**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

🙤🙧🟍🙥🙦

A blue logo with a planet in the center

Description automatically generated

**PHÂN TÍCH CÁC YẾU TỐ ẢNH HƯỞNG VÀ XÂY DỰNG MÔ HÌNH HỖ TRỢ ĐỊNH GIÁ VÉ MÁY BAY**

|  |  |  |  |
| --- | --- | --- | --- |
| Sinh viên thực hiện: | | | |
| STT | Họ tên | MSSV | Ngành |
| 1 | Võ Chí Trường | 20522092 | CNTT-CLC |
| 2 | Võ Huỳnh Anh Vũ | 20522169 | CNTT-CLC |

**TP. HỒ CHÍ MINH – 12/2023**

# GIỚI THIỆU

Đồ án hướng đến mục tiêu phân tích các yếu tố ảnh hưởng đến giá vé máy bay như thời điểm khởi hành ,hãng hàng không, thời gian cất cánh, thời gian hạ cánh, số cân hành lý,… Để từ đó xây dựng mô hình định giá vé máy hỗ trợ cho các nhà quản trị hàng không. Với thư viện *Selenium* dữ liệu giá vé máy bay đã được thu thập tự động trên website bán vé máy bay trực tuyến. Bên cạnh đó, các thư viện *Pandas*, *Numpy* cũng được sử dụng để xử lý dữ liệu. Ngoài ra, cũng được trực quan bằng thư viện *Matplotlib*, *Seaborn* và *Power BI*. Cuối cùng, sử dụng thư viện *Scikit-learn* để xây dựng mô hình. Các mô hình Machine Learning được sử dụng bao gồm: *Random Forest, Decision Tree, Gradient Boosting Regression, Linear Regression*. Sau khi thực hiện dự án đã thu được bộ dữ liệu về chuyến bay, bao gồm các thông tin như hãng hàng không, thời gian cất cánh, thời gian hạ cánh, số cân hành lý, hạng chỗ,…Bộ dữ liệu cũng được xử lý và trực quan giúp cho việc hiểu rõ hơn về mối quan hệ giữa các yếu tố ảnh hưởng đến giá vé. Cuối cùng là xây dựng được mô hình dự đoán giá vé máy bay với độ chính xác cao (Hơn 99% theo R2\_score).

Bộ dữ liệu sử dụng trong đồ án là tự thu thập tại website *baydep.vn.*

# THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU

## Thu thập bộ dữ liệu

Bộ dữ liệu phân tích tự thu thập tại website *https://baydep.vn/* . Bộ dữ liệu đã thu thập là thông tin chuyến bay trong một năm (từ 1/1/2024 đến 26/10/2024) của đường bay từ Thành phố Hồ Chí Minh (SGN) đi Hà Nội (HAN) cho chuyến bay thẳng một chiều. Với bộ dữ liệu ban đầu gồm 31929 dòng và 14 cột.

Sau đây là bảng mô tả bộ dữ liệu:

|  |  |  |  |
| --- | --- | --- | --- |
| **Cột** | **Mô tả ngắn** | **Minh họa** | **Dữ liệu Null** |
| flight-numb | ID chuyến bay | VJ198, VN204,… | 0 |
| airlines-name | Tên hãng hàng không | Vietjet Air, Vietnam Airlines,… | 0 |
| **price** | Giá vé máy bay – Biến cần  dự đoán | 1.959.500, 2.706.000,… | 1 |
| depart-airport | Sân bay khởi hành | SGN | 0 |
| arrival-airport | Sân bay đến | HAN | 0 |
| depart-hour | Thời gian cất cánh | 5:15, 16:00,… | 0 |
| arrival-hour | Thời gian hạ cánh | 7:25, 18:10,… | 0 |
| depart-date | Ngày khởi hành | 1/1/2024 đến 26/10/2024 | 0 |
| arrival-date | Ngày đến | 1/1/2024 đến 26/10/2024 | 0 |
| flight-time | Thời gian bay | 2h 10p, 2h 15p,… | 0 |
| seat-class | Hạng chỗ ngồi | Economy (B1\_ECO), Economy (V\_SBoss),… | 0 |
| plane | Loại máy bay | Airbus A350-900,… | 0 |
| cabin-baggage | Hành lý xách tay (số cân hành lý được mang lên máy bay đi kèm theo giá vé)(tính theo đơn vị kg) | 7kg, 12kg,… | 0 |
| checked-baggage | Hành lý ký gửi (số cân hành lý có thể ký gửi đi kèm theo giá vé)(tính theo đơn vị kg) | 20kg, 1 kiện x 23kg,… | 0 |

