UIT
Trường Đại học
Công nghệ Thông tin
Khoa Khoa học
và Kỹ thuật Thông tin

THU THẬP VÀ PHÂN TÍCH DỮ LIỆU VÉ MÁY BAY

ỨNG DỤNG HỌC MÁY TRONG VIỆC RA QUYẾT ĐỊNH MUA VÉ MÁY BAY

Môn: Tiền xử lý và xây dựng bộ dữ liệu
Lớp: DS108.P21
GVHD: TS. Nguyễn Gia Tuấn Anh
CN. Trần Quốc Khánh
SVTH: Đinh Bảo Thy - 23521563
Võ Ngọc Anh Thy - 23521565
Nguyễn Vũ Thùy Trâm - 23521617

12/06/2025

Nội dung báo cáo

Tổng quan

Mục tiêu

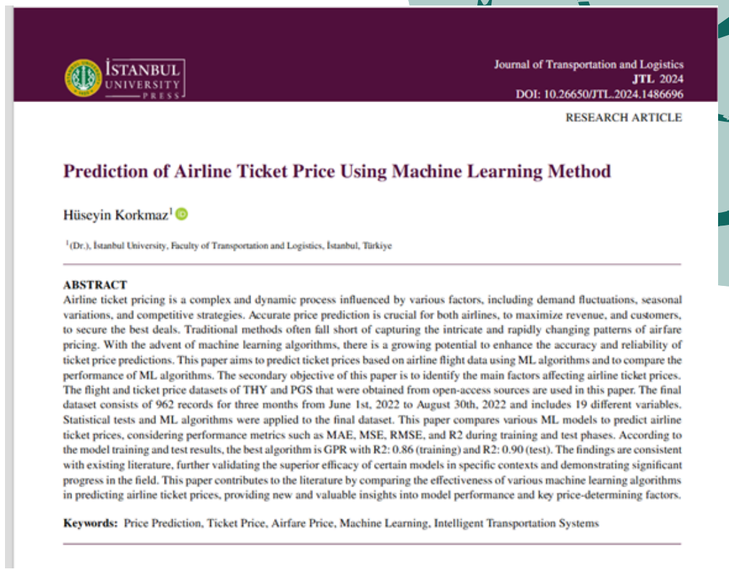
- Hệ thống dự đoán giá vé theo thời gian
- Khuyến nghị về thời điểm mua vé (mua ngay hoặc chờ)

Tại sao thu thập ?

- Tính phù hợp
- Tính cập nhật
- Tính chính xác

Khác biệt

- Bao gồm: dự đoán **+ hệ khuyến nghị**
- **Theo nhu cầu khách hàng**



İSTANBUL UNIVERSITY PRESS

Journal of Transportation and Logistics
JTL 2024
DOI: 10.26650/JTL.2024.1486696

RESEARCH ARTICLE

Prediction of Airline Ticket Price Using Machine Learning Method

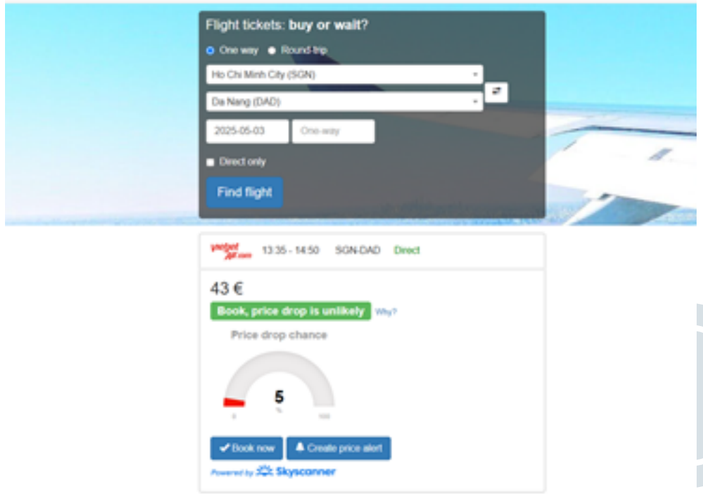
Hüseyin Korkmaz[1]

[1](Dr.), İstanbul University, Faculty of Transportation and Logistics, İstanbul, Türkiye

