ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

ỨNG DỤNG MÔ HÌNH XGBREGRESSOR KẾT HỢP EDA TRONG PHÂN TÍCH VÀ DỰ ĐOÁN GIÁ VÉ MÁY BAY

Nhóm 13			
Sinh viên thực hiện:			
STT	Họ tên	MSSV	Ngành
1	Nguyễn Hữu Lâm	20521516	CNTT
2	Đặng Quang Trung	20522067	CNTT
3	Lê Thành Phát	21522442	CNTT
4	Nguyễn Phú Kiệt	21522257	TMDT

TP. HỒ CHÍ MINH – 12/2024

1. GIỚI THIỆU

Trong những năm gần đây, sự phát triển mạnh mẽ của ngành du lịch, cùng với sự phục hồi của nền kinh tế, đã thúc đẩy nhu cầu di chuyển của người dân ngày càng cao. Trong đó, máy bay – với đặc điểm an toàn và tốc độ vượt trội – đã trở thành lựa chọn hàng đầu cho những chuyến đi xa, nhờ khả năng rút ngắn thời gian di chuyển và vận chuyển số lượng lớn hành khách. Hiện nay, máy bay đang dần khẳng định vị thế là phương tiện di chuyển hiệu quả và phổ biến nhất. Với đề tài này, nhóm mong muốn đưa ra các dự đoán chính xác nhất về giá vé máy bay, nhằm tối ưu hóa chi phí di chuyển và nắm bắt được xu hướng sử dụng phương tiện này trong thời gian tới.

Trong đề tài này, thông qua bộ dữ liệu tham khảo đã tìm được, cùng với những thông số mấu chốt, nhóm sẽ phân tích và đưa ra dự đoán về giá vé máy bay trong tương lai gần và có thể sẽ phát triển để dự đoán được trong tương lai càng xa càng tốt mà vẫn đảm bảo tỉ lệ chính xác.

2. MÔ TẢ BỘ DỮ LIỆU

Bộ dữ liệu Flight Price Prediction cung cấp thông tin chi tiết về giá vé máy bay ở Ấn Độ, tập trung vào các yếu tố ảnh hưởng đến giá vé từ nhiều hãng hàng không. Bao gồm các thông tin quan trọng như tên hãng hàng không, ngày giờ khởi hành và đến nơi, thời gian bay, số điểm dừng, và giá vé. Đây là một bộ dữ liệu lý tưởng cho các bài toán dự đoán và phân tích giá vé máy bay dựa trên các yếu tố về thời gian và khoảng cách chuyến bay. Bộ dữ liệu được tham khảo tại Kaggle [1].

Mô tả bộ dữ liệu:

Bảng 1. Bảng mô tả bộ dữ liệu

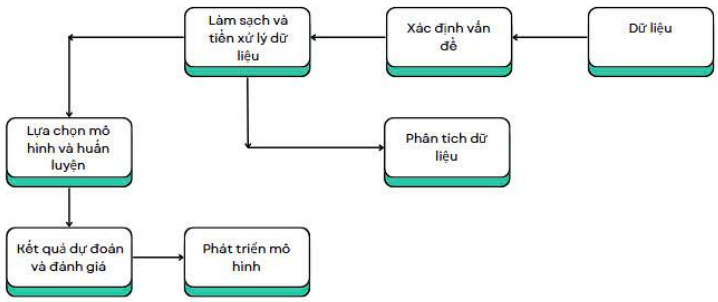
STT	Tên thuộc tính	Mô tả	Kiểu dữ liệu	Miền giá trị
1	Airline	Tên hãng hàng không khai thác chuyến bay	Object	‘IndiGo’ ‘AirIndia’ ‘Jet Airways’ ‘SpiceJet’ …
2	Source	Thành phố khởi hành hoặc sân bay	Object	‘Delhi’ ‘Konkata’ …
3	Destination	Thành phố hoặc sân bay đến	Object	‘Cochin’ ‘Banglore’ …
4	Total_Stops	Số điểm dừng của chuyến bay giữa điểm khởi hành và điểm đến	Float64	0 ➔ 4
5	Price	Chi phí của vé máy bay	Float64	1759 ➔ 79.5k
6	Date	Ngày chuyến bay khởi hành	Float64	1 ➔ 27
7	Month	Tháng của chuyến bay khởi hành	Float64	3 ➔ 6

8	Year	Năm của chuyến bay khởi hành	Float64	‘2019’
9	Dep_hours	Giờ trong ngày chuyến bay khởi hành	Float64	0 ➔ 23
10	Dep_min	Phút của chuyến bay khởi hành	Float64	0 ➔ 55
11	Arrival_hours	Giờ trong ngày chuyến bay đến	Float64	0 ➔ 23
12	Arrival_min	Phút của chuyến bay đến	Float64	0 ➔ 55
13	Duration_hours	Thời lượng chuyến bay tính bằng giờ	Float64	1 ➔ 47
14	Duration_min	Thời lượng chuyến bay tính bằng phút	Float64	0 ➔ 55

Thông tin ban đầu về bộ dữ liệu:

- Số hàng: 10683.
- Số cột: 14.
- Loại biến:
  + Biến phân loại: 3 biến (Airline, Source, Destination).
  + Biến số: 11 biến (Total_Stops, Price, Date, Month, Year, Dep_hours, Dep_min, Arrival_hours, Arrival_min, Duration_hours, Duration_min).
- Dữ liệu đã được làm sạch và không có giá trị khuyết.

