**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỆ THỐNG THÔNG TIN**

A blue logo with a black background

Description automatically generated

**FLIGHT PREDICTION**

**DỮ LIỆU LỚN**

Giảng viên hướng dẫn: **ThS. Nguyễn Hồ Duy Trí**

Lớp: **IS405.O22.HTCL**

Sinh viên thực hiện:

|  |  |  |
| --- | --- | --- |
| STT | Họ và tên | MSSV |
| 1 | Lê Thị Đoan Trang | 20522038 |
| 2 | Nguyễn Thị Thảo Trang | 20522040 |
| 3  4 | Lê Ngọc Mai Thanh  Trần Minh Ngọc | 20521913  20521669 |

**Thành phố Hồ Chí Minh, 2024**

**Mục Lục**

[**1.** **Tổng quan** 3](#_Toc168426483)

[**1.1.** **Lý do chọn đề tài** 3](#_Toc168426484)

[**1.2.** **Giới thiệu tập dữ liệu** 3](#_Toc168426485)

[**1.3.** **Mô tả bài toán** 4](#_Toc168426486)

[**1.4.** **Giới thiệu giải thuật** 5](#_Toc168426487)

[**2.** **Tiền xử lý dữ liệu** 6](#_Toc168426488)

[**2.1.** **Thu thập dữ liệu** 6](#_Toc168426489)

[**2.2.** **Mô tả dữ liệu** 16](#_Toc168426490)

[**2.3.** **Trực quan hóa dữ liệu** 19](#_Toc168426491)

[**3.** **Xây dựng thuật toán** 21](#_Toc168426492)

[**4.** **Kết quả đạt được** 24](#_Toc168426493)

[**4.1.** **Kết quả** 24](#_Toc168426494)

[**4.2.** **Đánh giá** 24](#_Toc168426495)

[**5.** **Tổng kết** 26](#_Toc168426496)

[**5.1.** **Ưu điểm** 26](#_Toc168426497)

[**5.2.** **Nhược điểm** 27](#_Toc168426498)

[**5.3.** **Hướng phát triển** 27](#_Toc168426499)

[**Bảng Phân Công** 29](#_Toc168426500)

[**Tài liệu tham khảo** 29](#_Toc168426501)

1. **Tổng quan**
   1. **Lý do chọn đề tài**

Trong ngành du lịch và hàng không hiện nay, việc dự đoán giá vé máy bay đang là một thách thức lớn bởi giá vé thường xuyên biến động, thay đổi do nhiều yếu tố khác nhau.

Đề tài "Dự đoán giá vé máy bay” giúp cung cấp cái nhìn sâu sắc vào cách thức xác định mức giá vé thông qua việc áp dụng các thuật toán Hồi quy Tuyến tính và các kiểm định giả thuyết trên dữ liệu thu thập được. Phân tích này sẽ giúp làm rõ các mối quan hệ và yếu tố ảnh hưởng tới giá vé, từ đó hỗ trợ việc đưa ra dự đoán chính xác hơn.

Bên cạnh việc đóng góp cho lĩnh vực nghiên cứu về dự đoán giá vé, nghiên cứu cũng hứa hẹn sẽ cải thiện trải nghiệm đặt vé trực tuyến cho người dùng, đặc biệt là trong bối cảnh du lịch trực tuyến ngày càng phát triển.

Ngoài ra, hiểu được các yếu tố ảnh hưởng đến giá vé có thể giúp các hãng hàng không và các nhà cung cấp dịch vụ tối ưu hóa chiến lược giá và cải thiện dịch vụ khách hàng.

* 1. **Giới thiệu tập dữ liệu**

Link dataset: <https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction/data>

Bộ dữ liệu chứa thông tin về các lựa chọn đặt vé máy bay từ trang web Easemytrip cho cho các chuyến bay giữa 6 thành phố lớn nhất của Ấn Độ. Dữ liệu được chia thành hai phần: một cho vé hạng phổ thông và một cho vé hạng thương gia. Có tổng cộng 300261 điểm dữ liệu và 11 đặc trưng trong bộ dữ liệu sau khi đã được tiền xử lý. [2]

Mô tả dữ liệu:

|  |  |  |
| --- | --- | --- |
| STT | Tên thuộc tính | Mô tả |
| 1 | Airline | Tên của 6 hãng hàng không khác nhau. |
| 2 | Flight | Mã chuyến bay. |
| 3 | Source City | Thành phố xuất phát của chuyến bay, có 6 thành phố khác nhau. |
| 4 | Departure Time | Thời gian xuất phát được chia thành các khoảng thời gian, có 6 nhãn thời gian khác nhau. |
| 5 | Stops | Số lượng điểm dừng giữa các thành phố xuất phát và thành phố đích đến, có 3 giá trị khác nhau. |
| 6 | Arrival Time | Thời gian đến được chia thành các khoảng thời gian, có 6 nhãn thời gian khác nhau. |
| 7 | Destination City | Thành phố đích đến của chuyến bay, có 6 thành phố khác nhau. |
| 8 | Class | Thông tin về hạng ghế, có hai giá trị khác nhau: Business và Economy. |
| 9 | Duration | Thời gian tổng cộng để di chuyển giữa các thành phố (tính bằng giờ). |
| 10 | Days Left | Đặc điểm được tạo ra bằng cách lấy ngày đi trừ cho ngày đặt vé. |
| 11 | Price | Biến mục tiêu lưu trữ thông tin về giá vé. |

* 1. **Mô tả bài toán**

Sử dụng thuật toán học máy tiên tiến Linear Regression, mô hình dự đoán giá vé của chúng tôi mang đến khả năng dự đoán chính xác giá vé cho mọi hành trình. Hệ thống tiếp thu dữ liệu từ nhiều nguồn khác nhau, bao gồm thông tin chi tiết về chuyến bay, đặc trưng của hãng hàng không, thời gian khởi hành, số lần dừng, hạng ghế và nhiều yếu tố khác ảnh hưởng đến giá vé.

**Lợi ích cho mọi đối tượng:**

* **Hành khách:** Dễ dàng so sánh giá vé, lựa chọn thời điểm đặt vé phù hợp để tiết kiệm chi phí.
* **Đại lý đặt vé:** Tự tin tư vấn cho khách hàng lựa chọn vé máy bay với giá tốt nhất.
* **Hãng hàng không:** Phân tích nhu cầu thị trường, đưa ra chiến lược giá cả hợp lý, tối ưu hóa doanh thu và nâng cao trải nghiệm khách hàng.

=> Mô hình giúp mọi đối tượng trong ngành hàng không đưa ra quyết định sáng suốt, mang lại lợi ích tối ưu cho cả hành khách, đại lý đặt vé và hãng hàng không.

