Thu Thập Và Phân Tích Dữ Liệu Vé Máy Bay – Ứng Dụng Học Máy Trong Việc Ra Quyết Định Mua Vé Máy Bay

Đinh Bảo Thy[1,1], Võ Ngọc Anh Thy[1,2], Nguyễn Vũ Thùy Trâm[1,3], Nguyễn Gia Tuấn Anh[1,4], Trần Quốc Khánh[1,5]
[1]Khoa Khoa học Kỹ thuật Thông tin, Trường Đại học Công nghệ Thông tin, Thành phố Hồ Chí Minh, Việt Nam
Đại học Quốc Gia, Thành phố Hồ Chí Minh, Việt Nam
Email: {[1]23521563, [2]23521565, [3]23521617, [4]anhngt, [5]khanhtq}@gm.uit.edu.vn

Tóm tắt

Nhóm nghiên cứu đã tiến hành thu thập và phân tích dữ liệu vé máy bay, sau đó ứng dụng các kỹ thuật học máy để xây dựng một hệ thống dự đoán giá vé theo thời gian. Mục tiêu của hệ thống này là đưa ra khuyến nghị về thời điểm mua vé tối ưu, nhằm hỗ trợ khách hàng đưa ra quyết định thông minh và tiết kiệm chi phí.

1 Giới thiệu

1.1 Tổng quan

Hiện nay, vận tải hàng không đóng vai trò quan trọng trong việc thúc đẩy giao lưu kinh tế và văn hóa giữa các vùng miền trong nước cũng như giữa các quốc gia trên thế giới. Tại Việt Nam, cùng với sự phát triển mạnh mẽ của ngành hàng không, nhu cầu đi lại bằng máy bay ngày càng tăng, kéo theo đó là sự biến động của giá vé tùy theo sự ảnh hưởng của tình hình thị trường, thời điểm đặt vé (các ngày đặc biệt như lễ, Tết,..). Tính cấp thiết của đề tài xuất phát từ nhu cầu tối ưu hóa chi phí đi lại cho khách hàng và đảm bảo được kế hoạch di chuyển thuận lợi trước sự biến động giá vé. Đối với mỗi chuyến bay và hãng hàng không khác nhau, giá vé máy bay sẽ có sự khác biệt riêng. Nhìn chung khi mua vé máy bay càng sớm thì khả năng mua được giá vé rẻ càng cao. Tuy nhiên điều này không phải lúc nào cũng đúng, đặc biệt với các ngày bay đặc biệt hay các hãng hàng không giá rẻ.

Chính vì lí do trên mà nhóm đã quyết định lựa chọn đề tài thu thập và phân tích dữ liệu vé máy bay với mục tiêu đưa ra hệ thống dự đoán giá vé theo thời gian và khuyến nghị về thời điểm mua vé cho khách hàng. Điểm nổi bật trong nghiên cứu của nhóm là việc phát triển một website khuyến nghị có giao diện thân thiện với người dùng, hướng đến thị trường Việt Nam và từ đó mang lại lợi ích thiết thực cho khách hàng khi tìm vé máy bay.

1.2 Giới thiệu bài toán

Bài toán gồm hai phần chính là phân tích dữ liệu và ứng dụng máy học để dự đoán giá vé. Đầu tiên, phân tích dữ liệu thông qua khai phá mối quan hệ giữa giá vé với các thuộc tính khác để đưa ra những quy luật ẩn. Tiếp theo sử dụng các mô hình học máy để đưa ra khuyến nghị về thời điểm đặt vé tối ưu nhất cho chuyến bay. Từ kết quả của hai bước chính, nhóm sẽ xây dựng một website khuyến nghị hỗ trợ người dùng chọn đúng thời điểm đặt vé với chi phí, ngân sách phù hợp ở thị trường Việt Nam. Dữ liệu đầu vào là các thuộc tính cơ bản của một vé máy bay và đầu ra sẽ là giá vé dự đoán để đưa ra được thời điểm đặt vé hợp lý và khuyến nghị nên mua hay chờ.

2 Công trình liên quan

Dự đoán giá vé máy bay là một lĩnh vực nghiên cứu với nhiều cách tiếp cận và phương pháp khác nhau thu hút nhiều sự quan tâm để nâng cao độ chính xác và hiệu suất dự đoán.

Bài báo *"Prediction of Airline Ticket Price Using Machine Learning Method"* (1) của tác giả *Huseyin Korkmaz* đã so sánh hiệu quả dự đoán giá vé bằng nhiều thuật toán học máy khác nhau. Kết quả cho thấy mô hình dự đoán tốt nhất là Gaussian Process Regression kết hợp hàm kernel rational Quadratic có R^2 score là 0.9 trên tập test và có RMSE thấp nhất trong các mô hình so sánh. Mô hình này giúp tăng độ chính xác trong dự đoán, hỗ trợ người dùng chọn được thời điểm mua vé hợp lý.

