Ứng dụng học máy để dự đoán giá vé chuyến bay

Nguyễn Ngọc Thanh Sang
Khoa Khoa học và Kỹ thuật thông tin
Trường đại học Công nghệ thông tin
Đại học Quốc gia Thành phố Hồ Chí Minh
21522544@gm.uit.edu.vn

Phan Cả Phát
Khoa Khoa học và Kỹ thuật thông tin
Trường đại học Công nghệ thông tin
Đại học Quốc gia Thành phố Hồ Chí Minh
21520389@gm.uit.edu.vn

Abstract*—Hệ thống Dự đoán Giá vé Chuyến bay là một giải pháp nhằm dự báo chính xác giá vé máy bay, cung cấp cho khách du lịch những hiểu biết có giá trị để lập kế hoạch và ra quyết định tốt hơn. Ngày nay, giá vé máy bay có thể thay đổi linh hoạt cho cùng một chuyến bay. Từ góc độ khách hàng, họ muốn tiết kiệm tiền nên chúng tôi đã đề xuất một mô hình dự đoán giá vé gần đúng. Hệ thống này sử dụng các thuật toán học máy và dữ liệu lịch sử của các chuyến bay để đưa ra dự đoán giá vé gần chính xác. Hệ thống này sử dụng một tập dữ liệu khổng lồ hơn 300.000 điểm dữ liệu bao gồm giá vé và lịch sử chuyến bay, bao gồm các yếu tố như ngày đi, điểm đến, hãng hàng không, thời gian khởi hành và nhiều tham số liên quan khác. Bằng cách phân tích dữ liệu này bằng các kỹ thuật học máy trên, hệ thống sẽ tìm hiểu các mô hình và mối quan hệ, cho phép đưa ra dự đoán đáng tin cậy về giá vé chuyến bay trong tương lai. Một tập hợp các thuật toán học máy, bao gồm các mô hình dựa trên hồi quy như Linear Regression, Decision Tree Regression và Lasso Regression, được sử dụng để nắm bắt các mẫu dữ liệu và mối quan hệ phức tạp trong dữ liệu. Hệ thống này sẽ cung cấp cho mọi người ý tưởng về xu hướng giá cả và cũng cung cấp giá trị dự đoán của giá mà họ có thể kiểm tra trước khi đặt chuyến bay để tiết kiệm tiền. Loại hệ thống hoặc dịch vụ này có thể được cung cấp cho khách hàng thông qua các công ty đặt vé máy bay để giúp họ đặt vé.

I. Giới thiệu

Mọi người đều biết rằng những ngày nghỉ luôn đòi hỏi một kỳ nghỉ rất cần thiết và việc lên kế hoạch cho hành trình du lịch trở thành một công việc tốn nhiều thời gian. Hoạt động kinh doanh hàng không thương mại đã phát triển vượt bậc và trở thành một thị trường được quản lý bởi sự phát triển của Internet và thương mại điện tử trên toàn thế giới. Do đó, để quản lý doanh thu của hãng hàng không, các chiến lược khác nhau như lập hồ sơ khách hàng, tiếp thị tài chính và các yếu tố xã hội được sử dụng để thiết lập cho giá vé. Khi đặt vé trước nhiều tháng, giá vé máy bay thường hợp lý nhưng khi đặt vé gấp, giá vé thường cao hơn. Tuy nhiên, số ngày/giờ cho đến khi khởi hành không phải là yếu tố duy nhất quyết định giá vé máy bay mà còn có rất nhiều yếu tố khác. Khách hàng cảm thấy khá khó khăn để có được một mức giá vé hoàn hảo và thấp nhất do phương pháp định giá phức tạp của ngành hàng không. Các công nghệ và mô hình dựa trên Machine Learning và Deep Learning đã được tạo ra để vượt qua thách thức này và nghiên cứu đáng kể cũng đang được thực hiện. Đồ án này thực hiện về Hệ thống dự đoán giá vé chuyến bay dựa trên Học máy sử dụng Linear Regression, Decision Tree Regression và Lasso Regression để dự đoán giá vé máy bay. Các đặc trưng khác nhau ảnh hưởng đến giá cả cũng được nghiên cứu cùng với phân tích thử nghiệm của hệ thống. Phần II bao gồm phần tổng quan về bộ dữ liệu, xem xét các phân tích sơ bộ về bộ dữ liệu thu thập được. Sự ảnh hưởng của từng đặc trưng đến kết quả cũng được xem xét. Trong Phần III, các mô hình máy học được đề xuất được mô tả chi tiết cùng với quy trình làm việc và các tính năng của nó. Trong Phần IV, các kết quả cũng như những so sánh khác nhau giữa các mô hình sẽ được xem xét dựa trên các thử nghiệm tinh chỉnh mô hình. Phần V đưa ra phân tích lỗi còn tồn tại trong quá trình huấn luyện và đề ra hướng phát triển. Trong Phần VI sẽ là kết luận về những kết quả đạt được trong quá trình thực hiện đồ án.