ABSTRACT

Airline ticket pricing is a complex and dynamic process influenced by various factors, including demand fluctuations, seasonal variations, and competitive strategies. Accurate price prediction is crucial for both airlines, to maximize revenue, and customers, to secure the best deals. Traditional methods often fall short of capturing the intricate and rapidly changing patterns of airfare pricing. With the advent of machine learning algorithms, there is a growing potential to enhance the accuracy and reliability of ticket price predictions. This paper aims to predict ticket prices based on airline flight data using ML algorithms and to compare the performance of ML algorithms. The secondary objective of this paper is to identify the main factors affecting airline ticket prices. The flight and ticket price datasets of THY and PGS that were obtained from open-access sources are used in this paper. The final dataset consists of 962 records for three months from June 1st, 2022 to August 30th, 2022 and includes 19 different variables. Statistical tests and ML algorithms were applied to the final dataset. This paper compares various ML models to predict airline ticket prices, considering performance metrics such as MAE, MSE, RMSE, and R2 during training and test phases. According to the model training and test results, the best algorithm is GPR with R2: 0.86 (training) and R2: 0.90 (test). The findings are consistent with existing literature, further validating the superior efficacy of certain models in specific contexts and demonstrating significant progress in the field. This paper contributes to the literature by comparing the effectiveness of various machine learning algorithms in predicting airline ticket prices, providing new and valuable insights into model performance and key price-determining factors.

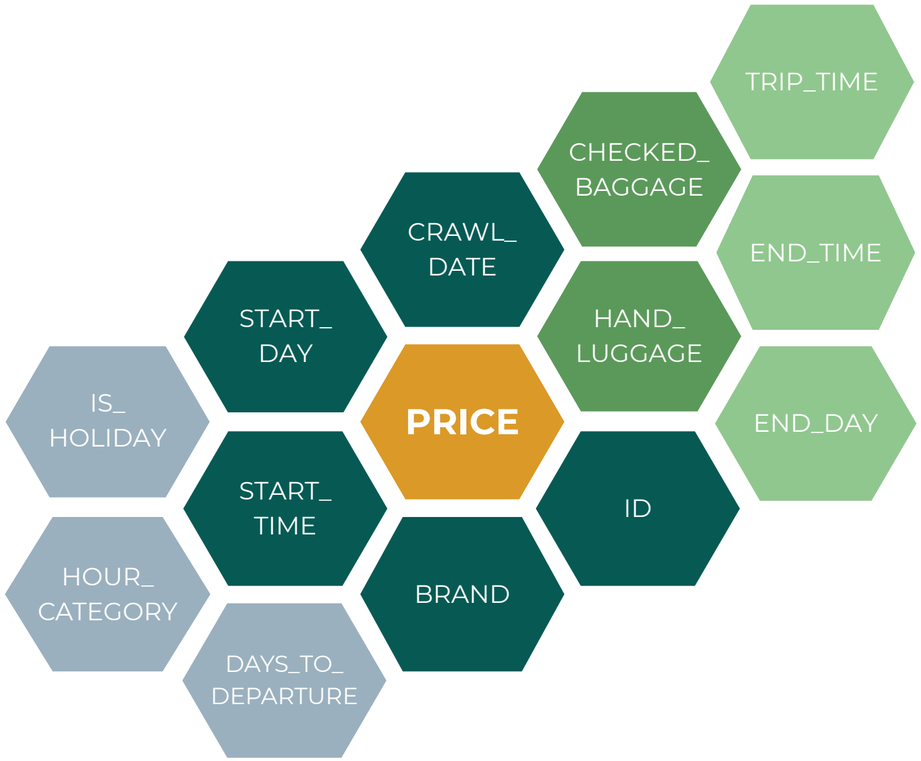
Keywords: Price Prediction, Ticket Price, Airfare Price, Machine Learning, Intelligent Transportation Systems



Bộ dữ liệu

- Website: traveloka
- Crawl dữ liệu: Selenium WebDriver
- Địa điểm: 6 thành phố lớn, du lịch ở Việt Nam
- Tính đa dạng: is_holiday

  (ngày lễ - cận lễ - cuối tuần - ngày thường)
- Bronze layer: 46,826 samples với 12 thuộc tính



brand	id	price	start_time	start_day	end_day	end_time	trip_time	destination	hand_luggage	checked_baggage	crawl_date
Bamboo Airways	PQC0002	1.245.267 VND/khách	16:30	02 thg 5	02 thg 5	17:30	1h 0m	Phú Quốc (PQC)\nSân bay Phú Quốc	Hành lý xách tay 7 kg	Hành lý 0 kg	2025-05-01
Vietravel Airlines	PQC0323	1.224.930 VND/khách	17:25	02 thg 5	02 thg 5	18:25	1h 0m	Phú Quốc (PQC)\nSân bay Phú Quốc	Hành lý xách tay 7 kg	Hành lý 0 kg	2025-05-01
Vietnam Airlines	PQC0292	1.535.463 VND/khách	17:25	02 thg 5	02 thg 5	18:25	1h 0m	Phú Quốc (PQC)\nSân bay Phú Quốc	Hành lý xách tay 1 x 12 kg	Hành lý 23 kg	2025-05-01
VietJet Air	PQC0109	1.567.000 VND/khách	15:40	02 thg 5	02 thg 5	16:40	1h 0m	Phú Quốc (PQC)\nSân bay Phú Quốc	Hành lý xách tay 7 kg	Hành lý 0 kg	2025-05-01
VietJet Air	PQC0132	1.567.000 VND/khách	17:10	02 thg 5	02 thg 5	18:15	1h 5m	Phú Quốc (PQC)\nSân bay Phú Quốc	Hành lý xách tay 7 kg	Hành lý 0 kg	2025-05-01