Ngoài ra, bài báo *"A Holistic Approach on Airfare Price Prediction Using Machine Learning Techniques"* (2) của nhóm tác giả *Theofanis Kalampokas* đã đề xuất một phương pháp dự đoán giá vé toàn diện bằng cách kết hợp ba nhóm mô hình (ML, DL và QML). Các mô hình được đánh giá theo hai hướng là theo điểm đến và theo hãng bay. Kết quả cho thấy mô hình DL đạt độ chính

xác cao và ổn định hơn với R^2 score đạt từ 89% đến 99%.

Bên cạnh đó, bài báo *"A regression model for predicting optimal purchase timing for airline tickets"* (3) của tác giả *William Groves and Maria Gini* đã đề xuất một mô hình hồi quy Partial Least Squares (PLS) để dự đoán thời điểm mua vé tối ưu nhất. Mô hình sử dụng dữ liệu lịch sử giá vé và các đặc trưng trích xuất có thời gian để dự đoán giá trong tương lai và đưa ra chính sách nên mua hay đợi. Kết quả của phương pháp này đã giúp giảm chi phí trung bình đến 12% so với các phương pháp truyền thống như mua ngay lập tức, theo cảm tính...

Cả ba bài báo trên đều góp phần khẳng định vai trò của học máy trong việc dự đoán giá và đưa ra thời điểm mua vé hợp lý. Tuy nhiên, những nghiên cứu này có một số hạn chế khi áp dụng tại thị trường Việt Nam, đặc biệt do bộ dữ liệu sử dụng là của quá khứ, chưa được cập nhật. Vì vậy, nhóm đã thu thập dữ liệu mới và phát triển mô hình dự đoán phù hợp hơn với người dùng ở thị trường Việt Nam.

3 Bộ dữ liệu

3.1 Tổng quan bộ dữ liệu

Bộ dữ liệu được thu thập từ trang web đặt vé uy tín https://www.traveloka.com/vi-vn/flight từ ngày 07/04/2025 đến 10/05/2025.

Cách thu thập dữ liệu: sử dụng thư viện "*selenium*" để thu thập các thành phần cơ bản của một vé máy bay khởi hành từ sân bay Tân Sơn Nhất (SGN), bao gồm hãng bay, mã chuyến, giá vé, thời gian khởi hành, thời gian hạ cánh, ngày khởi hành, ngày hạ cánh, thời gian bay, điểm đến, hành lý xách tay, hành lý ký gửi và ngày crawl dữ liệu.

Dữ liệu sau khi thu thập sẽ được lưu trữ trong một DataFrame, và bộ dữ liệu lưu thông tin các ngày trước ngày khởi hành 20 ngày với kết quả thu thập được 46,826 mẫu và 12 thuộc tính được mô tả như *bảng 1*.

Index	Feature	Description	Type Of Feature
1	brand	Hãng bay	Categorical
2	id	Mã chuyến bay	Categorical
3	price	Giá vé (VND/khách)	Categorical
4	start_time	Thời gian khởi hành	Categorical
5	start_day	Ngày khởi hành	Categorical
6	end_time	Thời gian hạ cánh	Categorical
7	end_day	Ngày hạ cánh	Categorical
8	trip_time	Thời gian bay	Categorical
9	destination	Điểm đến (PQC, DAD, HAN, HPH, CXR, DLI)	Categorical
10	hand_luggage	Hành lý xách tay (kg)	Categorical
11	checked_baggage	Hành lý ký gửi (kg)	Categorical
12	crawl_date	Ngày thu thập dữ liệu	Categorical

Bảng 1. Các thuộc tính của bộ dữ liệu

3.2 Khai phá và phân tích dữ liệu

3.2.1 Tiền xử lý bộ dữ liệu cơ bản

Trong quá trình thu thập dữ liệu, nhóm phát hiện một số bản ghi bị trùng lặp. Sau khi đối chiếu với các lịch bay liền kề, các chuyến bay này không xuất hiện ở cả ngày trước và ngày sau. Do đó, nhóm nhận định đây là lỗi hệ thống web và đã loại bỏ các bản ghi này.

Ngoài ra, hai thuộc tính *'hand_luggage'* và *'checked_baggage'* có một số giá trị bị thiếu do được cập nhật sau. Nhóm xử lý bằng cách duyệt theo từng chuyến bay và thay thế giá trị thiếu bằng giá trị phổ biến nhất.

Tiếp theo, dữ liệu được chuẩn hóa để thuận tiện cho phân tích và đảm bảo tính chính xác. Cụ thể, thuộc tính *'price'* được loại bỏ dấu phẩy và đơn vị 'VND/khách'; các thuộc tính *'hand_luggage'* và *'checked_baggage'* được làm sạch bằng cách loại bỏ các từ khóa mô tả và đơn vị 'kg'. Sau đó, cả ba thuộc tính được chuẩn hóa về định dạng số.

Đồng thời, để mô hình hiểu được ảnh hưởng của thời gian trong ngày đến giá vé, nhóm đã phân loại các khung giờ của hai đặc trưng *'start_hour'* và *'end_hour'* theo 5 khoảng thời gian:

- EarlyMorning (0-3h).
- Morning (3-9h).
- Afternoon (9-15h).
- Evening (15-21h).
- LateNight (21-24h).