* 1. **Giới thiệu giải thuật**

Linear Regression, hay hồi quy tuyến tính, là một kỹ thuật thống kê giúp dự đoán giá trị của một biến dựa trên các biến khác. Nó là một công cụ để tìm ra mối quan hệ giữa các yếu tố khác nhau và biến kết quả cần được quan tâm. Mục tiêu chính của Linear Regression là xây dựng một mô hình tuyến tính để dự đoán giá trị của biến phụ thuộc dựa trên các biến độc lập.

Mô hình này được sử dụng để tìm một phương trình dự đoán kết quả tốt nhất cho biến y như một hàm tuyến tính của các biến x. Tương tự như với dự đoán giá vé máy bay, Linear Regression sẽ tìm ra mối quan hệ giữa những yếu tố có thể ảnh hưởng đến giá vé và giúp định giá vé máy bay một cách phù hợp.

Lý do chọn giải thuật Linear Regression:

1. Trong việc dự đoán giá vé, có thể có những mối quan hệ tuyến tính giữa các đặc trưng như thời gian, số lượng điểm dừng, hãng hàng không,… với giá vé.
2. Linear Regression là một mô hình đơn giản và dễ hiểu, giúp dễ dàng triển khai và giải thích kết quả dự đoán để áp dụng vào các bài toán thực tế.
3. Vì giá vé là dạng dữ liệu liên tục (real-time) và lượng dữ liệu tương đối lớn, nên Linear Regression sẽ phản ánh tốt các mối quan hệ tuyến tính đối với giá vé máy bay, giúp giảm thời gian và tài nguyên tính toán.

Ta có phương trình hồi quy tuyến tính đa biến:

|  |
| --- |
|  |

Với

|  |  |
| --- | --- |
|  | Biến phụ thuộc (giá trị muốn dự đoán – vé máy bay) |
|  | Biến độc lập (thời gian bay, số điểm dừng, hãng hàng không) |
|  | Tung độ góc (hệ số chặn của giá trị Y, khi các biến độc lập = 0) |
|  | Hệ số góc riêng phần (mức độ ảnh hưởng của từng biến độc lập đối với biến phụ thuộc) |
|  | Sai số ngẫu nhiên |

Công thức với các ma trân

A picture containing diagram

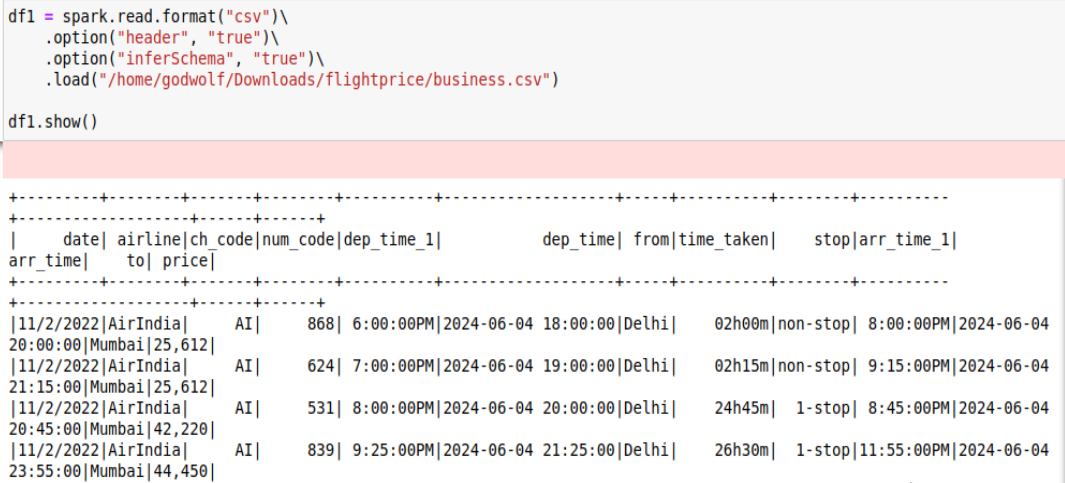
Description automatically generated

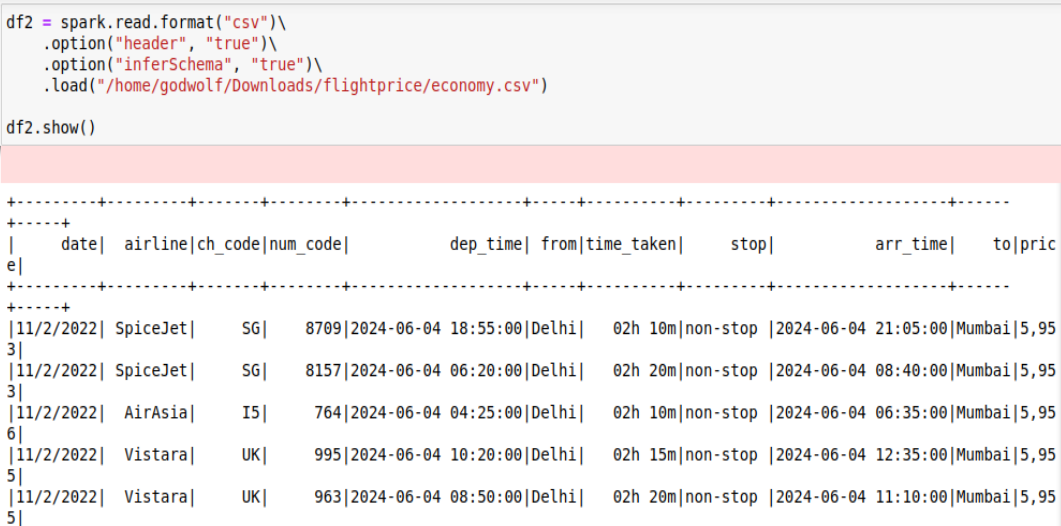
Với:

* B là ma trận cần tìm
* X là ma trận gồm các thuộc tính độc lập trong dữ liệu
* Y là ma trân 1 cột là cột cần dự đoán

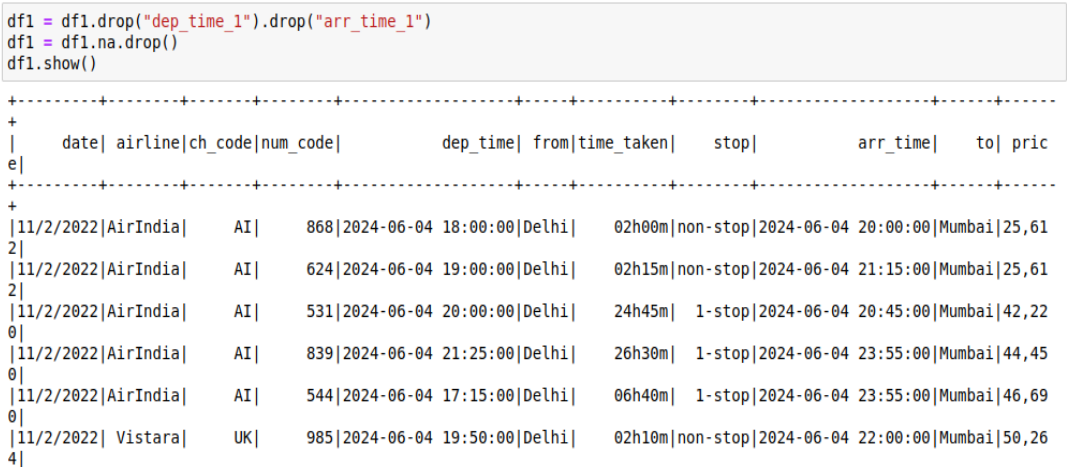
1. **Tiền xử lý dữ liệu**
   1. **Thu thập dữ liệu**