3. PHƯƠNG PHÁP PHÂN TÍCH



3.1. Tiền xử lý dữ liệu

Xử lý các dữ liệu trùng nhau nhằm đảm bảo rằng các mô hình không bị sai lệch do thiếu thông tin.

Các cột thời gian được kết hợp lại thành một cột duy nhất “Date_Month_Year. Các cột gốc “Date”, “Month”, “Year” sau đó được loại bỏ để tránh dư thừa dữ liệu.

Xóa cột "Duration_hours" và "Duration_min" để gộp lại thành "Duration".

Tiếp theo là gộp cột "Dep_hours" và "Dep_min" thành cột "Dep_time" (thời gian khởi hành tính bằng phút).

Tương tự, gộp cột "Arrival_hours" và "Arrival_min" thành cột "Arrival_time" (thời gian đến cũng tính bằng phút).

Đưa cột "Price" về cuối bảng để dễ quan sát biến mục tiêu.

	Airline	Source	Destination	Total_Stops	Date_Month_Year	Duration	Dep_time	Arrival_time	Price
0	IndiGo	Banglore	New Delhi	0	2019-03-24	170	1340	70	3897
1	Air India	Kolkata	Banglore	2	2019-05-01	445	350	795	7662
2	Jet Airways	Delhi	Cochin	2	2019-06-09	1140	565	265	13882
3	IndiGo	Kolkata	Banglore	1	2019-05-12	325	1085	1410	6218
4	IndiGo	Banglore	New Delhi	1	2019-03-01	285	1010	1295	13302

3.2. Phân tích thăm dò

Trong bài báo cáo này, nhóm chúng em tập trung khai thác và trực quan hóa dữ liệu dưới hai góc độ: đơn biến và đa biến, nhằm nhận diện các đặc điểm nổi bật và mối quan hệ giữa các yếu tố trong tập dữ liệu.

3.3. Xây dựng mô hình

Nhóm sử dụng mô hình dự đoán **XGBoost Regressor** (XGBRegressor) – một mô hình dự đoán thuộc thuật toán XGBoost (Extreme Gradient Boosting), được thiết kế để giải quyết các bài toán hồi quy (regression). Đây là một phương pháp học máy mạnh mẽ dựa trên kỹ thuật gradient boosting, trong đó nhiều mô hình cây quyết định (decision trees) được xây dựng tuần tự để cải thiện dần độ chính xác của dự đoán.

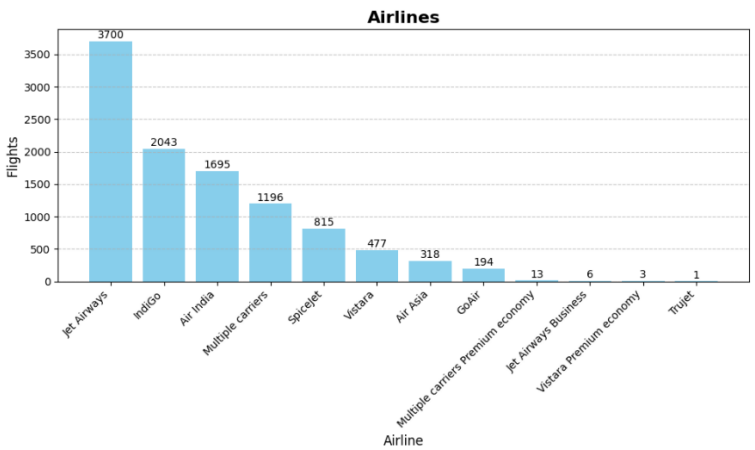
3.4. Đánh giá mô hình

Sử dụng các thông số như **Mean Absolute Error** (**MAE**) và **R^2 Score** để thực hiện đánh giá mô hình.

4. PHÂN TÍCH THĂM DÒ/SƠ BỘ

4.1. Phân tích và trực quan dữ liệu đơn biến

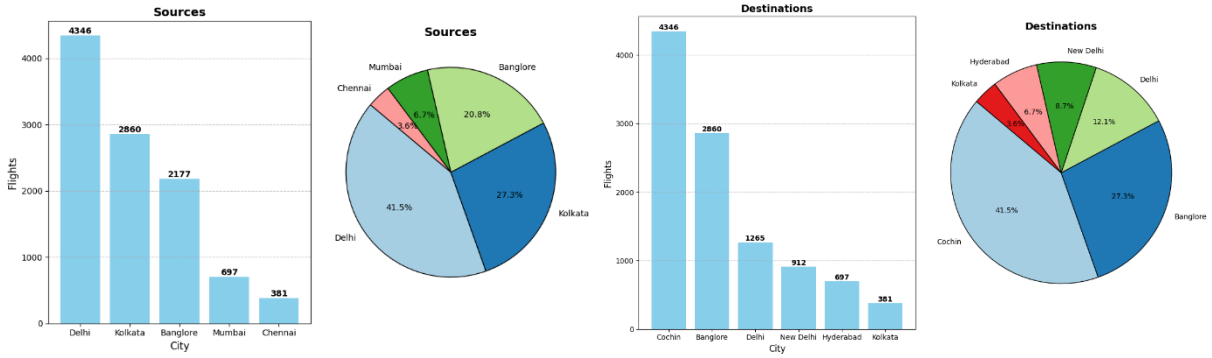
4.1.1. Thuộc tính "Airline" (Hãng hàng không)