Để nâng cao khả năng dự đoán độ biến động của giá vé theo thời gian, nhóm bổ sung hai đặc trưng '*days_left'* và *'is_holiday'*. Thuộc tính *'days_left'* biểu thị số ngày còn lại đến ngày khởi hành, tính bằng hiệu số giữa '*start_day'* và *'crawl_day',* giúp mô hình nhận biết xu hướng tăng giá khi ngày bay đến gần. Bên cạnh đó, ta có thuộc tính *'is_holiday'* dùng để phân loại mức độ đặc biệt của ngày bay, được mã hóa từ 0 đến 3:

- 0: Ngày thường
- 1: Ngày cuối tuần (thứ 6, thứ 7, chủ nhật)
- 2: Ngày cận lễ
- 3: Ngày lễ

Việc mã hóa thuộc tính *'is_holiday'* giúp mô hình học được xu hướng tăng giá vào các dịp cao điểm như lễ hoặc cuối tuần.

Cuối cùng là bước xử lý giá trị ngoại lai, được thực hiện sau khi chia tập train và test để tránh rò rỉ dữ liệu. Kết quả kiểm tra cho thấy các điểm ngoại lai chỉ xuất hiện ở ngưỡng trên, thường rơi vào dịp lễ hoặc sát ngày bay – những thời điểm giá vé

thường tăng cao. Khi so sánh với các chuyến bay tương tự (cùng hãng, điểm đến, days_left), giá vé cũng dao động gần mức ngưỡng trên. Do đó, nhóm quyết định thay các giá trị vượt ngưỡng bằng chính ngưỡng trên, nhằm giảm ảnh hưởng tiêu cực đến mô hình mà vẫn phản ánh đúng xu hướng thực tế.

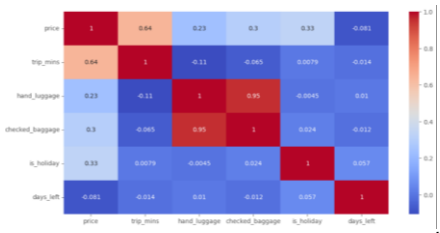
Sau khi hoàn tất trích xuất các đặc trưng, bộ dữ liệu đã được làm sạch, chuẩn hóa, bổ sung với 46,549 mẫu và 11 thuộc tính được mô tả như sau.

Index	Feature	Description	Type of Feature
1	id	Mã chuyến bay	Categorical
2	brand	Hãng bay	Categorical
3	price	Giá vé (VND/khách)	Numeric
4	destination	Điểm đến (PQC, DAD, HAN, HPH, CXR, DLI)	Categorical
5	hand_luggage	Hành lý xách tay (kg)	Numeric
6	checked_baggage	Hành lý ký gửi (kg)	Numeric
7	start_hour	Khung giờ khởi hành	Categorical
8	end_hour	Khung giờ hạ cánh	Categorical
9	trip_mins	Thời gian bay	Numeric
10	is_holiday	Phân loại ngày bay	Categorical
11	days_left	Số ngày trước ngày bay	Numeric

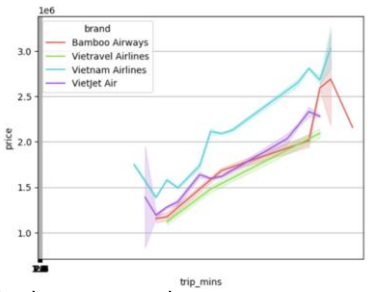
Bảng 2. Các thuộc tính của bộ dữ liệu sau khi tiền xử lý cơ bản

3.2.2 Mối tương quan giữa các đặc trưng với giá vé

Phân tích hệ số tương quan *(hình 1),* cho thấy biến "*trip_mins"* (thời gian bay) có mối tương quan khá cao với *"price"* (giá vé), hệ số đạt 0.64. Vì vậy, để phân tích kỹ hơn về mối liên hệ này, nhóm đã trực quan qua biểu đồ ở *hình 2* và thấy được hai thuộc tính "*price"* và "*trip_mins"* có sự tương quan đồng biến, có nghĩa là khi thời gian bay tăng thì giá vé cũng tăng theo.



Hình 1. Ma trận tương quan giữa các đặc trưng số



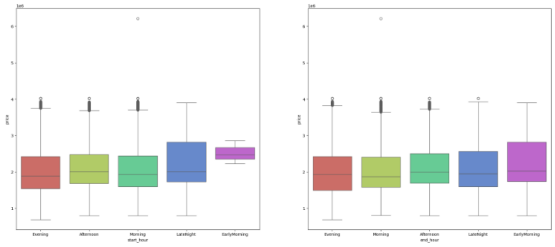
Hình 2. Biểu đồ đường phân bố giá vé và thời gian bay của 4 hãng

Do đó, có thể kết luận "*trip_mins"* là đặc trưng quan trọng có ảnh hưởng rõ rệt với giá vé và cần giữ lại trong quá trình huấn luyện mô hình mô hình dự đoán giá.

Bên cạnh đó, hai đặc trưng *"hand_luggage"* và *"checked_baggage"* có hệ số tương quan rất cao (xấp xỉ 1), cho thấy khả năng đồng biến giữa hai biến này gần như tuyệt đối. Tuy nhiên, khi chạy thử nghiệm mô hình với:

- Đầy đủ cả hai đặc trưng (1).
- Chỉ giữ lại một trong hai đặc trưng (2).