Tiền xử lý dữ liệu

Xử lý tổng thể

- Loại bỏ các dòng trùng lặp
- Xử lý missing value

Chuẩn hóa kiểu dữ liệu

- Chuyển đổi kiểu dữ liệu phù hợp, đồng nhất định dạng cho các đặc trưng

Trích xuất đặc trưng

- Tách và lấy các thông tin quan trọng
- Phân loại thời gian theo các nhóm
- Thêm đặc trưng
- Xác định và gán nhãn các ngày đặc biệt

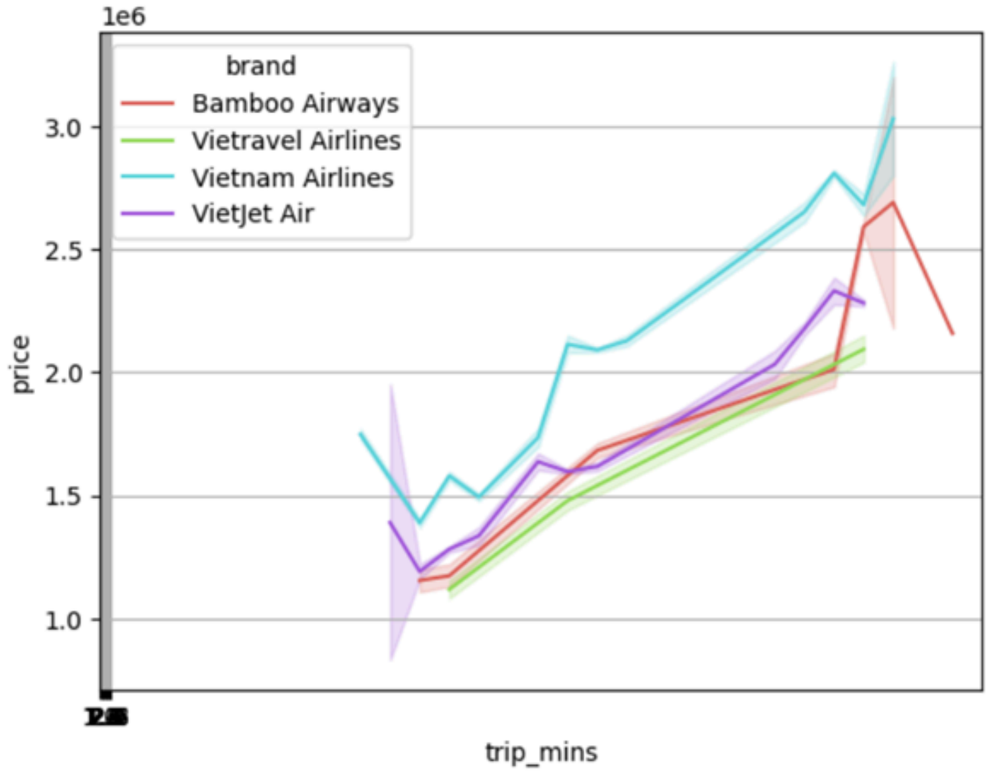
brand	id	price	start_time	start_day	end_day	end_time	trip_time	destination	hand_luggage	checked_baggage	crawl_date
Bamboo Airways	PQC0002	1.245.267 VND/khách	16:30	02 thg 5	02 thg 5	17:30	1h 0m	Phú Quốc (PQC)\nSân bay Phú Quốc	Hành lý xách tay 7 kg	Hành lý 0 kg	2025-05-01
Vietravel Airlines	PQC0323	1.224.930 VND/khách	17:25	02 thg 5	02 thg 5	18:25	1h 0m	Phú Quốc (PQC)\nSân bay Phú Quốc	Hành lý xách tay 7 kg	Hành lý 0 kg	2025-05-01

id	brand	price	destination	hand_luggage	checked_baggage	start_hour	end_hour	trip_mins	is_holiday	days_left
PQC0002	Bamboo Airways	1245267	Phú Quốc (PQC)\nSân bay Phú Quốc	7	0	Evening	Evening	60	2	1
PQC0323	Vietravel Airlines	1224930	Phú Quốc (PQC)\nSân bay Phú Quốc	7	0	Evening	Evening	60	2	1

→ Silver layer: 46,549 samples với 11 thuộc tính

Khai phá và phân tích dữ liệu:

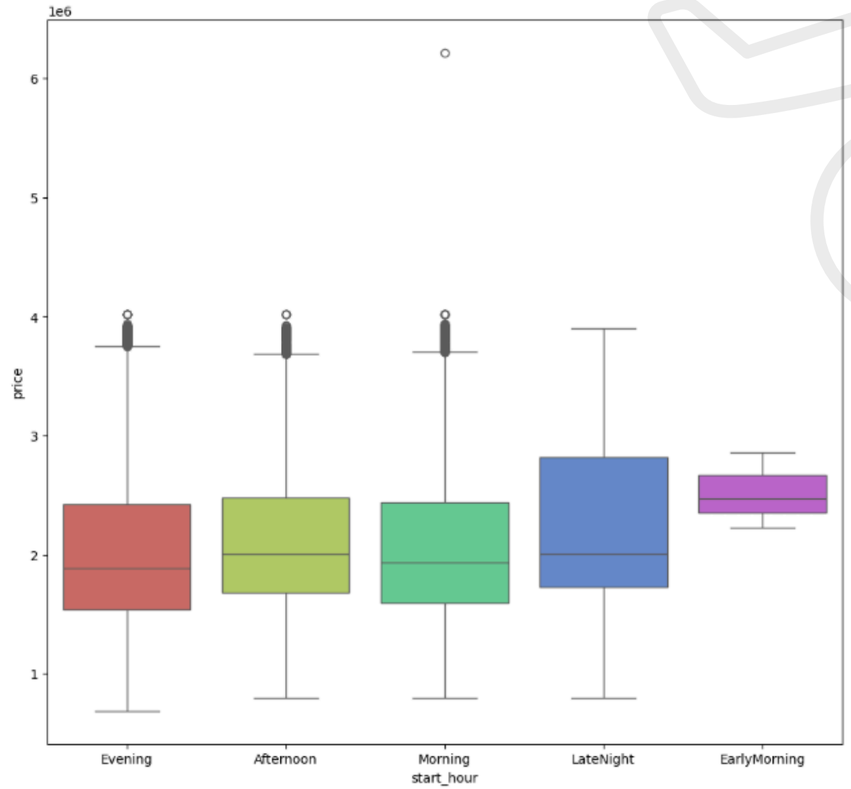
- Độ tương quan giữa các dữ liệu dạng số



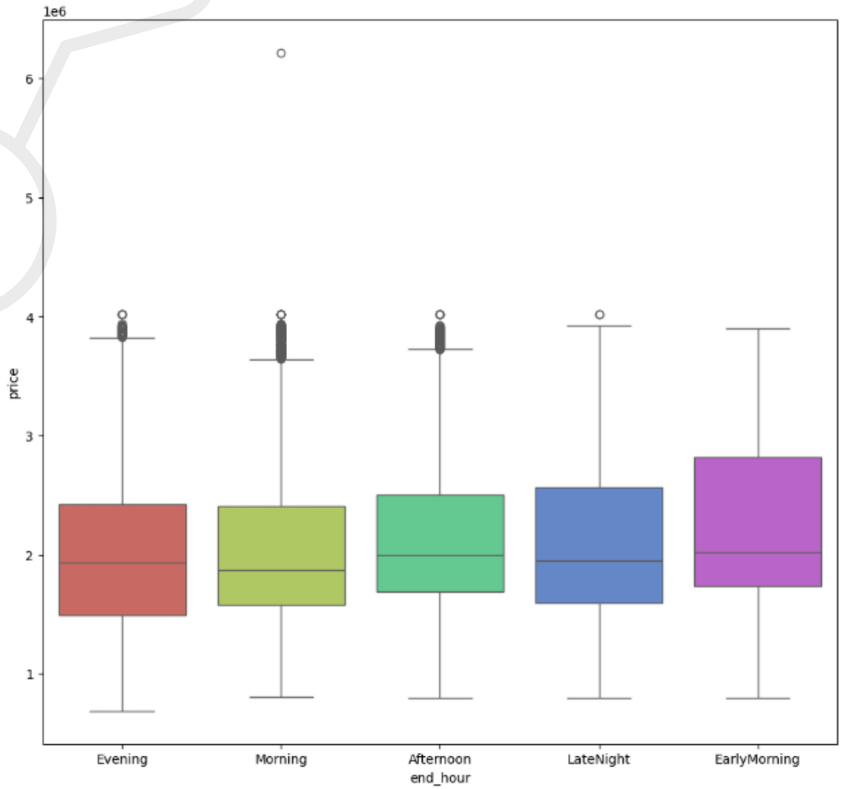
→ Giữa 2 biến **price** và **trip_mins** có sự tương quan đồng biến

Khai phá và phân tích dữ liệu:

Price & start_hour

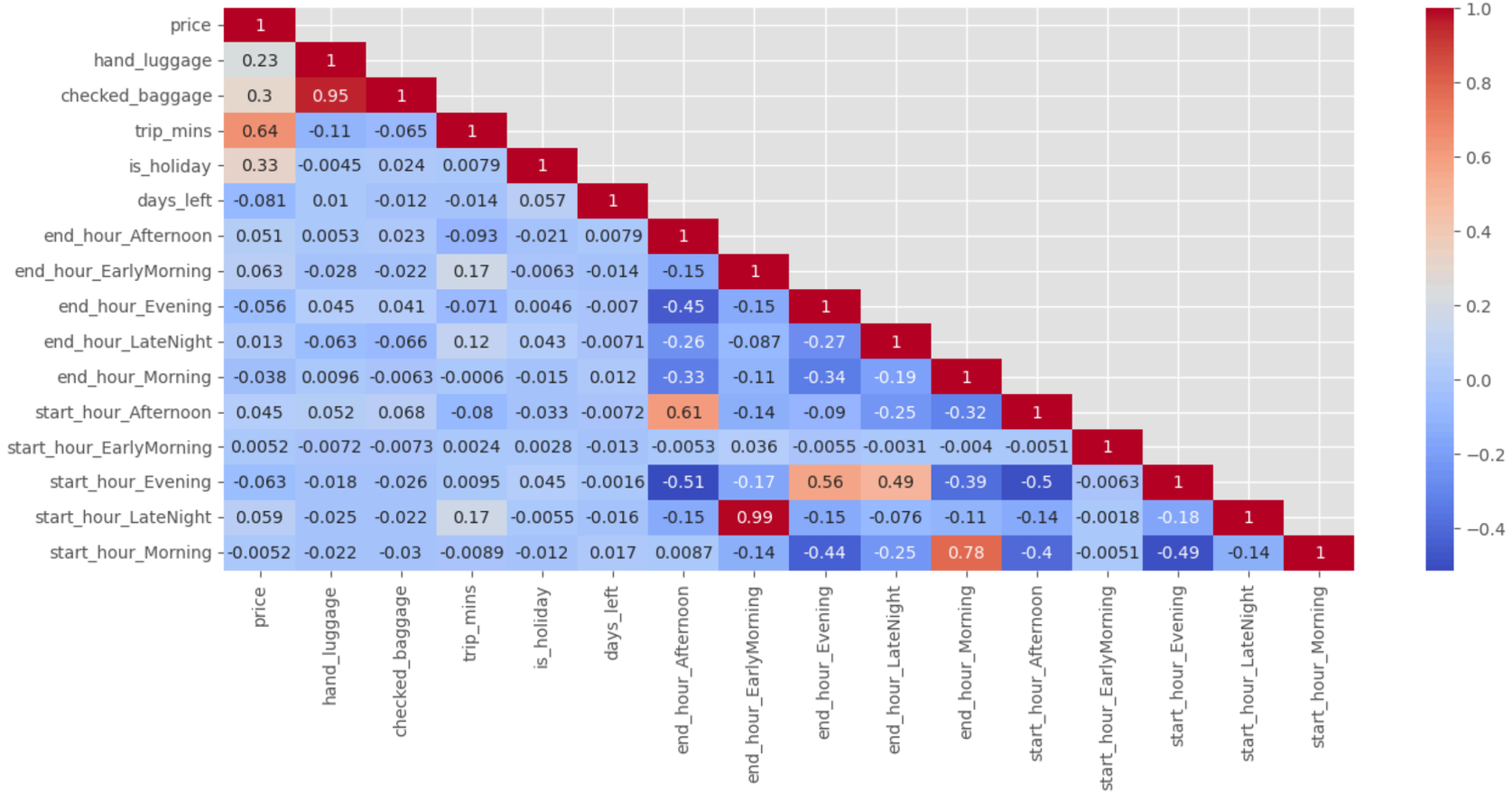


Price & end_hour



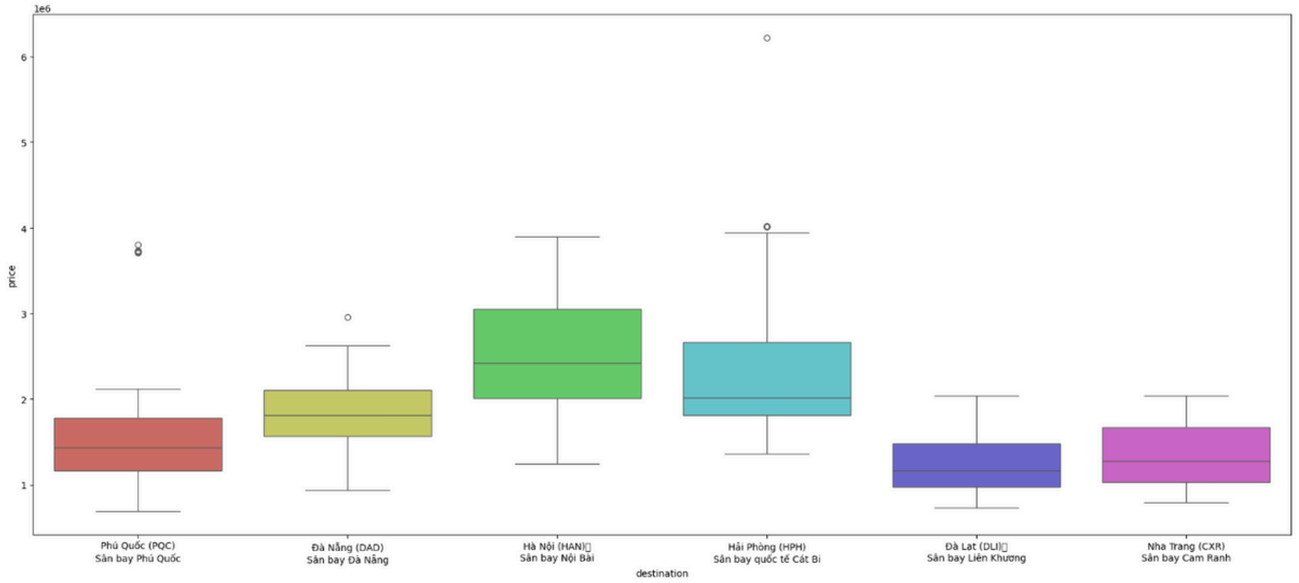
Khai phá và phân tích dữ liệu:

- Độ tương quan giữa các thuộc tính



→ Giữ lại **start_hour** và **end_hour**