Có thể thấy **Jet Airways** là hãng hàng không được sử dụng nhiều nhất với 3.700 chuyến bay, tiếp theo là **IndiGo** với 2.043 chuyến. Ngược lại, **Trujet** chỉ có 1 chuyến bay trong tập dữ liệu.

Điều này cho thấy **Jet Airways** là lựa chọn phổ biến nhất, trong khi **Trujet** ít được hành khách lựa chọn, có thể do phạm vi hoạt động hạn chế hoặc các yếu tố khác.

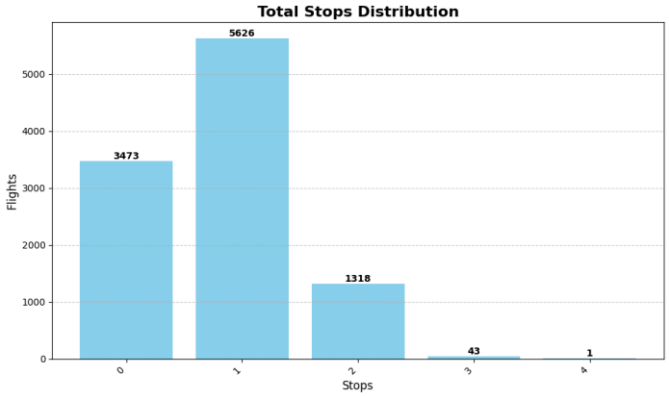
4.1.2. Thuộc tính "Source" và "Destination" (Thành phố xuất phát và điểm đến)



Delhi là thành phố xuất phát phổ biến nhất, chiếm 41,5% tổng số chuyến bay, cho thấy đây là trung tâm hàng không lớn và sôi động.

Cochin là điểm đến phổ biến nhất với tỷ lệ tương tự 41,5%, chứng minh sức hút du lịch và tiềm năng phát triển của khu vực này.

4.1.3. Thuộc tính "Total_Stops" (Số điểm dừng)



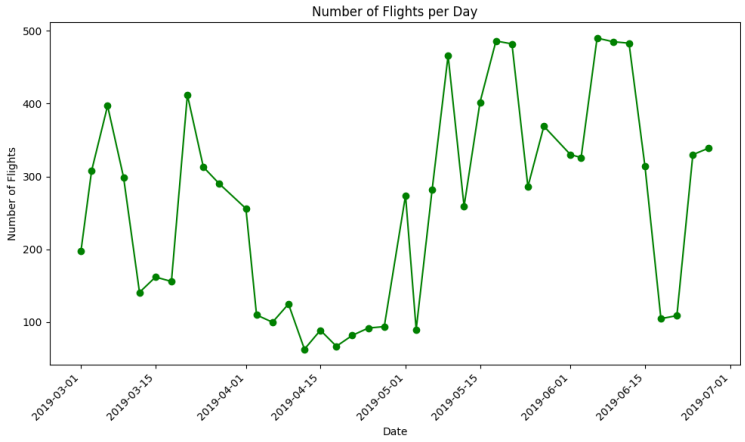
Các chuyến bay có 1 điểm dừng (1 Stop) chiếm tỷ lệ lớn nhất với 5.626 chuyến, thể hiện đây là loại hình di chuyển phổ biến nhất.

Các chuyến bay không điểm dừng (0 Stop) cũng có số lượng đáng kể với khoảng 3.000 chuyến, cho thấy nhu cầu cao đối với các hành trình nhanh và trực tiếp.

Các chuyến bay với 2 điểm dừng có số lượng ít hơn so với 1 điểm dừng và không điểm dừng, nhưng vẫn ở mức trung bình. Các chuyến bay với 3 điểm dừng trở lên rất hiếm, cho thấy nhu cầu thấp đối với những chuyến bay kéo dài với nhiều điểm dừng.

Có rất ít chuyến bay có 4 điểm dừng, chỉ xuất hiện một lần trong tập dữ liệu. Đây có thể là trường hợp đặc biệt hoặc liên quan đến các tuyến bay ít phổ biến.