**2.2. Phương pháp thu thập dữ liệu**

Bộ dữ liệu được thu thập tự động bằng thư viện Selenium.

A diagram of a flowchart

Description automatically generated*Hình 1: Quy trình thu thập dữ liệu*

**2.3. Tiền xử lý dữ liệu**

* Các bước tiền xử lý dữ liệu:
* Xử lý cột **price** (giá): Thay thế giá trị null dòng 23291 bằng dữ liệu trên website. Sau đó đồng nhất dữ liệu cho các giá trị '901.0', '966.0', '825', '834', '879'. Cuối cùng chuyển kiểu dữ liệu thành kiểu “float”.
* Điền khuyết cột dữ liệu **flight-time** (thời gian bay): với hãng hàng không Bamboo Airways cột **flight time** là *null*. Tiến hành điền khuyết bằng cách lấy *“arrival-hour (giờ hạ cánh)”* – (trừ) *“depart-hour (giờ cất cánh)”*. Sau đó chuyển về định dạng phút (kiểu số nguyên).
* Tách cột **depart-hour** (Thời gian cất cánh)thành các cột **Dep\_hour** (Giờ cất cánh), **Dep\_min** (Phút cất cánh). Sau đó chuyển kiểu dữ liệu thành kiểu “int”.
* Tách cột **arrival-hour** (Thời gian hạ cánh)thành các cột **Arr\_hour** (Giờ hạ cánh), **Arr\_min** (Phút hạ cánh). Sau đó chuyển kiểu dữ liệu thành kiểu “int”.
* Tách cột **depart-date** (Ngày khởi hành) thành **Dep\_Date** (Ngày khởi hành)**, Dep\_Month** (Tháng khởi hành)**, Dep\_Year** (Năm khởi hành). Sau đó chuyển kiểu dữ liệu thành kiểu “int”.
* Tách cột **arrival-date** (Ngày đến) thành **Arr\_Date** (Ngày đến)**, Arr\_Month** (Tháng đến)**, Arr\_Year** (Năm đến). Sau đó chuyển kiểu dữ liệu thành kiểu “int”.
* Đồng nhất dữ liệu cột **cabin-baggage** (hành lý xách tay) và **checked-baggage** (hành lý kí gửi): Xóa bỏ các kí tự là chuỗi, sau đó chuyển kiểu dữ liệu thành kiểu “int”.
* Tách cột **seat-class** thành các cột **rank\_seat\_class** (Hạng chỗ ngồi), **seat\_class\_type** (Loại ghế ngồi).
* Thêm cột **Total\_Stops** (Số điểm dừng trong chặng bay).
* Thêm cột **Total\_Passenger** (Số lượng hành khách trong vé).

# MÔ TẢ BỘ DỮ LIỆU

* Sau khi xử lý dữ liệu thu được bộ dữ liệu gồm 31929 dòng và 23 cột, dưới đây là bảng mô tả:

|  |  |  |  |
| --- | --- | --- | --- |
| **Cột** | **Mô tả ngắn** | **Kiểu dữ liệu** | **Minh họa** |
| flight-numb | ID chuyến bay | **object** | VJ198, VN204,… |
| airlines-name | Tên hãng hàng không | **object** | Vietjet Air, Vietnam Airlines,… |
| price | Giá vé máy bay – Biến cần dự đoán | **float64** | 1959500, 869000,… |
| depart-airport | Sân bay khởi hành | **object** | SGN |
| arrival-airport | Sân bay đến | **object** | HAN |
| Dep\_hour | Giờ cất cánh | **int64** | 5, 6, 7,… |
| Dep\_min | Phút cất cánh | **int64** | 0, 15, 30,… |
| Arr\_hour | Giờ hạ cánh | **int64** | 7, 8, 9,… |
| Arr\_min | Phút hạ cánh | **int64** | 10, 25, 40,… |
| Dep\_date | Ngày khởi hành | **int64** | 1, 2, 3, 4, … |
| Dep\_month | Tháng khởi hành | **int64** | 1, 2,…,9, 10 |
| Dep\_year | Năm khởi hành | **int64** | 2024 |
| Arr\_date | Ngày đến | **int64** | 1, 2, 3, 4,… |
| Arr\_month | Tháng đến | **int64** | 1, 2,…,9, 10 |
| Arr\_year | Năm đến | **int64** | 2024 |
| flight-time | Thời gian bay (tính theo đơn vị phút) | **int64** | 130, 135, 140,… |
| plane | Loại máy bay | **object** | Airbus A321,… |
| cabin-baggage | Hành lý xách tay (số cân hành lý được mang lên máy bay đi kèm theo giá vé)(tính theo đơn vị kg) | **int64** | 7, 10, 12, 18 |
| checked-baggage | Hành lý ký gửi (số cân hành lý có thể ký gửi đi kèm theo giá vé)(tính theo đơn vị kg) | **int64** | 0, 20, 23, 30, 32 |
| rank\_seat | Hạng chỗ ngồi (tiết kiệm, phổ thông, thương gia, ….) | **object** | economy, business classic,… |
| seat\_class | Loại ghế ngồi | object | a, a1\_eco, b, c,… |
| Total\_Stops | Số điểm dừng trong chặng bay | int64 | 1 |
| Total\_Passenger | Số lượng hành khách trong vé | int64 | 1 |

# PHÂN TÍCH THĂM DÒ

## Thống kê mô tả

* Các giá trị đặc trưng cho biến price:
* Mean: 2563721.444
* Median: 2207000.0
* Mode: 3546000.0

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | price | Dep\_hour | Dep\_min | Arr\_hour | Arr\_min | Dep\_Date | Dep\_Month | Dep\_Year |
| count | 31929.0 | 31929.0 | 31929.0 | 31929.0 | 31929.0 | 31929.0 | 31929.0 | 31929.0 |
| mean | 2,56372E+15 | 1,39257E+16 | 2,32716E+16 | 1,37538E+16 | 2,3276E+15 | 1,56274E+16 | 4,83751E+14 | 2024.0 |
| std | 1,23343E+16 | 5,18025E+15 | 2,07209E+16 | 6,2357E+16 | 1,7824E+16 | 8,5491E+15 | 2,97697E+15 | 0.0 |
| min | 800000.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 2024.0 |
| 25% | 1635500.0 | 10.0 | 0.0 | 100 | 10.0 | 8.0 | 2.0 | 2024.0 |
| 50% | 2207000.0 | 13.0 | 25.0 | 16.0 | 15.0 | 16.0 | 4.0 | 2024.0 |
| 75% | 3546000.0 | 18.0 | 45.0 | 18.0 | 40.0 | 23.0 | 8.0 | 2024.0 |
| max | 8762000.0 | 23.0 | 55.0 | 23.0 | 55.0 | 31.0 | 10.0 | 2024.0 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Arr\_Date | Arr\_Month | Arr\_Year | flight-time | cabin-baggage(kg) | checked-baggage(kg) | Total\_Stops | Total\_Passenger |
| count | 31929.0 | 31929.0 | 31929.0 | 31929.0 | 31929.0 | 31929.0 | 31929.0 | 31929.0 |
| mean | 1,56489E+16 | 4,84005E+15 | 2024.0 | 1,31713E+16 | 9,05813E+14 | 1,62446E+16 | 1.0 | 1.0 |
| std | 8,55894E+15 | 2,9769E+15 | 0.0 | 3,44946E+15 | 2,38044E+15 | 1,14783E+16 | 0.0 | 0.0 |
| min | 1.0 | 1.0 | 2024.0 | 115.0 | 7.0 | 0.0 | 1.0 | 1.0 |
| 25% | 8.0 | 2.0 | 2024.0 | 130.0 | 7.0 | 0.0 | 1.0 | 1.0 |
| 50% | 16.0 | 4.0 | 2024.0 | 130.0 | 7.0 | 20.0 | 1.0 | 1.0 |
| 75% | 23.0 | 8.0 | 2024.0 | 135.0 | 12.0 | 23.0 | 1.0 | 1.0 |
| max | 31.0 | 10.0 | 2024.0 | 150.0 | 18.0 | 32.0 | 1.0 | 1.0 |

