and Communication Technology, Hanoi University of Industry, Vietnam