4.1.4. *Thuộc tính "Date_Month_Year" (Thời gian)*

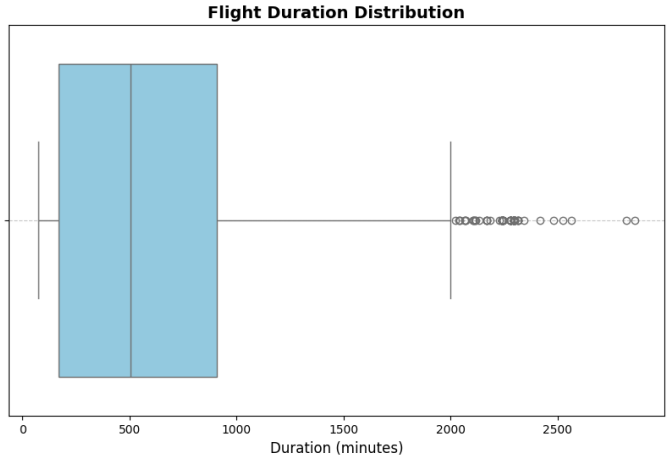


Số lượng chuyến bay dao động đáng kể trong khoảng thời gian từ tháng 3 đến cuối tháng 6 năm 2019. Xu hướng dường như tăng dần từ tháng 3 đến tháng 5 năm 2019, đạt đỉnh vào giữa tháng 5-6 với hơn 400 chuyến bay mỗi ngày, có khả năng do mùa cao điểm du lịch.

Thời điểm giữa tháng 4 ghi nhận số chuyến bay giảm mạnh xuống dưới 100 chuyến mỗi ngày, có thể do các yếu tố như thời tiết hoặc lịch nghỉ lễ.

Xu hướng biến động giữa các tuần và tháng cho thấy nhu cầu thị trường phụ thuộc nhiều vào mùa vụ và các sự kiện bên ngoài.

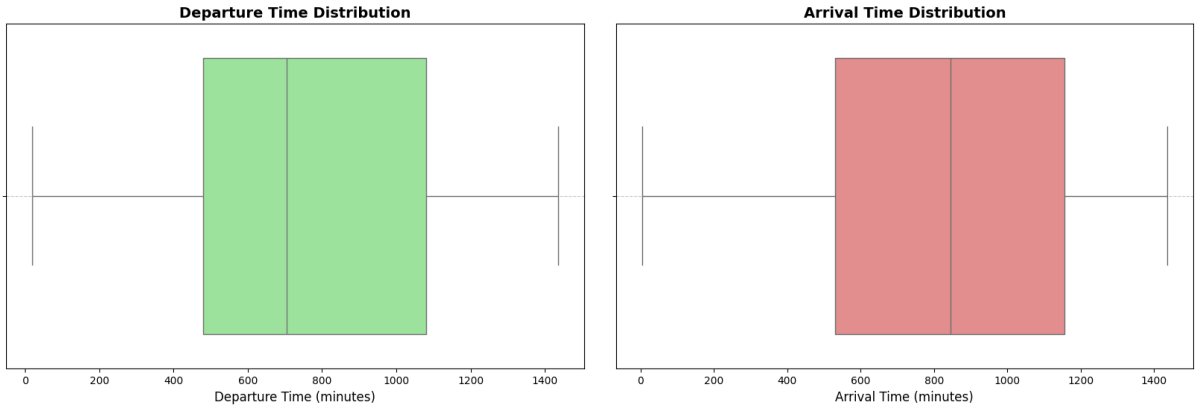
4.1.5. *Thuộc tính "Duration" (Thời gian bay)*



Phần lớn chuyến bay có thời gian bay từ 200-900 phút, phù hợp với các tuyến bay tầm trung.

Một số giá trị ngoại lệ xuất hiện, đại diện cho các chuyến bay đường dài hoặc hành trình phức tạp, cho thấy sự đa dạng trong loại hình chuyến bay.

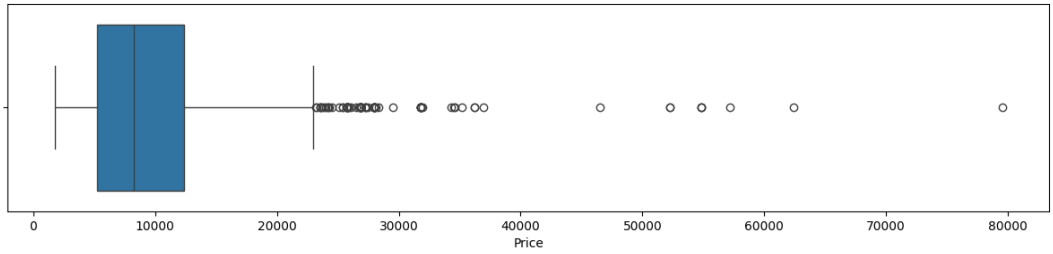
4.1.6. Thuộc tính "Dep_time" và "Arrival_time" (Thời gian khởi hành và hạ cánh)



Hầu hết các chuyến bay khởi hành vào buổi sáng, với tần suất đều đặn suốt cả ngày.

Các chuyến bay thường hạ cánh vào ban đêm, phản ánh lịch trình bay tối ưu hóa cho thời gian di chuyển của hành khách và hiệu quả hoạt động sân bay.