**4.2. Tương quan Pearson**

* Sử dụng độ tương quan pearson để tìm các giá trị tương quan của các biến số sau:
* Biến số ảnh hưởng nhiều nhất là: checked-baggage(kg) (Pearson Correlation: 0.75)
* Biến số ảnh hưởng ít nhất là: Dep\_min (Pearson Correlation: -0.26)
* Các biến số ảnh hưởng là: 'checked-baggage(kg)', 'cabin-baggage(kg)', 'Dep\_min', 'flight-time', 'Dep\_Month', 'Arr\_Month', 'Arr\_Date', 'Dep\_Date', 'Arr\_min', 'Dep\_hour', 'Arr\_hour'.

**4.3. Tương quan ANOVA:**

* Sử dụng tương quan ANOVA để tìm các giá trị tương quan của các biến phân loại sau:
* Biến phân loại ảnh hưởng nhiều nhất: seat-class-type (F-test: 17322.46).
* Biến phân loại ít ảnh hưởng nhất: flight-numb (F-test: 86.77).
* Các biến phân loại ảnh hưởng là: 'flight-numb', 'airlines-name', 'plane', 'rank-seat-class', 'seat-class-type'.

**4.4. Trực quan (Dashboard):**

* Sử dụng công cụ Power BI để trực quan các yếu tố ảnh hưởng đến giá vé máy bay. Thống kê được số lượng chuyến bay mỗi hãng khai thác theo từng tháng, biểu đồ giá vé các ngày trong tháng từ đó cho biết giá vé tốt nhất trong tháng, A screenshot of a graph

  Description automatically generatedgiá vé thấp nhất của từng hãng, từng hạng vé, từng hạng ghế.

*Hình 2: Trực quan các yếu tố ảnh hưởng đến giá vé máy bay*

# PHÁT TRIỂN MÔ HÌNH VÀ ĐÁNH GIÁ

**Phương pháp (Methodology):**

Quy trình Phân tích dữ liệu cho bộ dữ liệu chuyến bay:

A diagram of a data wring

Description automatically generated

*Hình 3. Quy trình PTDL.*

* Sau khi tìm được những biến ảnh hưởng dựa vào tương quan Pearson và tương quan ANOVA thì các thuộc tính sau được chọn để đưa vào mô hình: 'flight-numb', 'airlines-name', 'price', 'Dep\_hour', 'Dep\_min', 'Arr\_hour', 'Arr\_min', 'Dep\_Date', 'Dep\_Month', 'Arr\_Date', 'Arr\_Month', 'flight-time', 'plane', 'cabin-baggage(kg)', 'checked-baggage(kg)', 'rank-seat-class', 'seat-class-type'.

**Thiết kế thực nghiệm (Empirical design):**

* Chuẩn bị dữ liệu:
* Biến đầu vào bao gồm các biến số: 'Dep\_hour', 'Dep\_min', 'Arr\_hour', 'Arr\_min', 'Dep\_Date', 'Dep\_Month', 'Arr\_Date', 'Arr\_Month', 'flight-time', 'cabin-baggage(kg)', 'checked-baggage(kg)' và các biến phân loại: 'airlines-name', 'flight-numb', 'plane', 'rank-seat-class', 'seat-class-type'.
* Biến mục tiêu: ‘price’.
* Chuẩn hóa dữ liệu: Sử dụng StandardScaler để chuẩn hóa các biến số. Sử dụng OneHotEncoder để chuẩn hóa các biến phân loại.
* Chia dữ liệu thành tập huấn luyện và kiểm tra theo tỉ lệ 7:3 (70% là dữ liệu huấn luyện và 30% là dữ liệu kiểm tra).
* Lựa chọn và xây dựng mô hình: Qua quá trình phân tích, các mô hình: Linear Regression, Random Forest, Decision Tree, Gradient Boosting Regression, SVR được lựa chọn. Xây dựng các mô hình dựa trên phương pháp sử dụng Pipeline.