II. Dữ liệu

Chúng tôi đã thu thập tập dữ liệu Flight Price Predicion từ Kaggle, một nền tảng trực tuyến cho cộng đồng Machine Learning và Khoa học dữ liệu. Tập dữ liệu Flight Price bao gồm thông tin về lịch sử các chuyến bay được đặt từ trang web Easemytrip dành cho các chuyến bay giữa 6 thành phố đô thị hàng đầu của Ấn Độ. Dữ liệu được thu thập trong 50 ngày, từ ngày 11 tháng 2 đến ngày 31 tháng 3 năm 2022 để đảm bảo sự đa dạng và phản ánh chính xác của thị trường chuyến bay. Có 300153 điểm dữ liệu và 11 đặc trưng trong tập dữ liệu đã được làm sạch. Dữ liệu chuyến bay bao gồm cả dữ liệu định tính (Categorical Data) và dữ liệu định lượng (Numerical Data). Có 8 đặc trưng định tính bao gồm các cột như Airline, Flight, Source-city, Destination-city, Departure-time, Arrival-time, Stops, Class. Còn lại 3 dữ liệu định lượng bao gồm các cột Duration, Days-left, Price.

Trước khi đưa vào mô hình, chúng tôi đã thực hiện phân tích sơ bộ tập dữ liệu và thu được một số thống kê về các đặc trưng định lượng và định tính. Về Thời lượng chuyến bay (Duration), thời lượng trung bình của chuyến bay là khoảng 12.22 giờ, với độ lệch chuẩn là khoảng 7.19 giờ. Trong đó, thời lượng chuyến bay ngắn nhất và dài nhất lần lượt là 0.83 giờ và 49.83 giờ. Về Số ngày còn lại trước khi bay (Days-left), số ngày còn lại trung bình là khoảng 26 ngày, với độ lệch chuẩn là 13.56 ngày. Số ngày còn lại ít nhất là 1 ngày và nhiều nhất là 49 ngày. Về giá của chuyến bay, giá trung bình là khoảng 20,889.66 rupee, giá thấp nhất là 1,105 rupee và cao nhất là 123,071 rupee. Giá chuyến bay có độ lệch chuẩn cao, khoảng 22,697.77 rupee. Đặc trưng Duration có sự phân bố đều và sự biến động lớn. Đặc trưng Days-left và Price có sự phân tán

cao, ngoài ra thì một số giá trị của Price rất cao so với giá trị trung bình, đó là các giá trị ngoại lai.

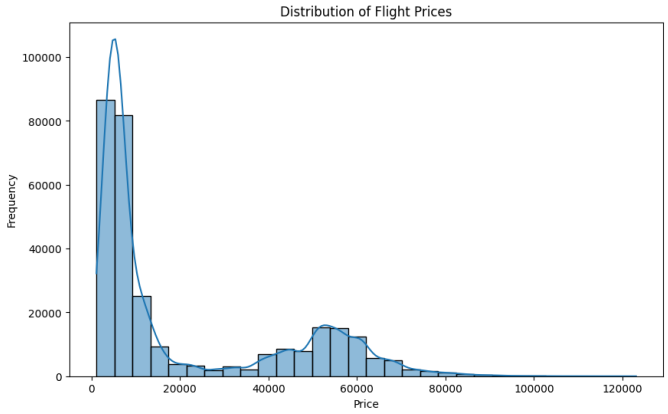


Fig. 1: Biểu đồ phân phối giá vé .