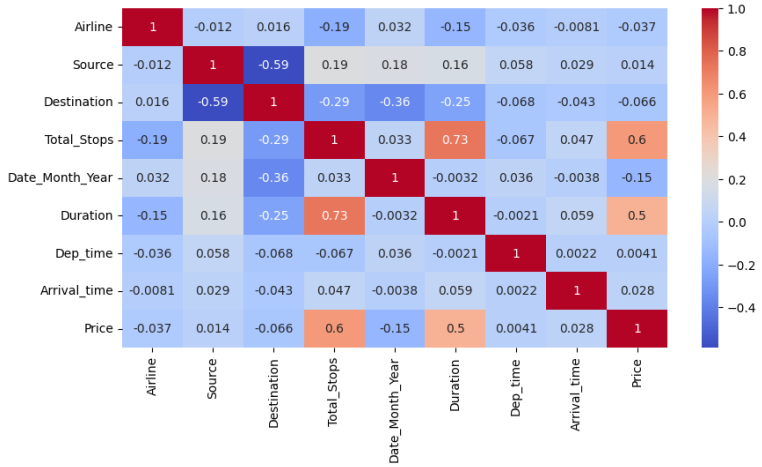
4.1.7. Thuộc tính "Price" (Giá vé)



Giá vé chủ yếu dao động trong khoảng từ 6.000 đến 13.000, cho thấy mức giá ổn định và hợp lý đối với đa số chuyến bay.

Tuy nhiên, có một số chuyến bay có giá vé cao bất thường lên đến gần 80.000, có thể liên quan đến dịch vụ cao cấp, tuyến bay đặc biệt, hoặc thời điểm đặt vé gấp.

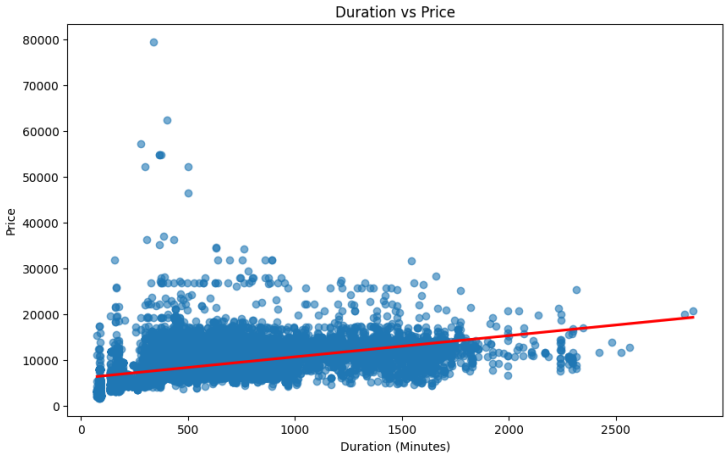
4.2. Phân tích và trực quan dữ liệu 2 biến



Dựa vào Heatmap, ta thấy được mối tương quan giữa "Price", "ToTal_Stops" và "Duration" cao hơn các features khác:

- Total_Stops (0.6): Đây là thuộc tính có mối tương quan cao nhất với giá vé. Giá vé càng cao khi số lượng điểm dừng tăng lên
- Duration (0.5): Giá vé có mối tương quan vừa phải với thời lượng chuyến bay. Chuyến bay dài hơn thường có giá cao hơn.

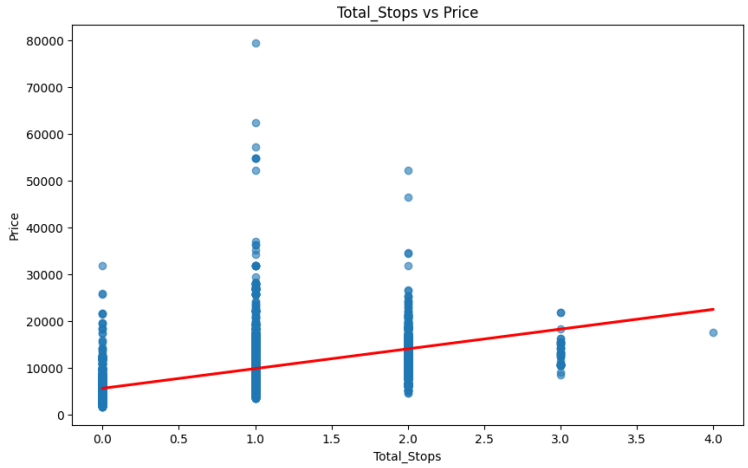
Ngoài ra ta thấy sự tương quan cao giữa "Duration" và "Total_Stops".



Biểu đồ thể hiện mối quan hệ giữa thời gian bay (Duration) và giá vé (Price).

Đường hồi quy đỏ cho thấy thời gian bay càng dài, giá vé có xu hướng cao hơn.