Hai file CSV kết hợp thành một tập dữ liệu ghi nhận việc đặt vé của hai hạng ghế khác nhau: hạng Thương gia (Business) và hạng Phổ thông (Economy)

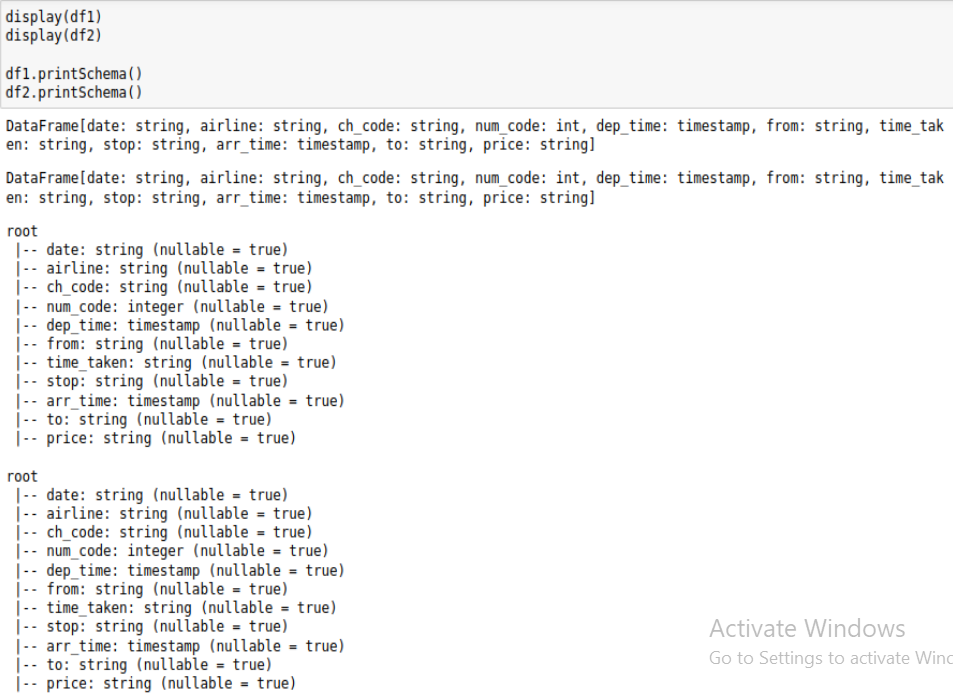


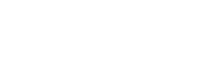
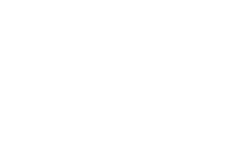


Xử lý xóa bỏ các cột dep\_time\_1 và arr\_time\_1 vì trùng với dep\_time và arr\_time trong tập dữ liệu “business”

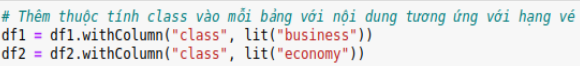


Mô tả chi tiết của hai tập dữ liệu

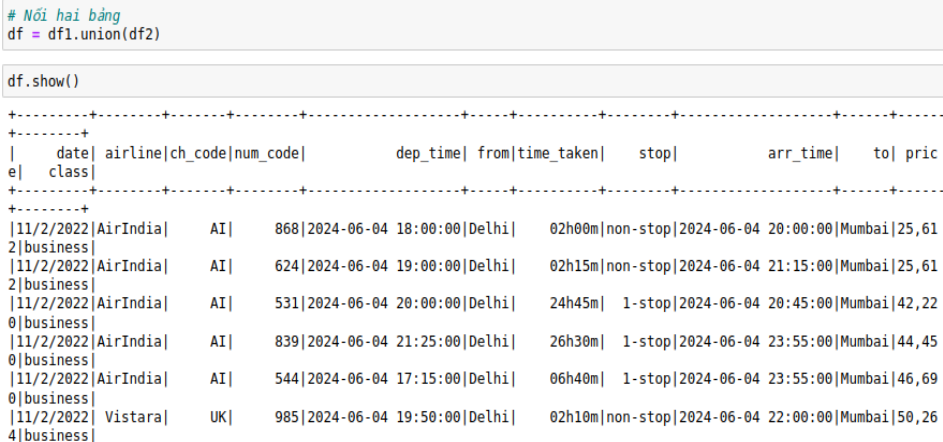




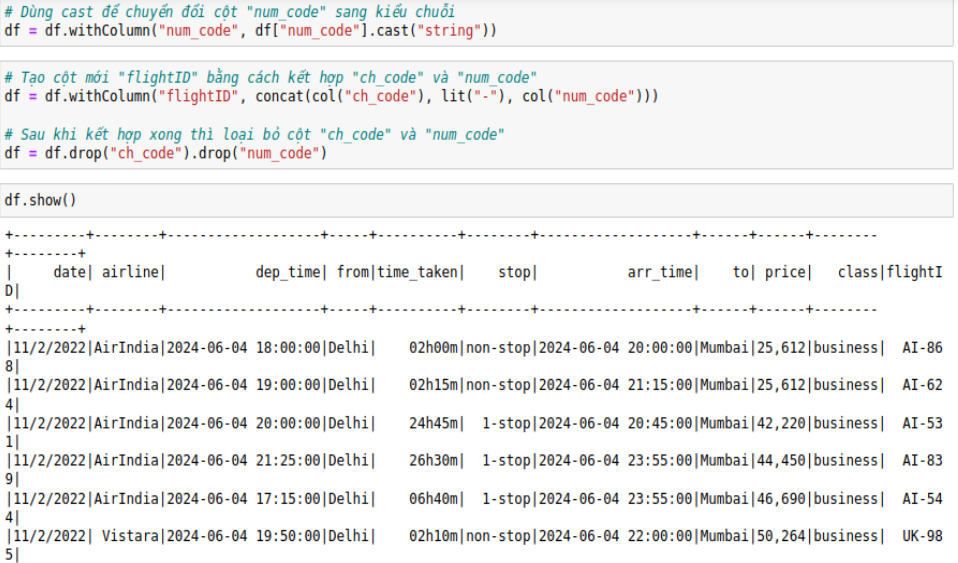
Thêm thuộc tính class vào mỗi bảng với nội dung tương ứng với hạng vé