Biểu đồ cột đã giúp chúng tôi phân tích sự phân phối của một số biến phân loại như hãng hàng không, thành phố xuất phát, số điểm dừng và loại ghế. Trong số tất cả các hãng hàng không trong tập dữ liệu, Vistara có số lượng chuyến bay nhiều nhất, trong khi SpiceJet có số chuyến bay ít nhất. Hai thành phố có số lượng chuyến bay xuất phát nhiều nhất là Delhi và Mumbai. Số lượng chuyến bay giữa các thành phố không có sự chênh lệch nhau quá lớn. Trong khi đó, đối với thuộc tính Loại ghế (Classes), hạng vé phổ thông có số lượng hơn gấp đôi so với hạng vé còn lại là thương gia.

Fig. 2: Biểu đồ phân phối loại ghế

Bằng cách sử dụng heatmap, chúng tôi đã xác định được mối liên hệ giữa các biến. Trong đó, sự tăng trưởng của giá vé và thời lượng chuyến bay có mối liên quan tích cực. Thêm vào đó số ngày còn lại của chuyến bay cũng có sự tương quan âm, tức là số ngày còn lại của chuyến bay sẽ tỉ lệ nghịch với giá vé. Vì vậy, chúng tôi đã xem xét tới tự ảnh hưởng tới giá vé bởi các thuộc tính khác.

Đầu tiên, chúng tôi xem xét đến sự ảnh hưởng của thời gian đặt vé (Days left) ảnh hưởng như thế nào đến giá vé. Chúng tôi nhận thấy rằng, khi thời gian đặt vé càng gần thì giá vé sẽ càng giảm. Điều này có thể được giải thích bởi vì nhiều hãng hàng không sử dụng chiến lược giá động, có nghĩa là giá vé thay đổi dựa trên nhiều yếu tố, chẳng hạn như khi thời gian còn lại ít, giá vé có thể được điều chỉnh để thu hút khách hàng. Bên cạnh đó, có thể xem như là một ưu đãi đối với những người muốn đặt vé ngay lập tức, hãng hàng không có thể cung cấp ưu đãi để tăng khả năng bán hết vé trước khi khởi hành chuyến bay.

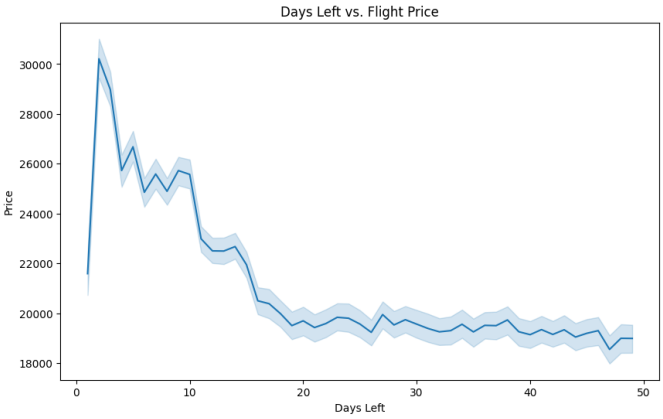


Fig. 3: Ảnh hưởng của thời gian đặt vé lên giá vé.

Bên cạnh đó, chúng tôi cũng xem xét đến ảnh hưởng của thời gian khởi hành và thời gian hạ cánh của chuyến bay. Khi quan sát biểu đồ chúng tôi nhận thấy rằng, có một sự tương đồng giữa thời gian khởi hành và thời gian hạ cánh của chuyến bay khi tác động lên giá vé là những khoảng thời gian vào sáng sớm, buổi tối và đêm thường có giá cao, tuy nhiên, giá của những chuyến bay vào đêm muộn lại có giá rất thấp. Một số lý do có thể lý giải cho hiện tượng này là vì khách hàng có nhu cầu cao vào thời điểm thuận lợi. Thời điểm sáng sớm và buổi tối thường là thời điểm thuận lợi cho nhiều hành khách do phù hợp với kế hoạch của họ, do đó nhu cầu đặt vé có thể tăng cao kéo theo giá vé. Ngược lại, vào đêm muộn, nhu cầu thường thấp hơn vì nhiều người không ưa việc đi lại vào thời điểm này.