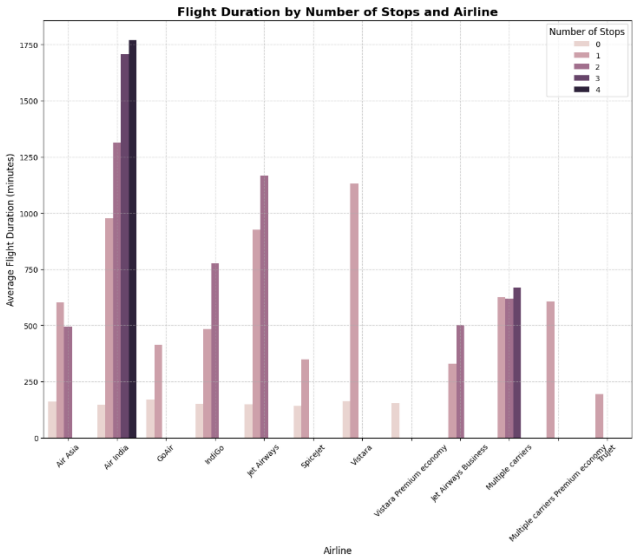
Tuy nhiên, cũng có sự phân tán khá lớn, đặc biệt ở mức giá thấp, cho thấy rằng không phải lúc nào thời gian bay dài hơn cũng đồng nghĩa với giá vé cao hơn.



Biểu đồ này thể hiện mối quan hệ giữa số điểm dừng (Total_Stops) và giá vé (Price).

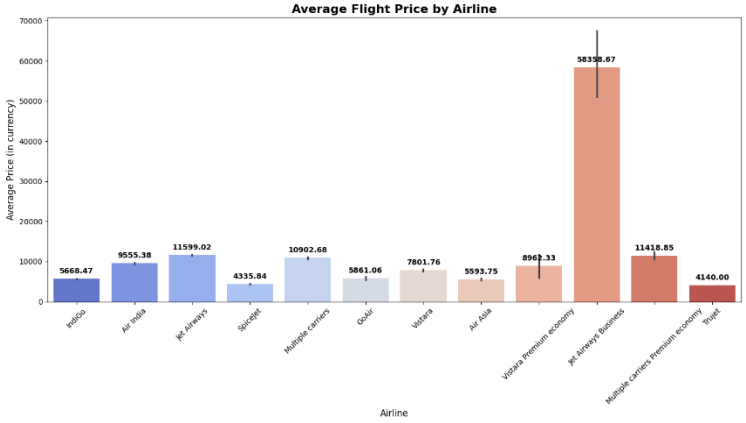
Đường hồi quy đỏ cho thấy xu hướng tổng quát: càng nhiều điểm dừng thì giá vé có xu hướng càng tăng.

Tuy nhiên, dữ liệu khá phân tán, đặc biệt ở mức giá cao, chứng tỏ rằng số điểm dừng không phải yếu tố duy nhất ảnh hưởng đến giá vé.

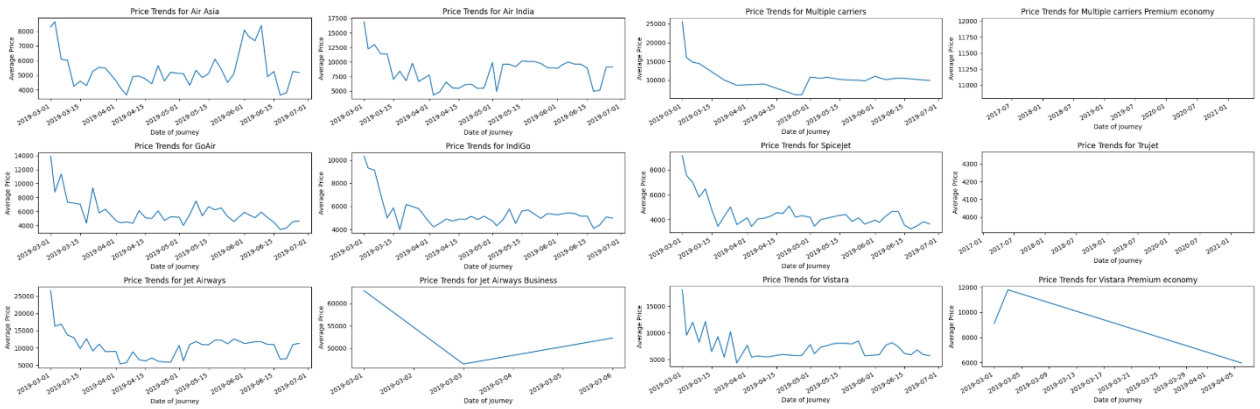


Sự ảnh hưởng khá giống nhau ở các hãng hàng không là số điểm dừng tăng thì thời gian bay cũng tăng và ngược lại.

Có 1 ngoại lệ đó là Multiple carriers thì không có chênh lệch nhiều.