**Kết quả thực nghiệm (Empirical results):**

* Sau khi huấn luyện các mô hình, kết quả thu được từ mô hình “Decision Tree” và “Random Forest” là 2 mô hình tốt nhất với r2\_score lần lượt là: 99.99 %, 99.98%. Vì vậy mô hình Decision Tree và Random Forest được lựa chọn làm mô hình chính cho việc dự đoán giá của các chuyến bay.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Linear Regression | Decision Tree | Gradient  Boosting Regression | Random Forest | SVR |
| r2\_score | 0.9969 | 0.9999 | 0.9870 | 0.9999 | -0.0774 |
| MSE | 4682827145.82 | 170893446.79 | 19827768900.30 | 146652542.73 | 1638722996658.62 |
| RMSE | 68431.04 | 13413.83 | 140780.94 | 13172.29 | 1280126.43 |
| MAE | 15206.53 | 293.77 | 89276.11 | 563.29 | 1063153.56 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Linear Regression | Decision Tree | Gradient  Boosting Regression | Random Forest | SVR |
| r2\_score | 0.9974 | 0.9999 | 0.9864 | 0.9998 | -0.0669 |
| MSE | 3924672119.05 | 38969125.17 | 20726740783.47 | 158754449.93 | 1632846583718.02 |
| RMSE | 62647.2 | 6242.52 | 143967.85 | 12599.78 | 1277828.85 |
| MAE | 14592.2 | 103.5 | 90533.32 | 803.02 | 1061145.98 |

* Kiểm chứng chéo (k-Fold Cross-Validation):
* Có thể thấy sau khi thực hiện kiểm chứng chéo thì 2 mô hình Random Forest và Decision Tree vẫn đạt hiệu suất rất cao. Về tổng thể có 4 mô hình có R2\_score cao (trên 98%) nhưng chỉ có 2 mô hình (Random Forest và Decision Tree) có sai số dự đoán thấp hơn so với các mô hình còn lại.

# KẾT LUẬN

Đồ án đã đạt được mục tiêu đề ra là phân tích các yếu tố ảnh hưởng đến giá vé máy bay và xây dựng mô hình dự đoán giá vé máy bay hỗ trợ cho các nhà quản trị hàng không. Trong quá trình thực hiện, nhóm đã sử dụng thư viện *Selenium* để thu thập dữ liệu tự động về các chuyến bay từ website bán vé máy bay trực tuyến. Sau đó, đã áp dụng được những kỹ thuật đã học trong việc sử dụng các thư viện *Pandas, Numpy, Matplotlib, Seaborn* và công cụ *Power BI* để xử lý, phân tích và trực quan hóa dữ liệu. Cuối cùng, sử dụng thư viện *Scikit-Learn* để xây dựng mô hình dự đoán giá vé máy bay.

Kết quả thu được là bộ dữ liệu về chuyến bay đã được thu thập và xử lý thành công. Mô hình dự đoán giá vé máy bay có độ chính xác cao (Cao nhất là Decision Tree và Random Forest với độ chính xác R2\_score lần lượt là 99.99% và 99.98%). Những kết quả này đã góp phần xác định các yếu tố ảnh hưởng đến giá vé máy bay và hỗ trợ cho các nhà quản trị hàng không trong việc đưa ra các quyết định định giá vé máy bay một cách hiệu quả.