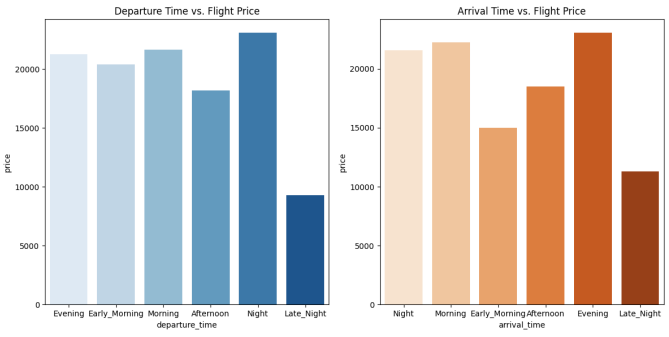


Fig. 4: Ảnh hưởng của thời gian chuyến bay đến giá vé

Một yếu tố khác cũng có ảnh hưởng đến giá vé đó là thời lượng chuyến bay. Mặc dù biểu đồ thể hiện rằng giá vé có sự dao động lớn, nhưng nhìn chung thì giá vé có sự tăng dần trong khoảng thời gian bay dưới 20 và sau đó giảm dần vào đoạn 20h-35h và sau đó tăng lại.

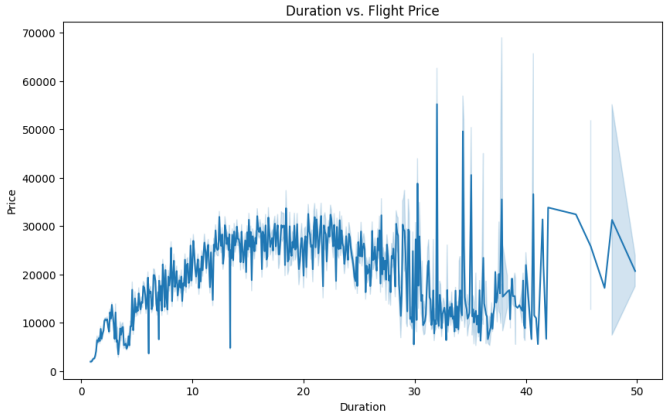


Fig. 5: Ảnh hưởng của thời lượng chuyến bay đến giá vé.

Một yếu tố nữa cũng không kém phần quan trọng là loại vé. Hạng vé thương gia có sự phân bố chủ yếu trên mức giá vé trung bình, tuy nhiên số lượng vé chủ yếu lại rơi vào hạng phổ thông.

Fig. 6: Ảnh hưởng của hạng ghế đến giá vé

III. Phương pháp thực hiện

A. Tiền xử lý dữ liệu

Quá trình chuẩn bị dữ liệu cho việc mô hình hóa bao gồm các bước như làm sạch dữ liệu, loại bỏ các giá trị ngoại lai, chuyển đổi kiểu dữ liệu, và tạo các thuộc tính mới. Tuy nhiên, vì đây là tập dữ liệu đã được làm sạch trước đó, nên trong đồ án này, chúng tôi sẽ tiến hành 2 bước: Biến đổi dữ liệu các biến phân loại và sau đó là chuẩn hóa toàn bộ dữ liệu. Khi thực hiện biến đổi dữ liệu, chúng tôi đã sử dụng công cụ LabelEncoder trong thư viện scikit-learn của Python để chuyển đổi giá trị của các biến phân loại thành giá trị số để có thể được sử dụng trong quá trình huấn luyện mô hình. Lý do chúng tôi chọn công cụ này là vì LabelEncoder giữ nguyên không gian lưu trữ so với OneHotEncoder. Thực tế là khi chúng tôi thử biến đổi dữ liệu bằng OneHotEncoder, số chiều của dữ liệu đã tăng lên đáng kể. Sau khi đã sử dụng LabelEncoder để mã hóa dữ liệu và quan sát kết quả thông qua hàm describe(), có một số điểm quan trọng khiến chúng tôi quyết định chuẩn hóa dữ liệu bằng công cụ StandardScaler. Đây là một phương pháp chuẩn hóa dữ liệu thuộc về nhóm biến đổi dữ liệu sao cho nó có mean (trung bình) bằng 0 và độ lệch chuẩn bằng 1. Điều này giúp giảm ảnh hưởng của outliers (giá trị ngoại lai) và tạo ra phân phối có đặc tính chuẩn hóa. Lý do mà chúng tôi chọn phương pháp chuẩn hóa này là vì Phân phối của mỗi đặc trưng có vẻ khá phân tán với độ lệch chuẩn tương đối lớn. Có sự chênh lệch lớn giữa giá trị max và percentiles 75%, đặc biệt là ở đặc trưng 'price'. Điều này có thể là một dấu hiệu của sự xuất hiện của outliers trong dữ liệu. Tuy nhiên, vì trong số các mô hình máy học mà chúng tôi lựa chọn có một số mô hình không bị ảnh hưởng bởi outliers, nên sự thử nghiệm trên cả MinMaxScaler là cần thiết. Việc sử dụng công cụ này nhằm giữ lại đặc điểm của các outliers nhiều hơn vì dữ liệu chỉ bị thu nhỏ về một khoảng cố định mà không loại bỏ đi điểm dữ liệu nào.