Kết quả đánh giá cho thấy mô hình (1) có hiệu suất tốt hơn. Vì vậy, nhóm quyết định vẫn giữ hai thuộc tính này cho mô hình.



Hình 3. Biểu đồ hộp thể hiện sự phân bố giá vé với thời gian khởi hành và thời gian hạ cánh

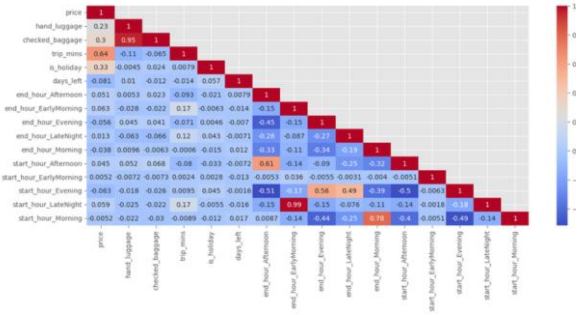
Bên cạnh đặc trưng là hành lý, nhóm cũng xem xét đánh giá yếu tố thời gian khởi hành (*start_hour*) và thời gian hạ cánh (*end_hour*) nhằm đánh giá ảnh hưởng của khung giờ bay đến giá vé. Kết quả phân tích dựa trên biểu đồ hộp giữa từng đặc trưng *(hình 3)* cho thấy:

- Đối với đặc trưng "*start_hour"*
  Phân bố khá tương tự giữa các khung giờ ngoại trừ LateNight có box lớn hơn thể hiện giá vé ở khung giờ này cao hơn các khoảng còn lại. EarlyMorning có phân bố giá hẹp, giá trị trung vị cao nhất cho thấy giá vé khung giờ này ổn định. Ngoài ra giá trị ngoại lai chỉ xuất hiện nhiều ở ba khoảng thời gian Evening, Afternoon và Morning.
- Đối với đặc trưng "*end_hour"*
  Các khoảng thời gian phân bố khá đồng đều và tương đồng nhau.

Để làm rõ hơn về hai đặc trưng này so với giá vé, nhóm đã thực hiện *one-hot-encoding* hai thuộc tính "*start_hour*" và "*end_hour*" và phân tích hệ số tương quan với giá thì ma trận tương quan cho thấy hệ số tương quan thấp (*hình 4*). Vì vậy, nhóm đã chạy thử nghiệm mô hình với hai trường hợp:

- Đầy đủ cả hai đặc trưng (1)
- Loại bỏ cả hai đặc trưng (2)

Kết quả đánh giá sau khi chạy mô hình cho cả hai trường hợp không chênh lệch nhiều (*hình 5*). Vậy nên, nhóm đã quyết định giữ lại hai đặc trưng này.

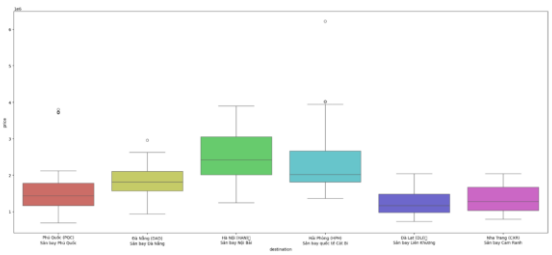


Hình 4. Ma trận tương quan giữa các đặc trưng

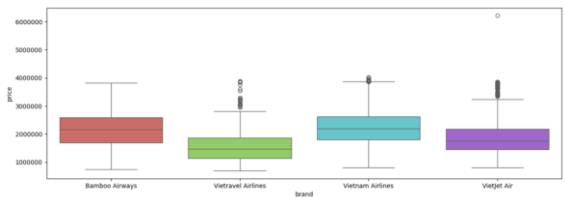
Brand	Bamboo Airways	VietJet Air	Vietnam Airlines	Vietravel Airlines
Model Name				
Mô hình vẫn giữ start_hour và end_hour				
GradientBoostingRegressor	0.909849	0.762464	0.778481	0.723546
Mô hình loại bỏ start_hour và end_hour				
GradientBoostingRegressor	0.904410	0.738583	0.748960	0.723546

Hình 5. Kết quả mô hình của hai trường hợp (1) và (2).

Tiếp theo, thuộc tính "*price"* cũng phụ thuộc vào hai thuộc tính "*destination*" và "*brand*". Vì vậy, để làm rõ hơn, hai biểu đồ hộp về sự phân bố giữa "*price"* và "*destination*" và sự phân bố giữa "*price*" và "*brand*" được trực quan.

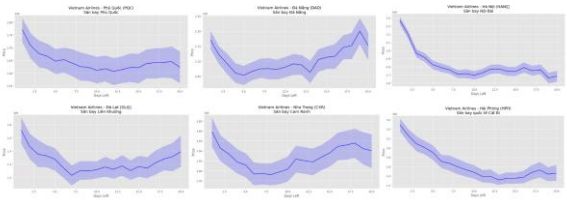


Hình 6. Biểu đồ phân bố giá vé theo từng điểm đến.