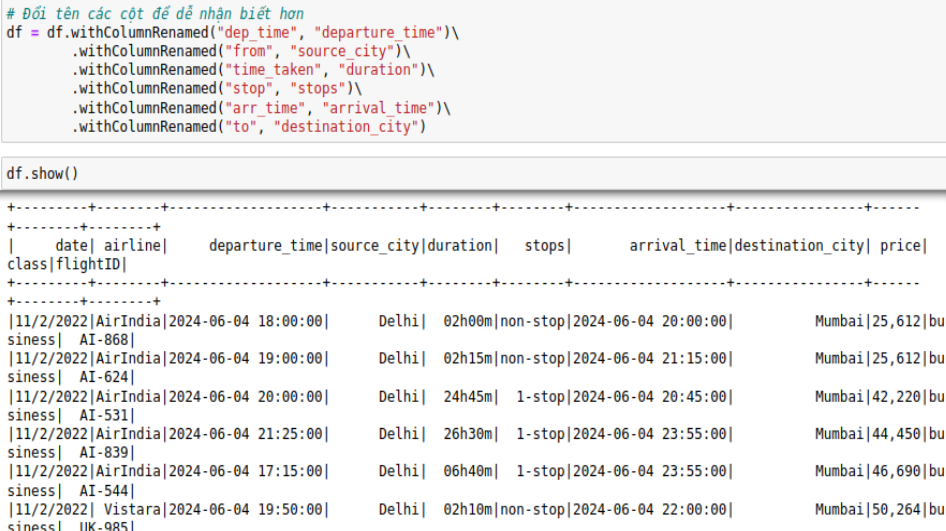
Ghép hai bảng lại thành một bảng



Tạo cột “flightID” bằng cách ghép hai cột “num\_code” và “ch\_code”



Đổi tên cột.

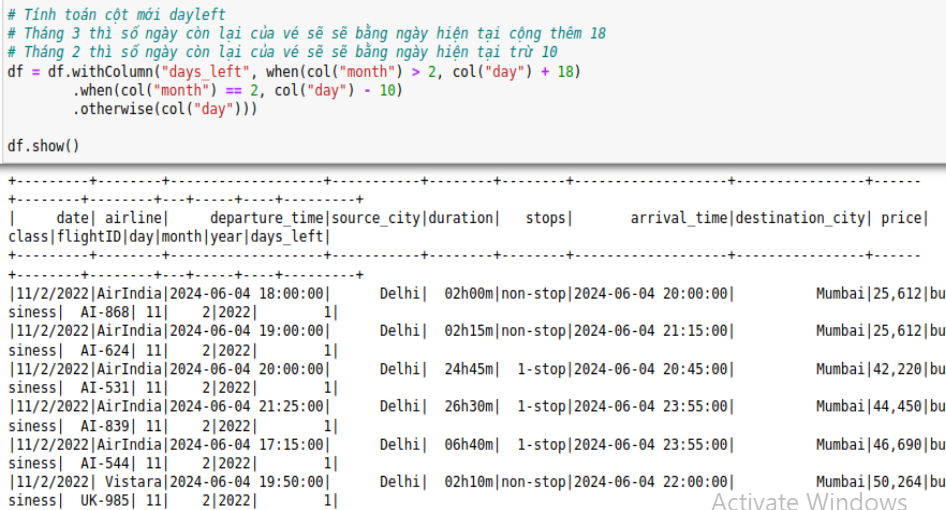


Thêm cột days\_left: hạn sử dụng của vé máy bay.

* B1: tách ngày tháng năm từ cột “date” và lưu vào các cột mới “day” “month” “year”.



* B2: Tính toán hạn sử dụng của vé máy bay.



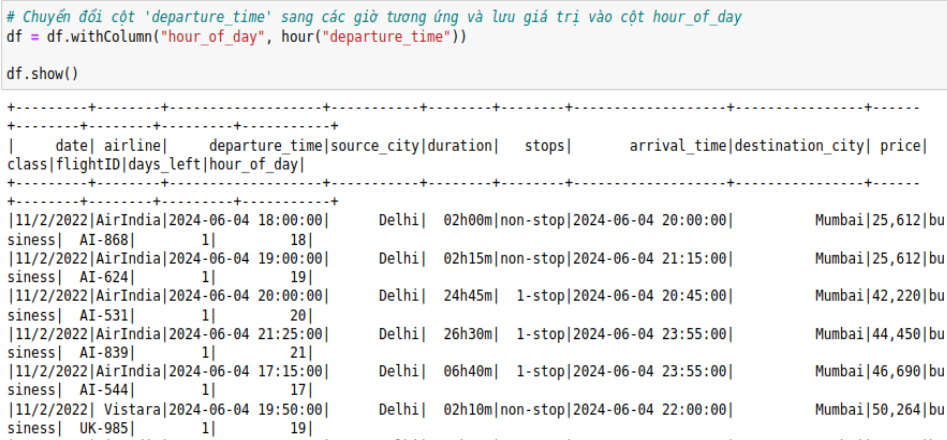
* B3: Sau khi ra được cột “day\_left” thì xóa các cột không cần nữa “day” “month” “year”.



Chuyển đổi thời gian ở các cột “departure\_time” “arrival\_time” thành các buổi trong ngày.

* Chuyển đổi “departure\_time”

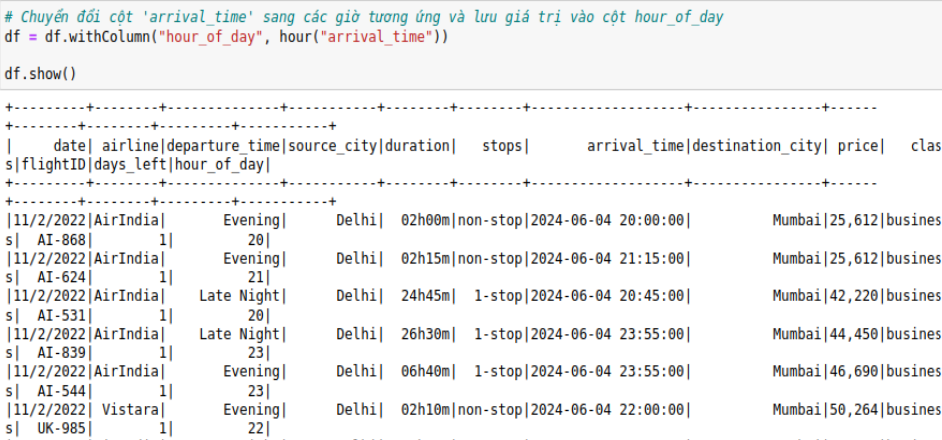
+ B1: Chuyển đổi giá trị trong cột “departure\_time” thành các giờ tương ứng. Tạo một cột “hour\_of\_day” lưu trữ giá trị giờ của “departure\_time” làm trung gian để chuyển đổi giá trị “departure\_time”.