Sau đó chúng tôi tiến hành chia dữ liệu ban đầu thành tập train và tập test với tỉ lệ là 70:30.

B. Lựa chọn thang đo

Đánh giá hiệu suất của các mô hình dự đoán là rất quan trọng trong nhiều lĩnh vực như dự báo giá cả. So với các phương pháp truyền thống, các mô hình học máy thường thể hiện độ chính xác cao hơn. Để đánh giá mức độ chính xác của các mô hình này một cách định lượng, nhóm sử dụng một số thang đo hồi quy từ module sklearn.metrics. Dưới đây là mô tả chi tiết về các thang đo quan trọng này:

Mean Absolute Error (MAE): Được tính bằng cách lấy trung bình giữa giá trị tuyệt đối của hiệu giữa giá trị dự đoán ($\hat{y}$) và giá trị thực tế (y), chia cho tổng số điểm dữ liệu (n).

$$\text{MAE} = \frac{1}{n}\sum|y - \hat{y}|$$

Giá trị MAE càng thấp, mô hình càng hiệu quả.

Mean Square Error (MSE): MSE liên quan đến việc bình phương hiệu giữa giá trị dự đoán và giá trị thực tế trước khi lấy trung bình, tạo ra một độ đo đánh phạt lỗi lớn hơn.

$$\text{MSE} = \frac{1}{n}\sum(y - \hat{y})^2$$

Giá trị MSE thấp hơn cho thấy độ chính xác của mô hình tốt hơn.

Root Mean Square Error (RMSE): RMSE được tính bằng cách lấy căn bậc hai của giá trị trung bình của bình phương hiệu giữa giá trị dự đoán và giá trị thực tế.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum(y - \hat{y})^2}$$

Giá trị RMSE, lớn hơn so với MAE, giúp đánh giá tốt mức độ điều chỉnh của biến độc lập trong mô hình, với giá trị thấp hơn là hiệu suất tốt hơn.

R2-Score (Hệ số xác định): R2-Score đo lường tỷ lệ phương sai của biến phụ thuộc (giá trị thực tế, y) có thể dự đoán được từ biến độc lập (giá trị dự đoán, $\hat{y}$).

$$R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}$$

index	airline	flight	source_city	departure_time	stops	arrival_time	destination_city	class	duration	days_left	price
count	300153.0	300153.0	300153.0	300153.0	300153.0	300153.0	300153.0	300153.0	300153.0	300153.0	300153.0
mean	3.10	1088.34	2.58	2.42	0.28	3.07	2.59	0.69	12.22	26.00	20889.66
std	1.83	426.69	1.75	1.75	0.67	1.74	1.74	0.46	7.19	13.56	22697.77
min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.83	1.0	1105.0
25%	1.0	783.0	1.0	1.0	0.0	2.0	1.0	0.0	6.83	15.0	4783.0
50%	3.0	1142.0	2.0	2.0	0.0	4.0	3.0	1.0	11.25	26.0	7425.0
75%	5.0	1486.0	4.0	4.0	0.0	5.0	4.0	1.0	16.17	38.0	42521.0
max	5.0	1560.0	5.0	5.0	2.0	5.0	5.0	1.0	49.83	49.0	123071.0

Bảng Mô Tả Dữ liệu

R2-Score nằm trong khoảng từ 0 đến 1, giá trị gần 1 cho thấy mô hình điều chuẩn tốt hơn. Nó giúp đánh giá mức độ tốt của mô hình so với các mô hình khác.

textbfMean Absolute Percentage Error (MAPE): MAPE tính giá trị trung bình của phần trăm khác biệt giữa giá trị dự đoán và giá trị thực tế, cung cấp thông tin về độ chính xác tương đối của mô hình.

$$\text{MAPE} = \frac{1}{n} \sum \left(\frac{|y - \hat{y}|}{|y|} \right) \times 100$$

MAPE thấp hơn cho thấy độ chính xác dự đoán tốt hơn.