# MỞ RỘNG (LÀM THÊM)

## Thu thập và tiền xử lý dữ liệu

**a) Thu thập dữ liệu**

* Bộ dữ liệu phân tích tự thu thập tại website *https://baydep.vn/* . Bộ dữ liệu đã thu thập là thông tin chuyến bay trong 3 tháng đầu năm 2024 (từ 1/1/2024 đến 31/3/2024) của các đường bay nội địa bao gồm:

1. Tp.HCM (SGN) đến Hà Nội (HAN).
2. Tp.HCM (SGN) đến Phú Quốc (PQC).
3. Tp.HCM (SGN) đến Đà Nẵng (DAD).
4. Tp.HCM (SGN) đến Đà Lạt (CLI).
5. Tp.HCM (SGN) đến Nha Trang (CXR).
6. Phú Quốc (PQC) đến Tp.HCM (SGN).
7. Đà Nẵng (DAD) đến Tp.HCM (SGN).
8. Đà Lạt (CLI) đến Tp.HCM (SGN).
9. Nha Trang (CXR) đến Tp.HCM (SGN).

* Với bộ dữ liệu ban đầu gồm 32755 dòng và 14 cột. Dưới đây là bảng mô tả bộ dữ liệu:

|  |  |  |  |
| --- | --- | --- | --- |
| **Cột** | **Mô tả ngắn** | **Minh họa** | **Dữ liệu Null** |
| flight-numb | ID chuyến bay | VJ198, VN1829,… | 0 |
| airlines-name | Tên hãng hàng không | Vietjet Air, Vietnam Airlines,… | 0 |
| **price** | Giá vé máy bay – Biến cần  dự đoán | 942.000, 1.544.000,… | 14 |
| depart-airport | Sân bay khởi hành | SGN, DAD, CLI,… | 0 |
| arrival-airport | Sân bay đến | HAN, DAD, PQC,… | 0 |
| depart-hour | Thời gian cất cánh | 20:50, 16:00,… | 0 |
| arrival-hour | Thời gian hạ cánh | 16:35, 18:10,… | 0 |
| depart-date | Ngày khởi hành | 1/1/2024 đến 31/3/2024 | 0 |
| arrival-date | Ngày đến | 1/1/2024 đến 31/3/2024 | 0 |
| flight-time | Thời gian bay | 1h 10p, 2h 15p, 0h 50p,… | 0 |
| seat-class | Hạng chỗ ngồi | Economy (B1\_ECO), Economy (V\_SBoss),… | 0 |
| plane | Loại máy bay | Airbus A350-900,… | 13 |
| cabin-baggage | Hành lý xách tay (số cân hành lý được mang lên máy bay đi kèm theo giá vé)(tính theo đơn vị kg) | 7kg, 12kg,… | 13 |
| checked-baggage | Hành lý ký gửi (số cân hành lý có thể ký gửi đi kèm theo giá vé)(tính theo đơn vị kg) | 20kg, 1 kiện x 23kg,… | 13 |

1. **Tiền xử lý dữ liệu**

* Các bước tiền xử lý dữ liệu tương tự so với bộ dữ liệu cũ. Tuy nhiên, có một số thay đổi sau:
* Xử lý cột **price** (giá): Drop những giá trị null. Sau đó đồng nhất dữ liệu cho các giá trị '825', '834', '879'. Cuối cùng chuyển kiểu dữ liệu thành kiểu “float”.
* Xử lý cột **plane** (loại máy bay), **cabin-baggage** (hành lý xách tay), **checked-baggage** (hành lý kí gửi): Drop những giá trị null.