Hình 7. Biểu đồ phân bố giá vé theo hãng bay

Từ biểu đồ về sự phân bố giá theo điểm đến, nhóm nhận thấy có sự khác biệt giữa các sân bay lớn (Hà Nội,..) và sân bay nhỏ (Nha Trang,..). Cụ thể hơn, các sân bay lớn có giá trị trung vị cao với chỉ số IQR lớn hơn thể hiện giá vé có sự dao động mạnh. Trong khi đó, sân bay nhỏ có chỉ số IQR hẹp hơn, ít biến động hơn cho thấy giá sẽ ổn định hơn. Ngoài ra, biểu đồ đã thể hiện rõ các giá trị ngoại lai, có một số nơi như Hải Phòng, Phú Quốc thì giá trị này cao hơn hẳn so với giá mặt bằng chung. Điều này có thể sẽ ảnh hưởng đến mô hình.



Hình 8. Biểu đồ thể hiện xu hướng giá dựa vào số ngày còn lại trước khi bay của Vietnam Airlines theo từng điểm đến.

Biểu đồ thể hiện sự phân bố giá theo hãng bay (*Hình 7*) cho thấy giá trị trung vị không chênh lệch nhiều và chỉ số IQR khá tương đương, thể hiện mức giá dao động không quá khác biệt như phân bố theo điểm đến. Vì vậy, khi phân bố giá theo hãng bay sẽ giúp mô hình ít bị nhiễu hơn.

Bên cạnh đó, khi phân tích xu hướng giá theo hãng bay thì nhóm thấy được giá có quy luật tăng hay giảm. Dựa vào *hình 8*, giá vé của hãng Vietnam Airlines ở cả 6 điểm đến đều có xu hướng tăng khi gần đến ngày bay. Ngược lại khi phân tích giá vé theo từng địa điểm sẽ có hạn chế về bộ dữ liệu, cụ thể hơn là những điểm đến như Đà Lạt, Nha Trang có số chuyến ít hơn vì vậy mô hình sẽ có ít dữ liệu của các điểm đến này để học và có thể đưa ra kết quả dự đoán không cao. Chính vì vậy, nhóm đã huấn luyện mô hình theo "*brand"* (hãng bay) thay vì theo "*destination"* (điểm đến) để dự đoán và đưa ra kết quả có độ chính xác cao hơn.

3.2.3 Đánh giá bộ dữ liệu

Sau khi phân tích, nhóm đã tiến hành đánh giá chất lượng bộ dữ liệu cũng như mức độ quan trọng của từng đặc trưng. Cụ thể hơn, bộ dữ liệu đã đáp ứng được năm đặc tính cơ bản: tính đầy đủ, tính cập nhật, tính đa dạng, tính chính xác và tính tin cậy. Ngoài ra, nhóm quyết định giữ lại toàn bộ 11 thuộc tính ban đầu vì chúng đều có tiềm năng quan trọng cho mô hình dự đoán.

4 THỰC NGHIỆM & ĐÁNH GIÁ

4.1 Mô hình

Từ kết quả phân tích dữ liệu, nhóm nhận thấy phần lớn các đặc trưng đầu vào có mối quan hệ phức tạp và không tuyến tính rõ ràng với giá vé (ngoại trừ thuộc tính "*trip_mins*"). Do đó, nhóm lựa chọn áp dụng các mô hình hồi quy phi tuyến tính sau đây:

Model	Algorithm type
AdaBoost Regressor	Boosting family
Bagging Regressor	Boosting family
Gradient Boost Regressor	Boosting family
Decision Tree Regressor	Tree based
Random Forest Regressor	Tree based
Extra Trees Regressor	Tree based

Bảng 3. Các mô hình phù hợp

AdaBoost Regressor (4) thuộc nhóm thuật toán 'Boosting', là mô hình kết hợp tuần tự các mô hình phân loại yếu để điều chỉnh trọng số cho các quan sát, từ đó tạo nên một mô hình tổ hợp mạnh giúp giảm sai lệch và phương sai trong dữ liệu huấn luyện. Tuy nhiên, AdaBoost khá nhạy cảm với nhiễu và có khả năng bị overfitting khi dữ liệu có số lượng đặc trưng lớn. *Bagging Regressor* (5) có cách tiếp cận giống như AdaBoost, nhưng các mô hình học yếu trong Bagging được huấn luyện song song và độc lập. Dự đoán cuối cùng là trung bình cộng kết quả từ các mô hình con. Nhược điểm của thuật toán là dễ nhạy cảm với nhiễu và hiệu suất giảm khi dữ liệu có độ phức tạp cao. Thuật toán cuối cùng từ nhóm Boosting là *Gradient Boost Regressor* (6). Thuật toán này kết hợp nhiều mô hình dự đoán yếu theo cách tuần tự nhằm tối ưu hàm mất mát với số lần lặp tối thiểu. Tuy nhiên, trên các bộ dữ liệu lớn, mô hình có thể dễ rơi vào cực tiểu cục bộ, gây ra hiện tượng underfitting.