Tóm lại, việc sử dụng các thang đo như MAE, MSE, RMSE, R2-Score và MAPE giúp đánh giá toàn diện hiệu suất của các mô hình dự đoán, hướng dẫn việc lựa chọn mô hình chính xác và đáng tin cậy nhất cho bộ dữ liệu cụ thể.

C. Lựa chọn mô hình máy học

Sau khi tiến hành khám phá dữ liệu, bước tiếp theo là sử dụng thuật toán máy học và phát triển mô hình. Nhóm sẽ sử dụng các kỹ thuật hồi quy trong học máy có giám sát. Mối quan hệ giữa các biến phụ thuộc và độc lập được đặc trưng bởi các mô hình hồi quy. Đây là những phương pháp học máy nhóm sẽ sử dụng trong đồ án:

Linear Regression: Linear Regression là một mô hình học máy đơn giản, dựa trên giả định rằng có một mối quan hệ tuyến tính giữa các biến độc lập và biến phụ thuộc. Trong bối cảnh dự đoán giá vé chuyến bay, nó có thể được áp dụng để tìm ra mối liên kết tuyến tính giữa các yếu tố như thời gian đặt vé trước, thời gian chuyến bay, và loại ghế với giá vé. Linear Regression được sử dụng khi giả định về mối quan hệ tuyến tính giữa các biến là hợp lý. Trong trường hợp dự đoán giá vé, nó có thể cung cấp một mô hình dễ hiểu và dễ giải thích, đặc biệt là khi có các yếu tố có ảnh hưởng tuyến tính lên giá vé.

Decision Tree Regressor: Decision Tree Regressor là một mô hình dựa trên cây quyết định, chia dữ liệu thành các nhóm con dựa trên các quy tắc quyết định. Mỗi lá của cây đại diện cho một dự đoán giá vé dựa trên các điều kiện tương ứng. Decision Tree Regressor thích hợp khi có sự phức tạp và không tuyến tính trong dữ liệu. Trong ngữ cảnh dự đoán giá vé, nó có thể tự động học được sự ảnh hưởng của các đặc trưng và kiểm tra tính tuyến tính trong mối quan hệ giữa các biến.

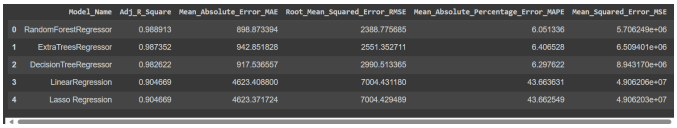
Random Forest Regressor: Random Forest Regressor là một tập hợp của nhiều cây quyết định (Random Forest). Nó tạo ra dự đoán bằng cách kết hợp đầu ra của các cây thành viên. Random Forest giảm nguy cơ overfitting và cung cấp tính ổn định hơn so với một cây quyết định đơn lẻ. Trong dự đoán giá vé, việc sử dụng Random Forest có thể giúp xử lý sự phức tạp và không chắc chắn trong dữ liệu.

Extra Trees Regressor: Extra Trees Regressor tương tự như Random Forest, nhưng nó sử dụng một số chiến lược khác nhau để xây dựng các cây quyết định. Extra Trees Regressor có thể đưa ra dự đoán hiệu quả với số lượng cây ít hơn so với Random Forest, điều này có lợi khi tài nguyên tính toán hạn chế.

Lasso Regression: Lasso Regression là một phương pháp hồi quy tuyến tính sử dụng hàm mất mát bổ sung regularization để ước lượng mô hình. Lasso Regression hữu ích khi có nhiều biến độc lập và chỉ một số ít trong số đó có ảnh hưởng đáng kể đến giá vé. Nó giúp giảm quá mức ảnh hưởng của các biến không quan trọng, làm cho mô hình trở nên đơn giản và dễ giải thích.