+ B2: Lập các điều kiện khoảng thời gian cho “hour\_of\_day” để thay đổi giá trị của cột “departure\_time” thành các buổi trong ngày tương ứng.



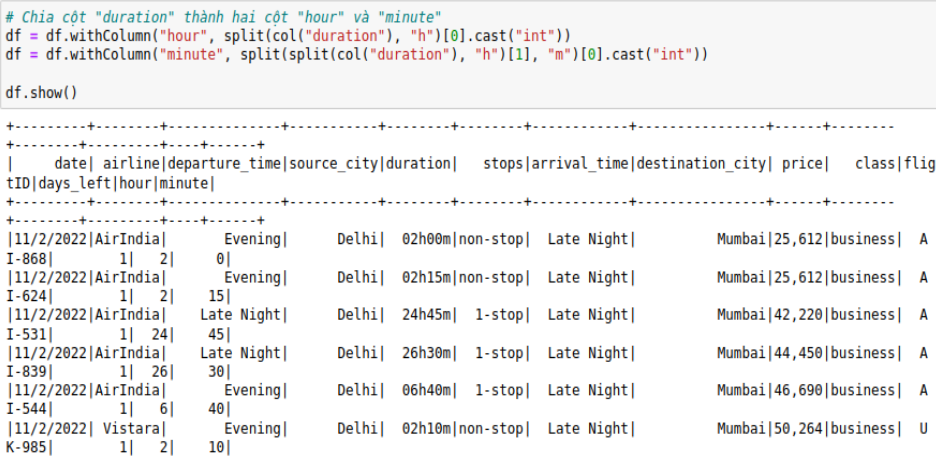
* Chuyển đổi “arrival\_time” tương đương chuyển đổi “departure\_time”





Chuyển đổi tổng thời gian di chuyển giữa các thành phố thành số thập phân tính theo giờ

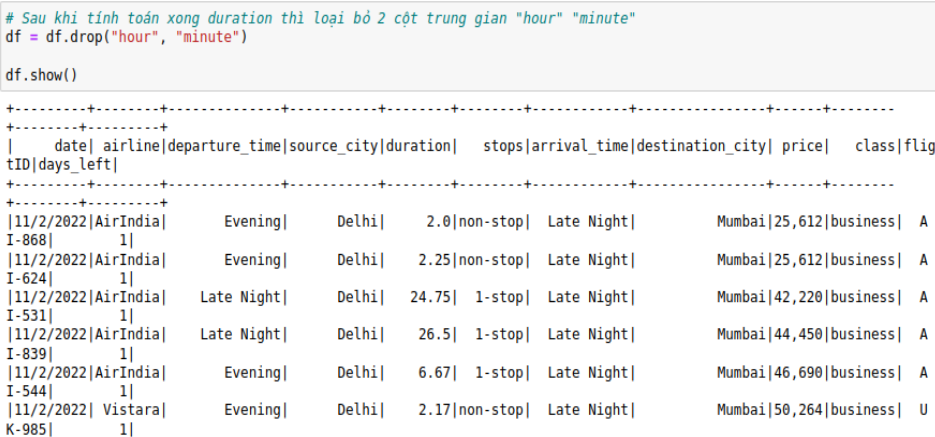
* B1: Chia giá trị cột “duration” thành hai cột “hour” “minute”



* B2: Chuyển đổi cột “minute” thành interger để tính toán đổi sang giờ. Tính lại giá trị của cột “duration”



* B3: Sau khi tính toán xong thì loại bỏ các cột không cần



Chuyển đổi giá trị cột stops thành các số nguyên với:

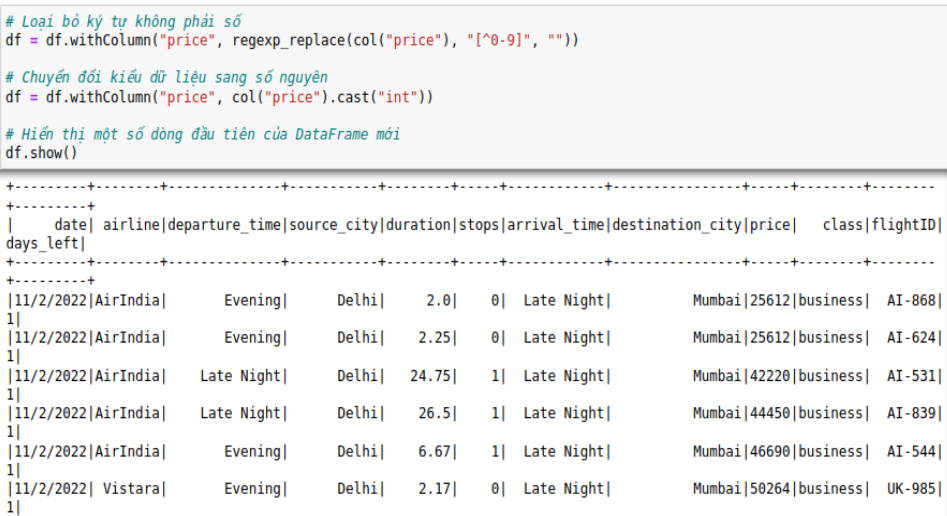
Non\_stop 🡪 0

1\_stop 🡪 1

2\_stop 🡪 2

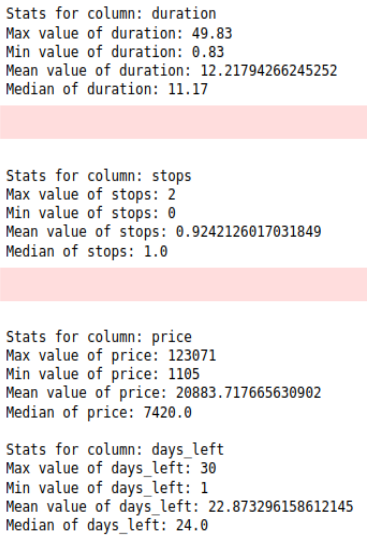


Chuyển đổi cột “price” thành kiểu dữ liệu số nguyên

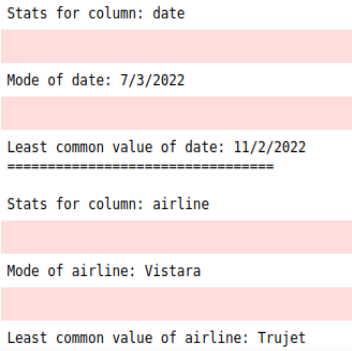


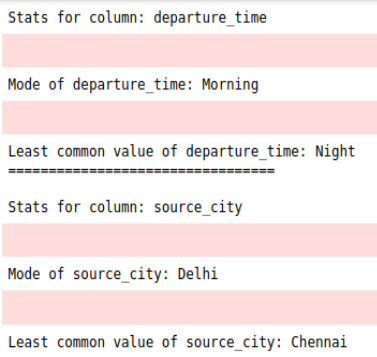
* 1. **Mô tả dữ liệu**

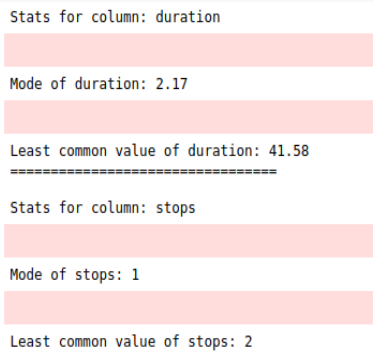
Tính toán các chỉ số cho các cột dữ liệu số “duration” “stops” “price” “day\_left”