Decision Tree Regressor (7) là mô hình phân tách dữ liệu dựa trên các đặc trưng chính, mỗi nút sẽ đặt câu hỏi về một đặc trưng, tiếp tục chia nhỏ dữ liệu đến khi đến các nút lá với dự đoán cuối cùng. Nhược điểm là mô hình có thể trờ nên quá lớn và thiếu ổn định khi số lượng đặc trưng và dữ liệu tăng cao. *Random Forest Regressor* (8) là sự kết hợp của nhiều cây quyết định được xây dựng ngẫu nhiên. Dự đoán được đưa ra bằng cách lấy trung bình hoặc bỏ phiếu đa số từ các cây. Mô hình này khắc phục phần nào hạn chế của cây quyết định đơn lẻ và thường cho hiệu suất cao hơn. *Extra Trees Regressor* (9) là một thuật toán tương tự với Random Forest, tuy nhiên quá trình lựa chọn giá trị phân tách và quy tắc quyết định thì được thực hiện một cách ngẫu nhiên, giúp tăng tốc độ huấn luyện và cải thiện khả năng tổng quát hóa.

Từ các phân tích trên, nhóm tiến hành áp dụng 6 mô hình đã đề cập để dự đoán giá vé máy bay và phân tích so sánh kết quả.

4.2 Độ đo đánh giá

Để định lượng mức độ hiệu quả của mô hình, nhóm sử dụng 3 chỉ số:

Độ đo R^2 (R^2 Score) cho biết phần trăm phương sai của biến phụ thuộc (biến giá vé) có thể được giải thích bởi các biến độc lập trong mô hình. Nói cách khác, nó thể hiện mức độ giá trị dự đoán $\hat{y}_i$ xấp xỉ giá trị thực tế y_i. Kết quả R^2 càng gần 1 nghĩa là mô hình càng dự đoán chính xác:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

	Bamboo Airways	Vietnam Airlines	VietJet Air	Viettravel Airlines
AdaBoost	0.8765	0.7181	0.6824	0.6532
Bagging	0.8819	0.7404	0.7313	0.6542
GradientBoosting	**0.9053**	**0.7691**	**0.7543**	**0.7235**
DecisionTree	0.8727	0.7315	0.7248	0.6150
RamdomForest	0.8903	0.7482	0.7494	0.6488
ExtraTrees	0.8762	0.7332	0.7256	0.6178

Bảng 4. Kết quả mô hình trên R^2 Score

Độ đo MAE (Mean Absolute Error) cho biết giá trị trung bình của sai số tuyệt đối giữa giá trị dự đoán $\hat{y}_i$ và giá trị thực tế y_i của bộ dữ liệu. MAE cho biết trực tiếp mức độ sai lệch trung bình của mô hình theo đơn vị của biến mục tiêu. Kết quả của MAE càng thấp thì mô hình càng dự đoán tốt:

$$MAE = \frac{1}{n} * \sum |y_i - \hat{y}_i|$$

	Bamboo Airways	Vietnam Airlines	VietJet Air	Viettravel Airlines
AdaBoost	11.001	16.1846	14.4888	18.5233
Bagging	9.5901	12.4926	12.4213	17.5011
GradientBoosting	**8.5731**	**12.1597**	**12.3612**	**15.4177**
DecisionTree	9.8340	12.6003	12.6238	17.5175
RamdomForest	9.5090	12.4102	12.3684	17.1729
ExtraTrees	9.8868	12.5927	12.5879	17.4471

Bảng 5. Kết quả mô hình trên MAE

Độ đo MAPE (Mean Absolute Percentage Error) cho biết phần trăm dự đoán lỗi của mô hình. Kết quả MAPE càng thấp thì mô hình càng phù hợp:

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

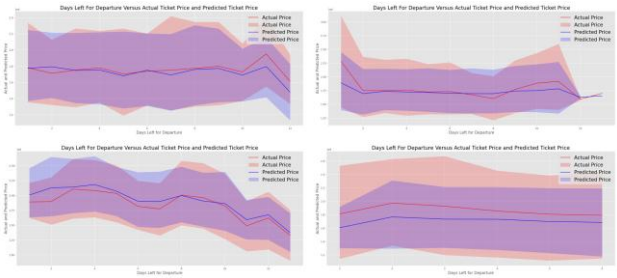
	Bamboo Airways	Vietnam Airlines	VietJet Air	Viettravel Airlines
AdaBoost	206785	289256	307256	278160
Bagging	185987	236955	271727	274564
GradientBoosting	**174936**	**229862**	**267609**	**239927**
DecisionTree	188767	239435	275491	275971
RamdomForest	184460	235651	270198	269020
ExtraTrees	189915	239354	274330	276073

Bảng 6. Kết quả mô hình trên MAPE

Lựa chọn mô hình

Sau quá trình thử nghiệm và đánh giá các mô hình khác nhau dựa trên 3 độ đo, mô hình *Gradient Boosting Regressor* cho kết quả tốt vượt trội.

Dựa trên kết quả trên *hình 9,* mô hình cho thấy khả năng dự đoán tốt các dao động trong giá vé. Vì vậy, *Gradient Boosting Regressor* được lựa chọn thực hiện bài toán và sử dụng cho hệ khuyến nghị.