IV. KẾT QUẢ

Xem xét kết quả khi sử dụng cả 2 phương pháp chuẩn hóa là StandardScaler và MinMaxScaler, ta thấy ở 2 lần chạy có kết quả gần giống nhau. Cả phương phương pháp đều cho ra kết quả là mô hình Random Forest Regressor có kết quả cao nhất và 2 mô hình hồi quy thông thường là Linear và Lasso cho kết quả gần như tương tự nhau và thấp nhất. Điều này có thể được giải thích như sau: Thuật toán Random Forest có khả năng làm việc tốt mà không cần đến sự chuẩn hóa đặc biệt nếu các biến độc lập có giá trị có thang đo khác nhau. Random Forest chủ yếu dựa vào việc xây dựng nhiều cây quyết định và kết hợp chúng, nó ít bị ảnh hưởng bởi sự thay đổi độ lớn của các biến độc lập. Còn về Linear và Lasso Regression, cả hai mô hình này có thể bị ảnh hưởng lớn hơn bởi sự chuẩn hóa, đặc biệt là nếu có sự khác biệt đáng kể về độ lớn giữa các biến độc lập. Tuy nhiên, nếu phương pháp chuẩn hóa giữ cho các biến độc lập có thang đo tương đồng, Linear và Lasso Regression có thể cho kết quả gần nhau. Hơn hết thì cả 2 đều là mô hình học máy yếu.



	Model_Name	Adj_R_Square	Mean_Absolute_Error_MAE	Root_Mean_Squared_Error_RMSE	Mean_Absolute_Percentage_Error_MAPE	Mean_Squared_Error_MSE
0	RandomForestRegressor	0.988913	898.873394	2388.775685	6.061336	5.706249e+06
1	ExtraTreesRegressor	0.987352	942.851828	2551.352711	6.406528	6.509401e+06
2	DecisionTreeRegressor	0.982622	917.536557	2990.513365	6.297622	8.943170e+06
3	LinearRegression	0.904669	4623.408800	7004.431180	43.663631	4.906206e+07
4	Lasso Regression	0.904669	4623.371724	7004.429489	43.662549	4.906203e+07

Fig. 7: Kết quả của các mô hình khi sử dụng StandardScaler

Dựa vào các kết quả trên, ta thấy rằng khi sử dụng MinMaxScaler trên tập dữ liệu này sẽ cho ra kết quả dự đoán tốt hơn so với StandardScaler. Thêm vào đó, với các mô

```
result.describe()
```

	duration	days_left	price	Price_actual	Price_pred	price_diff
count	90046.000000	90046.000000	90046.000000	90046.000000	90046.000000	90046.000000
mean	12.242090	25.946649	20874.981410	20874.981410	20922.085382	896.575002
std	7.200407	13.543974	22686.388372	22686.388372	22599.438853	2203.777574
min	0.830000	1.000000	1105.000000	1105.000000	1105.000000	0.000000
25%	6.830000	14.000000	4784.000000	4784.000000	4892.160000	2.590000
50%	11.250000	26.000000	7425.000000	7425.000000	7489.345000	123.845000
75%	16.170000	38.000000	42521.000000	42521.000000	42572.250000	653.730000
max	47.750000	49.000000	115211.000000	115211.000000	105922.320000	44277.110000

Fig. 8: Mô tả kết quả khi sử dụng StandardScaler

	Model_Name	Adj_R_Square	Mean_Absolute_Error_MAE	Root_Mean_Squared_Error_RMSE	Mean_Absolute_Percentage_Error_MAPE	Mean_Squared_Error_MSE
0	RandomForestRegressor	0.989013	893.798116	2377.976912	6.008956	5.654774e+06
1	ExtraTreesRegressor	0.987263	944.963227	2560.289553	6.424583	6.555083e+06
2	DecisionTreeRegressor	0.982829	907.396945	2972.643197	6.163993	8.836606e+06
3	LinearRegression	0.904669	4623.408800	7004.431180	43.663631	4.906206e+07
4	Lasso Regression	0.904669	4623.301220	7004.425521	43.660725	4.906198e+07

Fig. 9: Kết quả của các mô hình khi sử dụng MinMaxScaler

hình học máy mà nhóm đã chọn thì mô hình Random Forest Regressor cho kết quả cao nhất. Nhóm cũng đã sử dụng hàm feature_importances_ của mô hình Random Forest Regressor để chọn ra thuộc tính nào có đóng góp nhiều nhất trong việc dự đoán. Kết quả cho thấy là thuộc tính Loại ghế có tác động lớn nhất, tiếp theo là những thuộc tính như Thời lượng chuyến bay, Mã chuyến bay, Thời gian đặt vé (Số ngày còn lại) cũng có ảnh hưởng đáng kể. Còn những thuộc tính như Tên hãng bay, Số lần dừng lại, Thời gian khởi hành hầu như không có ảnh hưởng đến việc dự đoán. Thực tế thì khi nhóm thử nghiệm việc bỏ những thuộc tính đó ra khỏi tập dữ liệu thì kết quả dự đoán có sự thay đổi không đáng kể.