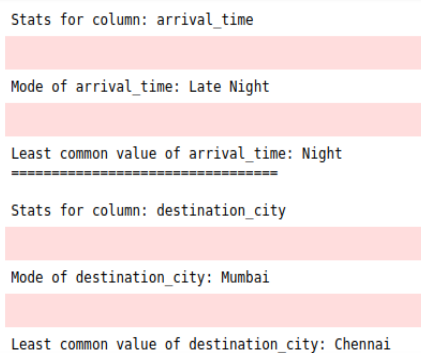


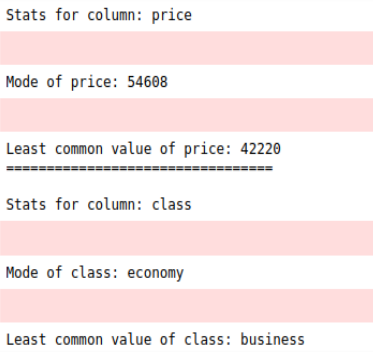
Giá trị xuất hiện nhiều nhất và giá trị hiếm nhất của các cột dữ liệu

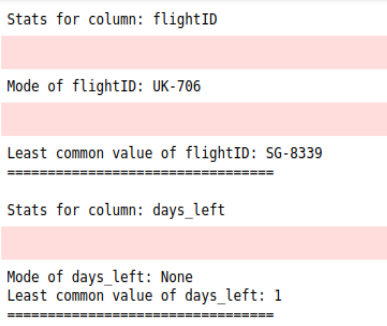






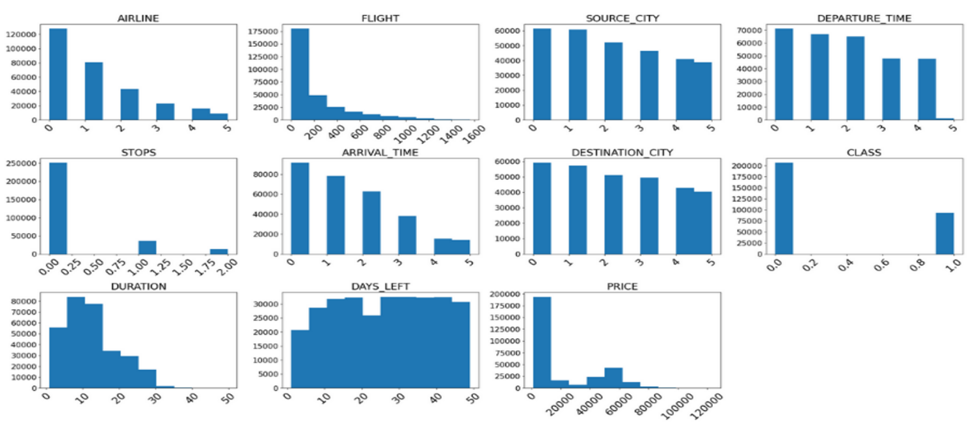




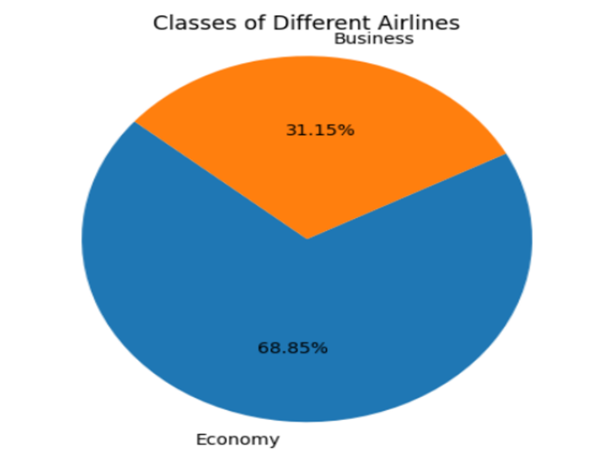


* 1. **Trực quan hóa dữ liệu**

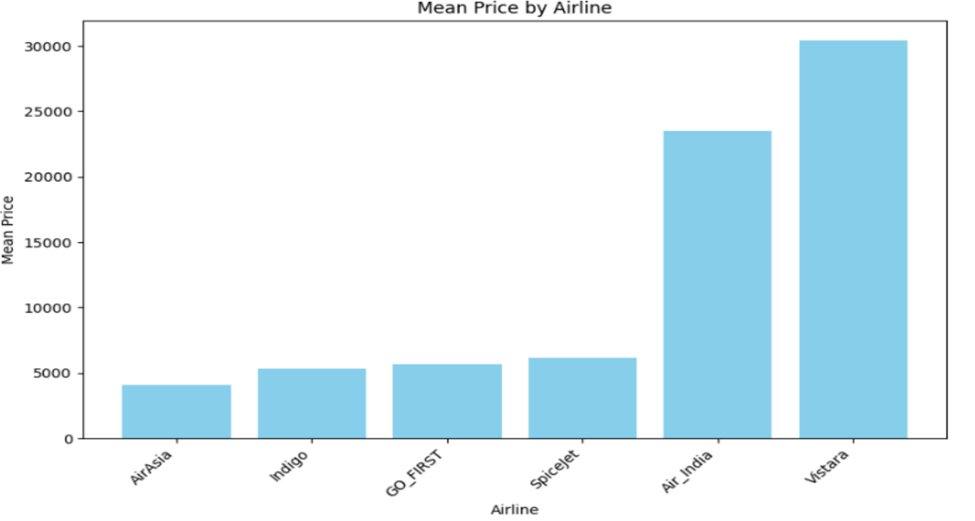
Mô hình các cột dữ liệu



Tỉ lệ khác nhau giữa hạng vé Business và Economy



Trung bình giá vé theo các hãng máy bay

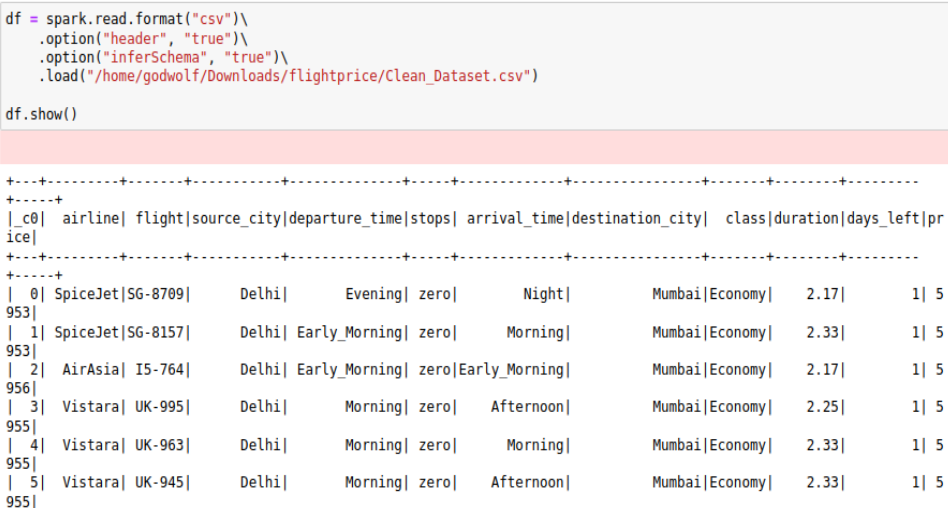


1. **Xây dựng thuật toán**

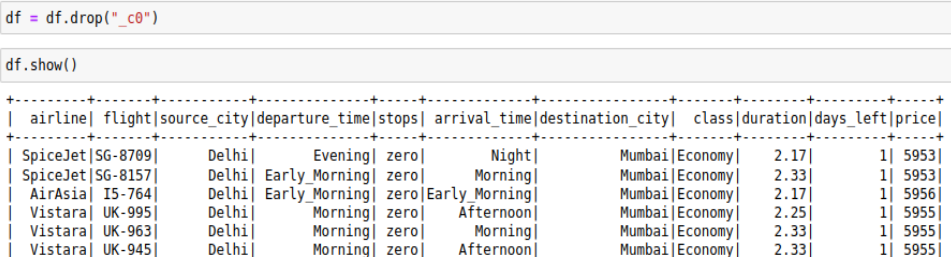
Trong quá trình xử lý dữ liệu, việc chuyển đổi các giá trị chuỗi (String) thành số là rất quan trọng, do đó, cần sử dụng phương pháp StringIndexer.

StringIndexer giúp biến đổi các giá trị phân loại (categorical) thành số nguyên bằng cách ánh xạ mỗi giá trị duy nhất trong một cột thành một số nguyên duy nhất. Nó phù hợp với thuật toán Linear Regression vì thuật toán này yêu cầu dữ liệu đầu vào ở dạng số nguyên, bao gồm cả các thuộc tính phân loại. Nhờ vậy, dữ liệu sẽ dể đọc và tương thích hơn hơn khi sử dụng các thuật toán yêu cầu dữ liệu dạng số.

Sau khi dữ liệu đã được xử lí, sử dụng tập dữ liệu mới để thực hiện dự đoán



Xóa đi cột không sử dụng “\_c0”



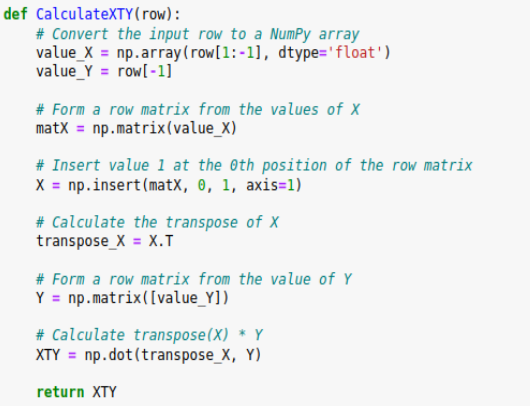
Chia tập dữ liệu thành tập huấn luyện và tập kiểm tra theo tỉ lệ 8:2



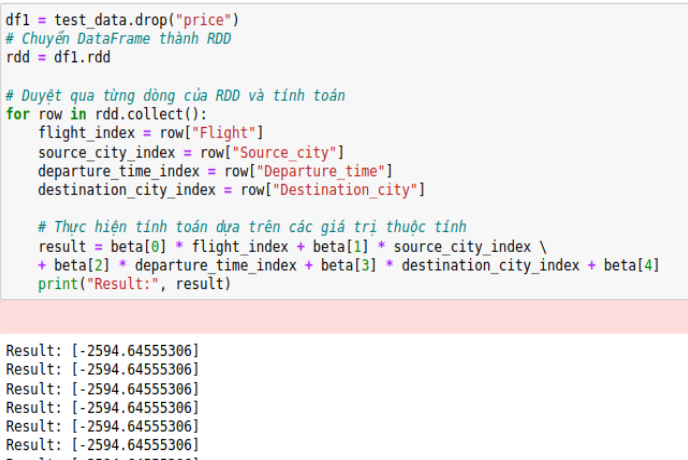
Định nghĩa hàm tính ma trân hoán vị XTX



Định nghĩa hàm tính XTY



1. **Kết quả đạt được**
   1. **Kết quả**

****

* 1. **Đánh giá**

a) Phương pháp Root Mean Squared Error (RMSE)

\* Định nghĩa:

RMSE là độ lệch chuẩn của phần dư (lỗi dự đoán). Phần dư là thước đo khoảng cách từ các điểm dữ liệu đường hồi quy. RMSE được sử dụng để đo lường sự chênh lệch giữa các giá trị dự đoán và các giá trị thực tế. Nó đo lường độ lớn của sai số trung bình giữa các giá trị dự đoán và thực tế, trong đó sai số lớn hơn sẽ ảnh hưởng lớn hơn đến kết quả cuối cùng của RMSE. RMSE thường được sử dụng trong khí hậu học, dự báo và phân tích hồi quy để xác minh kết quả thí nghiệm.

\* Công thức:

Text

Description automatically generated with low confidence

Trong đó:

- f là dự báo (giá trị dự kiến hoặc kết quả không xác định)

- o là giá trị quan sát (giá trị đã biết)

Công thức tương tự:

Shape

Description automatically generated with medium confidence

Trong đó:

- y^i: Giá trị Y tính được từ mô hình

- yi: Giá trị thực tế quan sát được

- n: Số dòng dữ liệu

Ngoài ra, có thể tính RMSE bằng cách:

- Bình phương các phần dư

- Tính trung bình của phần dư

- Lấy căn bậc hai của kết quả

RMSE là một phép đo tốt để đánh giá sự chênh lệch giữa dự đoán và giá trị thực tế trong mô hình học máy. Tuy nhiên, nó có thể dễ dàng bị ảnh hưởng bởi phạm vi của biến phụ thuộc. Khi biến phụ thuộc có phạm vi biến thiên hẹp, RMSE thường thấp hơn, và khi biến phụ thuộc có phạm vi rộng, RMSE thường cao hơn. Do đó, khi so sánh RMSE giữa các mô hình hoặc các lần lặp lại khác nhau của cùng một mô hình, cần phải cẩn thận và cân nhắc sự ảnh hưởng của phạm vi biến phụ thuộc.

b) Phương pháp Mean Absolute Error (MAE)

\* Định nghĩa:

MAE tính trung bình của giá trị tuyệt đối của sự chênh lệch giữa các giá trị dự đoán và các giá trị thực tế. Điều này có nghĩa là các sai số đều được coi trọng mà không phân biệt giữa sai số dương và sai số âm.