Hình 9. Kết quả dự đoán của mô hình Gradient Boosting Regressor với dữ liệu thực tế của hãng bay BamBoo Airways, VietJet Air, VietNam Airlines, Viettravel Airlines theo thứ tự từ trái sang phải

5 ỨNG DỤNG THỰC TIỄN

5.1 Hệ khuyến nghị

Hệ khuyến nghị của bài toán chia làm 2 giá trị:

- 0: nên chờ
- 1: nên mua

Quyết định khuyến nghị đưa ra bằng cách so sánh giá vé thực tế với ngưỡng giá (threshold) được tính toán.

Quá trình tính toán ngưỡng:

Bước 1: Dự đoán giá vé chuyến bay các ngày tiếp theo bằng mô hình Gradient Boosting Regressor đã được huấn luyện:

$$P = \{p_1, p_2, p_3, \dots, p_n\}$$

Bước 2: Chọn 10% giá vé thấp nhất (k là số lượng vé):

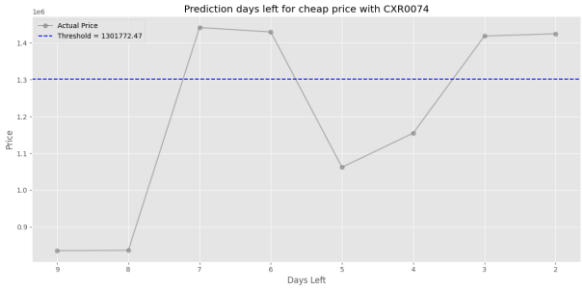
$$P_{low} = \{p_1, p_2, p_3, \dots, p_k\}$$

Bước 3: Tính threshold là giá trị trung bình của P_{low}:

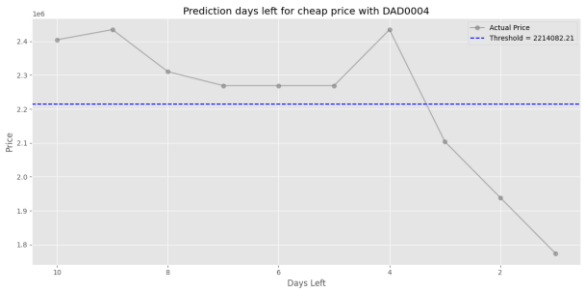
$$threshold = \frac{1}{k}\sum_{i=1}^{k} p_i$$

Dựa vào kết quả của threshold hệ thống sẽ đưa ra khuyến nghị phù hợp:

- Như minh họa trong *hình 10,* khi giá vé lớn hơn ngưỡng, mô hình sẽ khuyến nghị khách hàng nên chờ (label = 0).
- Tương tự trong *hình 11,* với giá vé bé hơn ngưỡng, khuyến nghị mô hình đưa ra là nên mua (label = 1).



Hình 10. Ngưỡng chọn mua của chuyến bay CXR0074



Hình 11. Ngưỡng chọn mua của chuyến bay DAD0004

5.2 Triển khai giao diện tương tác

Để đưa mô hình dự đoán giá vé máy bay đến gần hơn với người dùng, nhóm đã xây dựng một web demo trực quan, tích hợp các chức năng chính cho phép người dùng tương tác và nhận khuyến nghị một cách dễ dàng.

Quy trình hoạt động:

Bước 1: Nhập dữ liệu

Người dùng nhập thông tin cần thiết thông qua giao diện *(hình 12),* bao gồm:

- Điểm đến
- Ngày/ giờ bay mong muốn
- Hãng bay/ Hành lý (tùy nhu cầu)

Bước 2: Thu thập và xử lý dữ liệu

Hệ thống tự động thu thập dữ liệu từ trang web Traveloka dựa trên thông tin người dùng đã nhập và lọc theo tiêu chí của khách hàng.

Sau đó, dữ liệu qua tiền xử lý sẽ đưa vào chạy mô hình dự đoán đã huấn luyện từ trước.

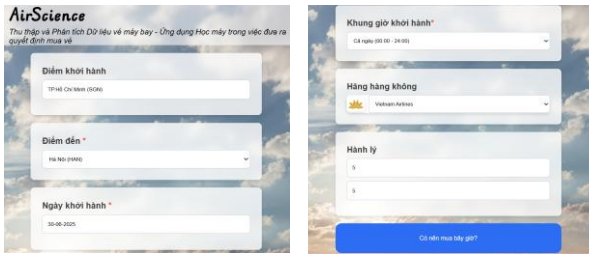
Bước 3: Dự đoán và khuyến nghị

Hệ thống sử dụng kết quả từ mô hình để in ra khuyến nghị phù hợp *(hình 13)*. Khuyến nghị sẽ dựa trên phân tích giá vé dự kiến và ngưỡng được tính toán, giúp người dùng quyết định đây có phải thời điểm thích hợp để mua vé không.