V. Kết luận

Kết luận: Đồ án này không chỉ giúp hiểu rõ hơn về các đặc trưng quyết định giá vé chuyến bay mà còn thể hiện quá trình triển khai mô hình học máy để dự đoán trong tình huống tương tự.

Học được: Trong quá trình thực hiện dự án, nhóm đã có cơ hội áp dụng kiến thức lý thuyết, học được cách xử lý dữ liệu trước khi đưa vào mô hình, chọn mô hình phù hợp và điều chỉnh để cải thiện kết quả dự đoán.

```
result.describe()
```

	duration	days_left	price	Price_actual	Price_pred	price_diff
count	90046.000000	90046.000000	90046.000000	90046.000000	90046.000000	90046.000000
mean	12.242090	25.946649	20874.981410	20874.981410	20918.274128	893.676312
std	7.200407	13.543974	22686.388372	22686.388372	22599.764171	2201.934709
min	0.830000	1.000000	1105.000000	1105.000000	1105.000000	0.000000
25%	6.830000	14.000000	4784.000000	4784.000000	4896.000000	2.300000
50%	11.250000	26.000000	7425.000000	7425.000000	7488.010000	120.485000
75%	16.170000	38.000000	42521.000000	42521.000000	42600.520000	651.250000
max	47.750000	49.000000	115211.000000	115211.000000	108914.680000	41672.150000

Fig. 10: Mô tả kết quả khi sử dụng MinMaxScaler

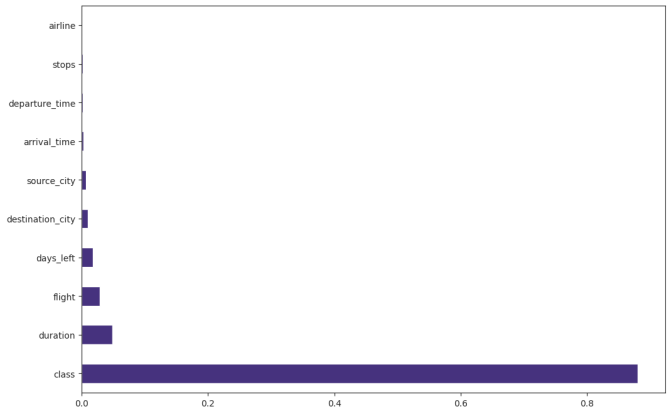


Fig. 11: Minh họa mức ảnh hưởng của các thuộc tính

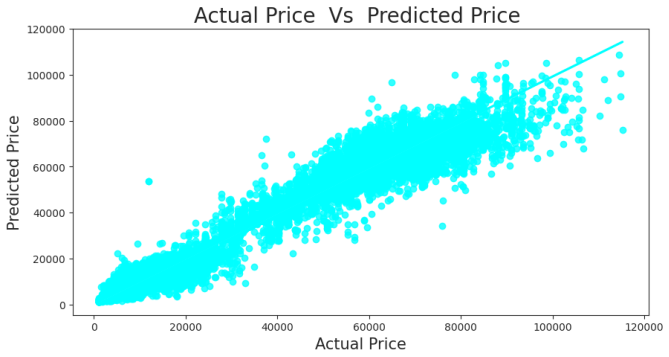


Fig. 12: Biểu đồ phân phối giữa các điểm dữ liệu thực tế và dự đoán

Hướng Phát Triển: Đồ án mở ra nhiều khả năng phát triển, từ việc phân tích các đặc trưng đến thử nghiệm với các mô hình phức tạp hơn để nâng cao độ chính xác. Nên thu thập thêm các đặc trưng khác có thể ảnh hưởng đến giá chuyến bay như là Thời tiết, Mùa, Sự kiện, Ngày lễ...

References