\* Công thức:

A black and white symbol

Description automatically generated with medium confidence

Trong đó:

- n: số lượng lỗi

- |yj – y^j |: các lỗi tuyệt đối

Cách tính:

- Tìm tất cả các lỗi tuyệt đối: yj – y^j

- Tính tổng của các lỗi tuyệt đối

- Chia cho số lượng lỗi. Ví dụ, nếu có 10 độ đo, ta chia cho 10

\* Ý nghĩa:

Khi giá trị của MAE càng nhỏ, điều này biểu thị rằng sự chênh lệch trung bình giữa các dự đoán và giá trị thực tế càng ít. Do đó, phương trình hồi quy có độ chính xác cao hơn. Ngược lại, khi MAE càng lớn, sự chênh lệch trung bình giữa dự đoán và giá trị thực tế càng lớn, dẫn đến độ chính xác của phương trình hồi quy giảm đi.

1. **Tổng kết**
   1. **Ưu điểm**

Mô hình Linear Regression giúp dự đoán giá vé máy bay dựa trên các yếu tố như thời gian bay, số lượng điểm dừng, và hãng hàng không. Điều này tạo ra một công cụ hữu ích cho nhiều đối tượng, từ hành khách muốn tìm vé rẻ, đại lý đặt vé đến nhà quản lý của các hãng hàng không và giúp các nhà quản lý hãng hàng không hiểu rõ hơn về các yếu tố ảnh hưởng đến giá vé. Nhờ đó, họ có thể điều chỉnh chiến lược giá linh hoạt hơn theo các biến động của thị trường, giúp tăng hiệu quả kinh doanh và đưa ra các mức giá cạnh tranh hơn.

* 1. **Nhược điểm**

Bên cạnh những ưu điểm kể trên, mô hình Hồi quy tuyến tính cẫn có một số hạn chế như sau:

- Độ chính xác không cao: Mô hình này chưa thể cung cấp kết quả dự đoán với độ chính xác cao, có thể do tính đơn giản của mô hình không phù hợp với việc mô tả các mối quan hệ phức tạp giữa các thuộc tính và giá vé máy bay.

- Chưa thể đánh giá về hiệu suất: Nhóm chưa thực hiện được quá trình đánh giá độ chính xác của kết quả dự đoán so với dữ liệu thực tế. Điều này dẫn đến thiếu minh bạch và khó có thể đánh giá chính xác mức độ hiệu quả của mô hình.

* 1. **Hướng phát triển**

Để khắc phục nhược điểm và nâng cao hiệu suất của mô hình Linear Regression, nhóm đề xuất các hướng phát triển sau:

- Đánh giá và kiểm soát hiệu suất: Thực hiện quá trình đánh giá đầy đủ để đo lường độ chính xác của mô hình so với dữ liệu thực tế. Điều này bao gồm sử dụng các phương pháp đánh giá như MAE và RMSE để đo lường độ chính xác của mô hình so với dữ liệu thực tế. MAE được sử dụng để đo lường trung bình của sai số tuyệt đối giữa dự đoán và giá trị thực tế, trong khi RMSE tính toán căn bậc hai của trung bình bình phương của sai số, đặt trọng số cao hơn cho các sai số lớn. Bằng cách này, có thể có cái nhìn toàn diện hơn về sự khớp giữa dự đoán và thực tế, từ đó cung cấp thông tin hữu ích để điều chỉnh và cải thiện mô hình. [4]

- Tối ưu hóa tham số: Để tối ưu hóa hiệu suất của mô hình Linear Regression, cần tiếp tục thực hiện nghiên cứu và thử nghiệm để điều chỉnh các tham số của mô hình một cách cẩn thận. Quá trình này bao gồm việc tinh chỉnh hệ số (beta), tức là các trọng số được áp dụng cho các biến đầu vào, để phản ánh chính xác mối quan hệ giữa các biến và biến mục tiêu. Ngoài ra, cần thực hiện các biện pháp để ngăn chặn hiện tượng quá mức (overfitting) của mô hình.

- Mở rộng dữ liệu đầu vào: Mở rộng dữ liệu đầu vào bằng cách thu thập và tích hợp thêm nhiều thông tin hơn, có thể sử dụng các biến mô tả phong phú hơn, hoặc thêm các biến tương quan có thể ảnh hưởng đến biến mục tiêu.

- Sử dụng mô hình phức tạp hơn: Để mô phỏng mối quan hệ phức tạp giữa các biến động và giá vé một cách hiệu quả, có thể xem xét sử dụng các mô hình dự đoán phức tạp hơn như Decision Trees, Random Forests, hoặc Neural Networks. Bằng cách sử dụng các mô hình này, mô hình sẽ có khả năng mô phỏng mối quan hệ phức tạp hơn và dự đoán mức giá vé một cách chính xác hơn trên cả dữ liệu huấn luyện và dữ liệu mới.

# **Bảng Phân Công**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Lê Ngọc Mai Thanh  20521913 | Lê Thị Đoan Trang  20522038 | Trần Minh Ngọc  20521669 | Nguyễn Thị Thảo Trang  20522040 |
| Tìm hiểu đề tài | X | X | X | X |
| Xây dựng outline | X | X | X | X |
| Viết giới thiệu tổng quan | X | X |  |  |
| Mô tả bài toán |  |  | X | X |
| Giới thiệu giải thuật |  | X | X |  |
| Triển khai trên spark | X | X | X | X |
| Xử lý dữ liệu | X | X |  | X |
| Trực quan hóa dữ liệu | X |  | X | X |
| Xây dựng thuật toán | X | X | X | X |
| % Hoàn Thành | 100% | 100% | 100% | 100% |

# **Tài liệu tham khảo**

[1] Thinh, N. T. (2023, December 20). *Đánh giá model trong Machine Learing*. Viblo.

[2] *Flight price prediction*. (2022, February 25). Kaggle.

[3] *Linear-Regression*. (n.d.). github.com.

[4] LamTien. (2020, July 29). MSE và RMSE là gì và cách tính trên STATA | Dịch vụ phân tích và xử lý số liệu. *Dịch vụ phân tích và xử lý số liệu*.