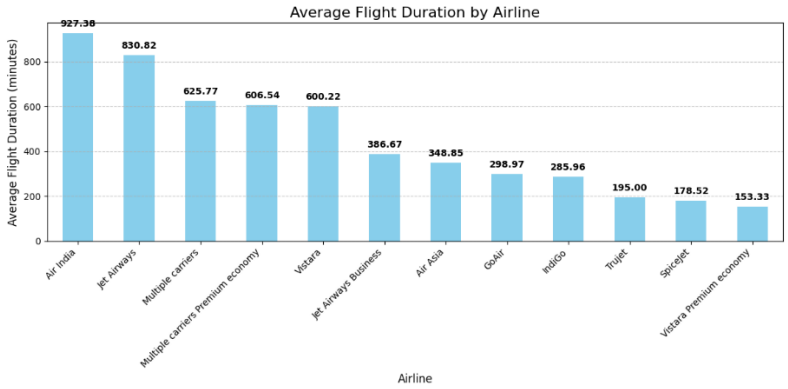


Giá của hãng **Spicejet** và **Trujet** khá tương đồng và có giá thấp nhất. Trong khi đó, giá của hãng **Jet Airways Business** có giá cao nhất.



Dựa vào các biểu đồ trên ta có thể quan sát thấy:

- Về xu hướng chung: Giá vé có xu hướng theo mùa rõ ràng, giá thường tăng vào nửa đầu năm và giảm dần về cuối năm. Xu hướng này nhất quán ở tất cả các hãng hàng không.
- Vào mùa cao điểm (từ tháng 1 đến tháng 3): **Jet Airways Business** có giá vé cao nhất trong thời gian này. Tiếp đến là **Jet Airways** và **Multiple carriers**.
- Vào ngoài mùa (các tháng còn lại): Giá vé trung bình thấp cho hầu hết các hãng hàng không trong năm.



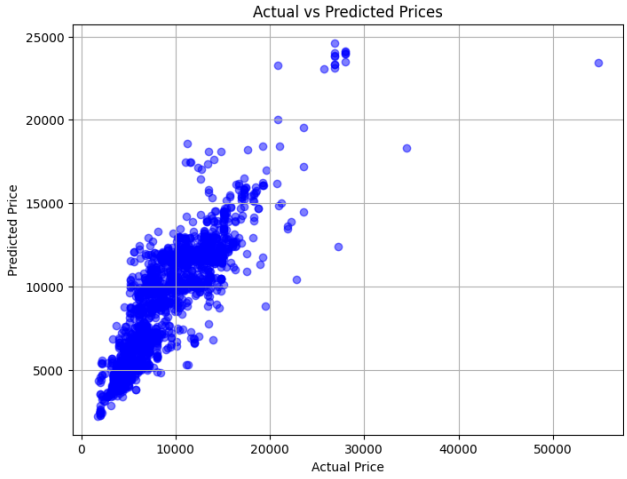
Qua biểu đồ ta có thể thấy **Vistara premium economy** có thời gian bay ngắn nhất và **Air India** có thời gian bay dài nhất.

5. KẾT QUẢ PHÂN TÍCH

Kết quả tính toán của mô hình XGBRegressor:

- Mean Absolute Error (MAE): 1533.396857059805
- R^2 Score: 0.7657318115234375

Biểu đồ phân tán giữa giá thực tế và giá dự đoán:



- Đây là biểu đồ so sánh giá thực tế và giá dự đoán của mô hình:

+ Phân bố các điểm khá tập trung dọc theo đường chéo (y = x), cho thấy mô hình có xu hướng dự đoán khá chính xác trong phần lớn trường hợp.
+ Tuy nhiên, có một số điểm lệch xa khỏi đường chéo, cho thấy mô hình gặp khó khăn trong việc dự đoán các trường hợp có giá trị lớn hoặc nhỏ bất thường.

6. CHỈNH SỬA SAU BÁO CÁO

6.1. Thêm phần cam kết

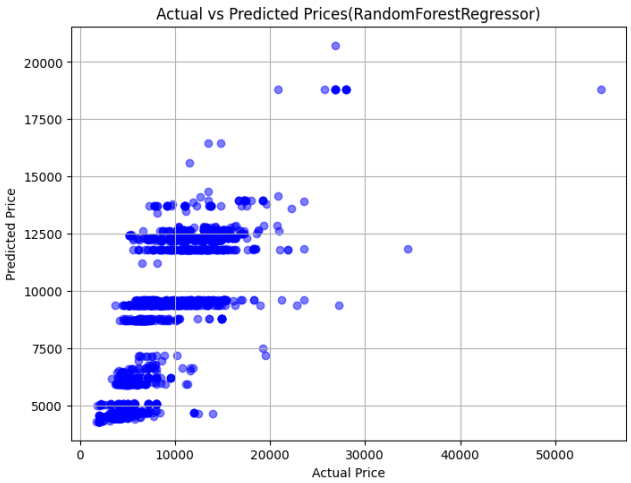
Chúng tôi xin cam kết bộ dữ liệu và đề tài này do nhóm tự phân tích thiết kế và không dựa trên đề tài nào khác. Bộ dữ liệu phân tích tự thu thập và được tham khảo tại Kaggle [1].