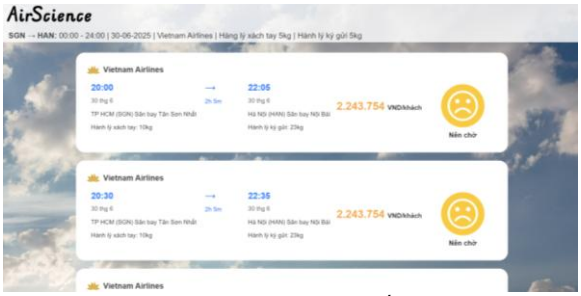
6 KẾT LUẬN

6.1 Kết luận

Nghiên cứu đã thành công trong việc xây dựng một mô hình dự đoán giá vé máy bay, cung cấp cái



Hình 12. Giao diện nhập dữ liệu của website



Hình 13. Giao diện các dự đoán và khuyến nghị của website

nhìn tổng quan về các yếu tố ảnh hưởng đến biến động giá.

Mặc dù còn tồn tại một số hạn chế, nhóm sẽ đề xuất các giải pháp khả thi trong tương lai, từ đó mang lại giá trị thiết thực cho người dùng trong việc tối ưu hóa chi phí di chuyển.

6.2 Hạn chế

Nghiên cứu này còn một số hạn chế do giới hạn về tài nguyên và thời gian:

Bộ dữ liệu chỉ bao gồm thông tin các chuyến bay ***trước ngày khởi hành 20 ngày*** và di chuyển từ ***TP HCM (SGN) Sân bay Tân Sơn Nhất đến 6 địa điểm*** tại Việt Nam. Điều này cho thấy kết quả phân tích chỉ phù hợp với các dự đoán ngắn hạn và không thể mô tả toàn bộ mạng lưới đường bay nội địa.

Bên cạnh đó, trong quá trình thu thập dữ liệu, nhóm nghiên cứu đã gặp trở ngại lớn là ***chưa thể tự động hóa*** hoàn toàn quy trình do vấn đề về chống bot và reCAPTCHA. Việc giải quyết vấn đề này là cần thiết để đảm bảo tính liên tục và hiệu quả của hệ thống khi triển khai.

6.3 Hướng phát triển

Để khắc phục các hạn chế đã nêu, nhóm nghiên cứu đã đưa ra một số giải pháp chiến lược:

Đầu tiên, chúng tôi sẽ mở rộng quy mô và phạm vi của bộ dữ liệu. Điều này bao gồm việc tăng cường số lượng điểm đến cũng như tổng số mẫu dữ liệu chuyến bay. Với một bộ dữ liệu đủ lớn và đa dạng, mô hình dự đoán sẽ có khả năng không chỉ phân tích xu hướng giá vé theo hãng hàng không mà còn chia nhỏ phân tích theo từng địa điểm đến cụ thể của mỗi hãng. Điều này sẽ mang lại cái nhìn chi tiết và chính xác hơn về biến động giá, giúp đưa ra các dự đoán tinh vi hơn.

Thứ hai, nhóm sẽ tập trung vào việc giảm thiểu độ trễ trong quá trình dự đoán. Mục tiêu là rút ngắn đáng kể thời gian từ khâu thu thập dữ liệu đến khi trả về kết quả dự đoán. Song song với đó, một ưu tiên hàng đầu là giải quyết vấn đề chặn bot và reCAPTCHA. Việc tự động hóa hoàn toàn quy trình thu thập dữ liệu là then chốt để đảm bảo tính kịp thời và hiệu quả của hệ thống.

Cuối cùng, phương pháp học tăng cường (Reinforcement Learning) có thể áp dụng vào hệ thống để tự học và liên tục tối ưu hóa các khuyến nghị dựa trên so sánh giữa kết quả dự đoán với dữ liệu thực tế thu được. Điều này sẽ giúp mô hình ngày càng trở nên thông minh và chính xác hơn theo thời gian, thích nghi với các thay đổi của thị trường.

TÀI LIỆU THAM KHẢO

1. *Prediction of Airline Ticket Price Using Machine Learning Method.* **Korkmaz, Huseyin.** 2024, Journal of Transportation and Logistics, Vol. 9.

2. **T. Kalampokas, K. Tziridis, N. Kalampokas, A. Nikolaou, E. Vrochidou and G. A. Papakostas.** A Holistic Approach on Airfare Price Prediction Using Machine Learning Techniques. *IEEE Access.* 2023, pp. 46627-46643.

3. **Groves, William and Gini, Maria.** *A regression model for predicting optimal purchase timing for airline tickets.* s.l. : Retrieved from the University Digital Conservancy, 2011.

4. *Experiments with a new boosting algorithm.* **Freund, Y. and Schapire, R. E.** Proceedings of the 13th International Conference on Machine Learning : s.n., 1996. pp. 148–156.

5. *Bagging predictors.* **Breiman, L.** 2, 1996, Machine Learning, Vol. 24, pp. 123-140.

6. *Greedy function approximation: A gradient boosting machine.* **Friedman, J. H.** 5, 2001, Annals ò Statistics, Vol. 29, pp. 1189-1232.

7. *Classification and regression trees.* **Gordon, A. D., et al.** 3, 1984, Biometrics, Vol. 40, p. 874.

8. *Random forests.* **Breiman, L.** 2001, Machine Learning, Vol. 45, pp. 5-32.

9. *Extremely randomized trees.* **Geurts, P. and Ernst, D. and Wehenkel, L.** 1, 2006, Machine Learning, Vol. 63, pp. 3-42.