## Mô tả dữ liệu

* Sau khi xử lý dữ liệu thu được bộ dữ liệu gồm 32741 dòng và 23 cột

|  |  |  |  |
| --- | --- | --- | --- |
| **Cột** | **Mô tả ngắn** | **Kiểu dữ liệu** | **Minh họa** |
| flight-numb | ID chuyến bay | **object** | VJ198, VN1829,… |
| airlines-name | Tên hãng hàng không | **object** | Vietjet Air, Vietnam Airlines,… |
| **price** | Giá vé máy bay – Biến cần  dự đoán | **float64** | 942000, 1544000,… |
| depart-airport | Sân bay khởi hành | **object** | SGN, DAD, CLI,… |
| arrival-airport | Sân bay đến | **object** | HAN, DAD, PQC,… |
| Dep\_hour | Giờ cất cánh | **int64** | 5, 16, 20,… |
| Dep\_min | Phút cất cánh | **int64** | 0, 15, 30,… |
| Arr\_hour | Giờ hạ cánh | **int64** | 7, 16, 18,… |
| Arr\_min | Phút hạ cánh | **int64** | 10, 25, 40,… |
| Dep\_date | Ngày khởi hành | **int64** | 1, 2, 3, 4, … |
| Dep\_month | Tháng khởi hành | **int64** | 1, 2, 3 |
| Dep\_year | Năm khởi hành | **int64** | 2024 |
| Arr\_date | Ngày đến | **int64** | 1, 2, 3, 4,… |
| Arr\_month | Tháng đến | **int64** | 1, 2, 3 |
| Arr\_year | Năm đến | **int64** | 2024 |
| flight-time | Thời gian bay (tính theo đơn vị phút) | **int64** | 55, 70, 130,… |
| plane | Loại máy bay | **object** | Airbus A321,… |
| cabin-baggage | Hành lý xách tay (số cân hành lý được mang lên máy bay đi kèm theo giá vé)(tính theo đơn vị kg) | **int64** | 7, 10, 12, 18 |
| checked-baggage | Hành lý ký gửi (số cân hành lý có thể ký gửi đi kèm theo giá vé)(tính theo đơn vị kg) | **int64** | 0, 20, 23, 30, 32 |
| rank\_seat | Hạng chỗ ngồi (tiết kiệm, phổ thông, thương gia, ….) | **object** | economy, business classic,… |
| seat\_class | Loại ghế ngồi | **object** | a, a1\_eco, b, c,… |
| Total\_Stops | Số điểm dừng trong chặng bay | **int64** | 1 |
| Total\_Passenger | Số lượng hành khách trong vé | **int64** | 1 |

## Phân tích thăm dò

1. ***Thống kê mô tả***

* Các giá trị đặc trưng cho biến price:
* Mean: 1975322.32
* Median: 1713000.0
* Mode: 3546000.0

1. ***Tương quan Pearson***

* Sử dụng độ tương quan pearson để tìm các giá trị tương quan của các biến số sau:
* Biến số ảnh hưởng nhiều nhất là: flight-time (Pearson Correlation: 0.58)
* Biến số ảnh hưởng ít nhất là: Dep\_min (Pearson Correlation: -0.15)
* Các biến số ảnh hưởng là: 'flight-time', 'checked-baggage(kg)', 'cabin-baggage(kg)', 'Dep\_min', 'Arr\_min', 'Arr\_Date', 'Dep\_Date', 'Dep\_Month', 'Arr\_Month', 'Arr\_hour', 'Dep\_hour'.

1. ***Tương quan ANOVA***

* Sử dụng tương quan ANOVA để tìm các giá trị tương quan của các biến phân loại sau:
* Biến phân loại ảnh hưởng nhiều nhất: arrival-airport (F-test: 3057.88).
* Biến phân loại ít ảnh hưởng nhất: flight-numb (F-test: 86.45).
* Các biến phân loại ảnh hưởng là: 'flight-numb', 'airlines-name', 'depart-airport', 'arrival-airport', 'plane', 'rank-seat-class', 'seat-class-type'.

## Phát triển mô hình và đánh giá

***Phương pháp (Methodology):***

Sau khi tìm được những biến ảnh hưởng dựa vào tương quan Pearson và tương quan ANOVA thì các thuộc tính sau được chọn để đưa vào mô hình: 'flight-numb', 'airlines-name', 'price', 'depart-airport', 'arrival-airport', 'Dep\_hour', 'Dep\_min', 'Arr\_hour', 'Arr\_min', 'Dep\_Date', 'Dep\_Month', 'Arr\_Date', 'Arr\_Month', 'flight-time', 'plane', 'cabin-baggage(kg)', 'checked-baggage(kg)', 'rank-seat-class', 'seat-class-type'.

***Thiết kế thực nghiệm (Empirical design):***

Tương tự so với bộ dữ liệu cũ.