6.2. Huấn luyện thêm mô hình Random Forest Regressor

Kết quả tính toán của mô hình Random Forest Regressor:

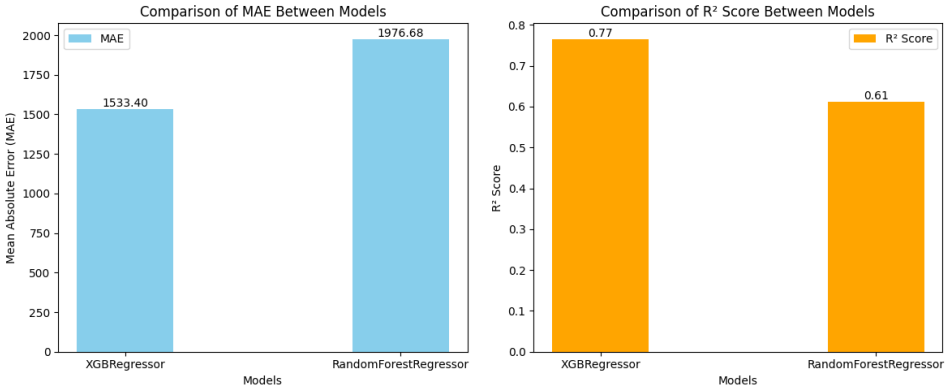
- Mean Absolute Error (MAE): 1976.676476933425
- R^2 Score: 0.6125363217728258

Biểu đồ phân tán giữa giá thực tế và giá dự đoán:



- Đây là biểu đồ so sánh giá thực tế và giá dự đoán của mô hình:
  + Các điểm dữ liệu phân tán hơn và có sự chồng lấn mạnh mẽ giữa các cụm giá trị dự đoán, đặc biệt ở mức giá thấp và trung bình.
  + Điều này cho thấy mô hình bị giới hạn trong việc dự đoán chính xác các giá trị thực cao hơn.

6.3. So sánh 2 biểu đồ XGBRegressor và Random Forest Regressor



Qua biểu đồ ta có thể thấy được:

- Mô hình XGBRegressor có chỉ số MAE là 1533.40, thấp hơn so với Random Forest Regressor (1976.68) ➔ Điều đó cho thấy XGBRegressor dự đoán tốt hơn về mặt sai số trung bình.
- Còn đối với chỉ số R^2 Score thì XGBRegressor đạt 0.77 và cao hơn Random Forest Regressor (0.61) ➔ Chứng minh XGBRegressor giải thích tốt hơn sự biến thiên của giá vé máy bay.

6.4. Huấn luyện mô hình Linear Regression

Mặc dù XGBRegressor và RandomForestRegressor đã mang lại hiệu suất khá tốt, nhưng độ chính xác vẫn chưa đạt kỳ vọng, thể hiện qua giá trị MAE còn tương đối cao (lần lượt là 1533.40 và 1976.68). Điều này phản ánh những hạn chế riêng của từng mô hình khi thực hiện dự đoán. Để khắc phục vấn đề này và khai thác tối đa điểm mạnh của cả hai phương pháp, việc sử dụng mô hình Linear Regression Stacking Ensemble là một giải pháp hiệu quả. Cách tiếp cận này không chỉ cải thiện đáng kể độ chính xác dự đoán mà còn giúp giảm MAE xuống mức thấp nhất và nâng cao R^2 lên đáng kể. Kết quả này minh chứng rằng mô hình ensemble thực sự phát huy được tiềm năng của cả hai mô hình thành phần (XGB và RF), tạo ra một hệ thống dự đoán vượt trội hơn so với từng mô hình riêng lẻ.

7. KẾT LUẬN

Trong quá trình phân tích và trực quan hóa dữ liệu chuyến bay, nhóm đã thực hiện các bước quan trọng, bao gồm xử lý dữ liệu, khám phá dữ liệu đơn biến và đa biến để hiểu rõ từng thuộc tính, cũng như phân tích các yếu tố ảnh hưởng giá vé máy bay, mối tương quan giữa thời gian bay và số điểm dừng. Kết quả tính toán của mô hình XGBRegressor cho thấy rằng mô hình có hiệu quả khá tốt trong việc dự đoán giá vé máy bay và nhóm có thể nghiên cứu để áp dụng các phân tích này với giá vé ở Việt Nam.

Tuy nhiên nhóm vẫn chưa phân tích sâu cũng như các yếu tố ảnh hưởng bên ngoài như thời tiết... Dữ liệu chỉ tập trung vào một khoảng thời gian nhất định, thiếu tính toàn

diện. Và (MAE) đạt 1533.39 cho thấy rằng trung bình mô hình sai lệch khoảng 1533 đơn vị tiền tệ so với giá vé thực tế. Đây là một mức sai lệch cần được xem xét.

Nhóm em sẽ tiếp tục quá trình nghiên cứu và phát triển để dự ăn tăng tính ứng dụng và đưa ra các đề xuất chi tiết hơn.

TÀI LIỆU THAM KHẢO

[1] Kaggle. link**:** https://www.kaggle.com/datasets/viveksharmar/flight-price-data

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Nguyễn Hữu Lâm	Phân tích trực quan dữ liệu
2	Đặng Quang Trung	Viết báo cáo, làm slide
3	Lê Thành Phát	Mô hình dự đoán đánh giá
4	Nguyễn Phú Kiệt	Làm dashboard