***Kết quả thực nghiệm (Empirical results):***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Linear Regression** | **Decision Tree** | **Gradient**  **Boosting Regression** | **Random Forest** | **SVR** |
| **r2\_score** | 0.9775 | 0.9970 | 0.9601 | 0.9983 | -0.0745 |
| **MSE** | 20535166360.39 | 2708576224.17 | 36481238028.99 | 1544153627.45 | 982777622078.45 |
| **RMSE** | 143300.96 | 52043.98 | 191000.62 | 39295.72 | 991351.41 |
| **MAE** | 90593.54 | 2978.88 | 141927.36 | 3701.82 | 696966.52 |

* Kiểm chứng chéo (k-Fold Cross-Validation):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Linear Regression** | **Decision Tree** | **Gradient**  **Boosting Regression** | **Random Forest** | **SVR** |
| **r2\_score** | 0.9786 | 0.9988 | 0.9624 | 0.9993 | -0.0742 |
| **MSE** | 19567528316.89 | 1120193562.91 | 34357007988.73 | 666206080.02 | 982612416000.34 |
| **RMSE** | 139884.13 | 34701.15 | 185358.27 | 26502.10 | 991268.13 |
| **MAE** | 89281.21 | 1771.07 | 140139.14 | 2597.60 | 697431.94 |

## Kết luận

Sau quá trình chỉnh sửa và cập nhật lại bộ dữ liệu thì bộ dữ liệu về chuyến bay đã được hoàn thiện. Với sự phong phú hơn trong các tuyến bay (có thêm các sân bay mới như: PQC, DAD, CLI, CXR). Nhóm vẫn tiếp tục sử dụng lại quy trình crawl dữ liệu và baseline dự đoán giá vé máy bay như trong buổi báo cáo. Tuy nhiên, với bộ dữ liệu mới phát sinh một số vấn đề khác so với bộ dữ liệu cũ nên ở bước “Sắp xếp và làm sạch dữ liệu” đã có sự thay đổi nhỏ. Mô hình sau khi được huấn luyện lại với bộ dữ liệu mới thì vẫn cho kết quả rất tốt (Với R2\_score trên 99% ở 2 mô hình Random Forest và Decision Tree). Qua đó, kết quả của đồ án này một lần nữa đã góp phần xác định các yếu tố ảnh hưởng đến giá vé máy bay và hỗ trợ cho các nhà quản trị hàng không trong việc đưa ra các quyết định định giá vé máy bay một cách hiệu quả hơn.

**TÀI LIỆU THAM KHẢO**

Bộ dữ liệu phân tích tự thu thập tại website: [*https://baydep.vn/*](https://baydep.vn/).

Đường dẫn source code hoàn chỉnh: [*https://github.com/chuongprg-it/data-flight*](https://github.com/chuongprg-it/data-flight)

**Vũ Hữu Tiệp**, *Machine Learning cơ bản, 2020.*

**Daniel Nelson**, *Data Visualization in Python, 2020*.

Selenium with Python. Link: [*https://selenium-python.readthedocs.io/*](https://selenium-python.readthedocs.io/)

(15/10/2023).

Get started with Power BI Desktop. Link: [*https://learn.microsoft.com/en-us/power-bi/fundamentals/desktop-getting-started*](https://learn.microsoft.com/en-us/power-bi/fundamentals/desktop-getting-started)

(04/12/2023).

**PHỤ LỤC PHÂN CÔNG NHIỆM VỤ**

|  |  |  |
| --- | --- | --- |
| **STT** | **Thành viên** | **Nhiệm vụ** |
| 1 | Võ Chí Trường | - Crawl data.  - Tiền xử lý dữ liệu (làm sạch và sắp xếp).  - Thăm dò dữ liệu.  - Mô hình hóa.  - Đánh giá mô hình.  - Trực quan hóa bằng python.  - Viết báo cáo, làm silde và tham gia thuyết trình. |
| 2 | Võ Huỳnh Anh Vũ | - Crawl data.  - Thăm dò dữ liệu.  - Tìm các biến ảnh hưởng.  - Lựa chọn mô hình.  - Trực quan hóa bằng PowerBI.  - Viết báo cáo, làm silde và tham gia thuyết trình. |