Khai phá và phân tích dữ liệu:

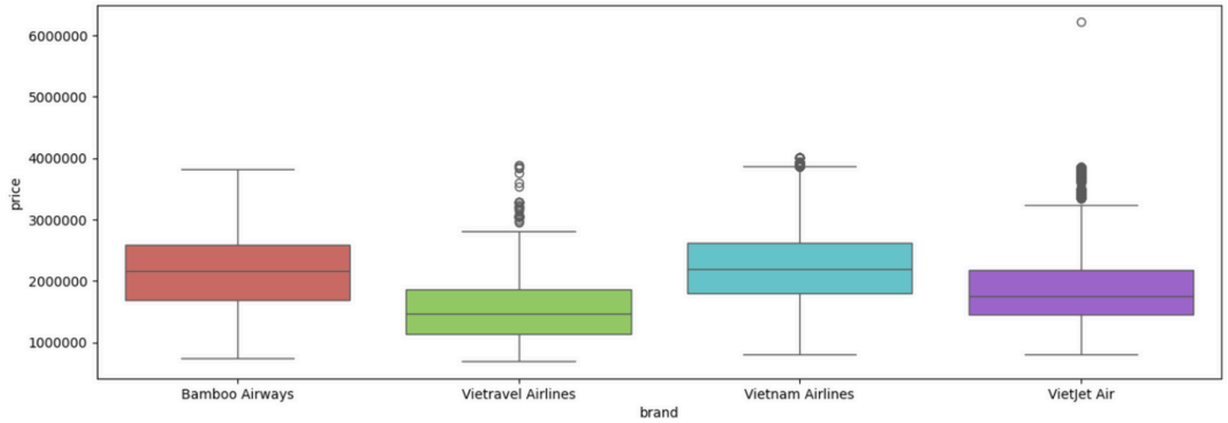


Phân bố giá vé theo điểm đến:

- Có sự khác biệt giữa các sân bay lớn và nhỏ

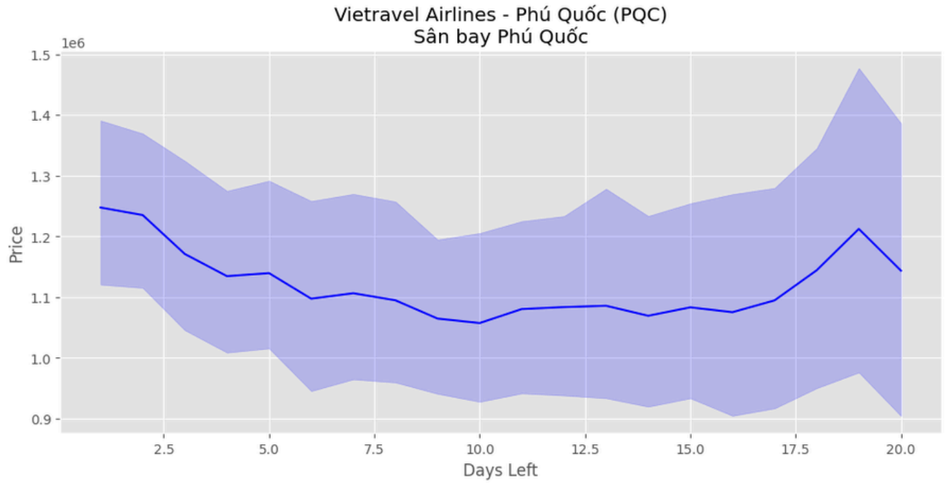
Phân bố giá vé theo hãng bay:

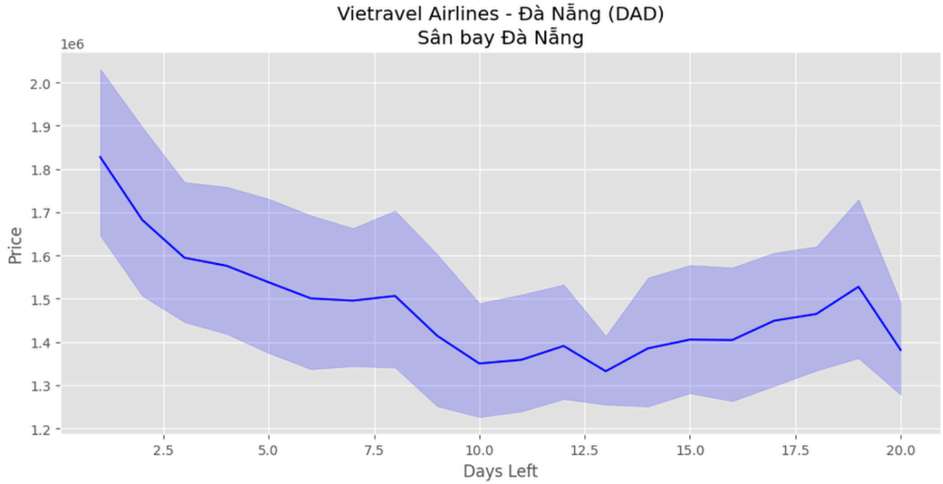
- Phân bố có sự tương đồng

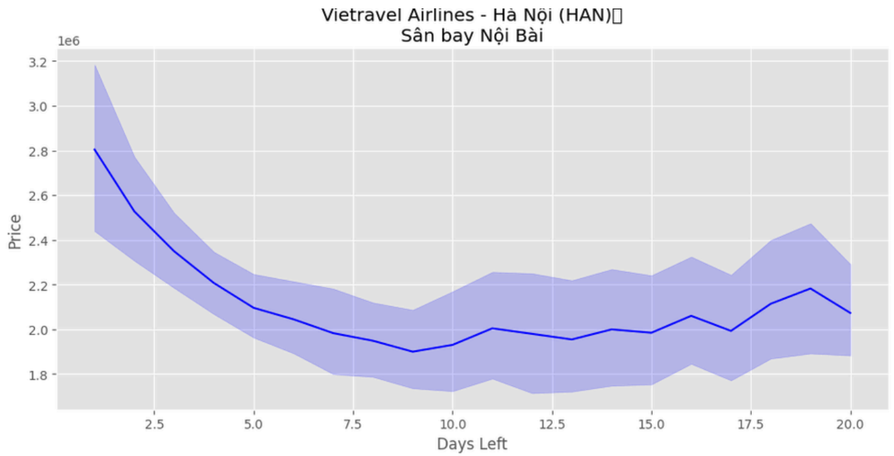


Khai phá và phân tích dữ liệu:

Biểu đồ phân bố price và days_left của brand theo destination

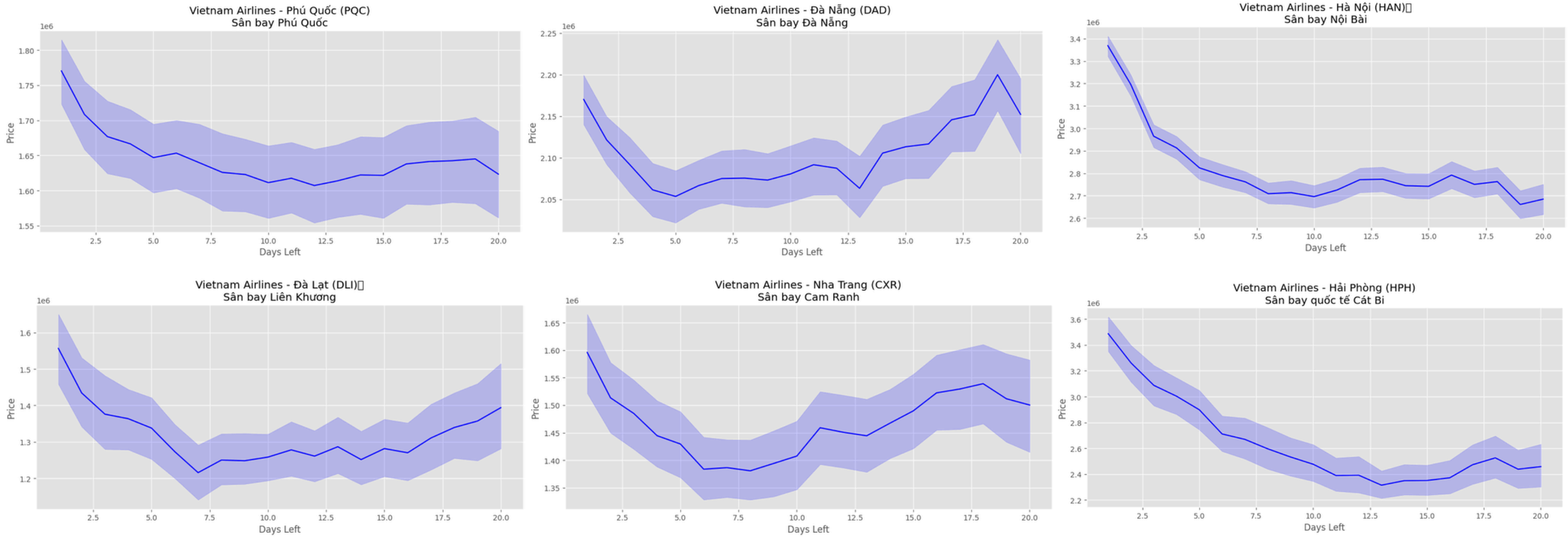






Khai phá và phân tích dữ liệu:

Biểu đồ phân bố price và days_left của brand theo destination



- Theo brand, giá vé tăng, giảm theo quy luật

→ Huấn luyện mô hình theo brand có khả năng cho kết quả tốt hơn theo từng destination.

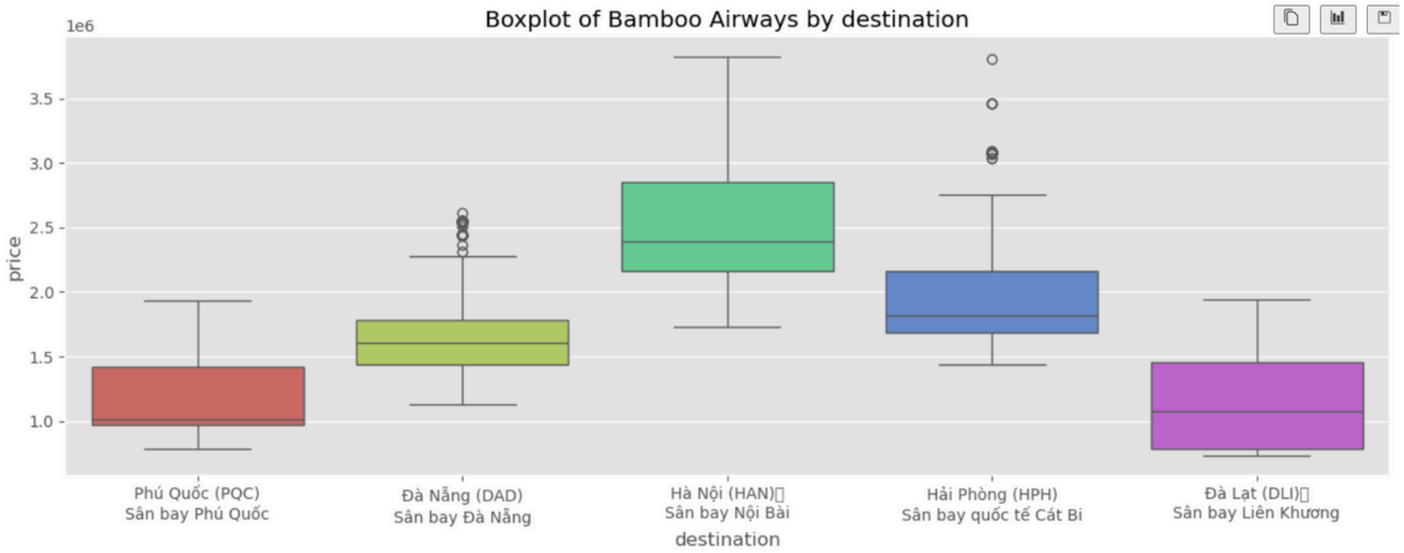
Tách tập train/test

Test_data:  +  +  +  = 40 id

Test= 18%*(bộ dữ liệu - Test_data) + Test_data

Train= bộ dữ liệu - Test

Xử lý outliers:

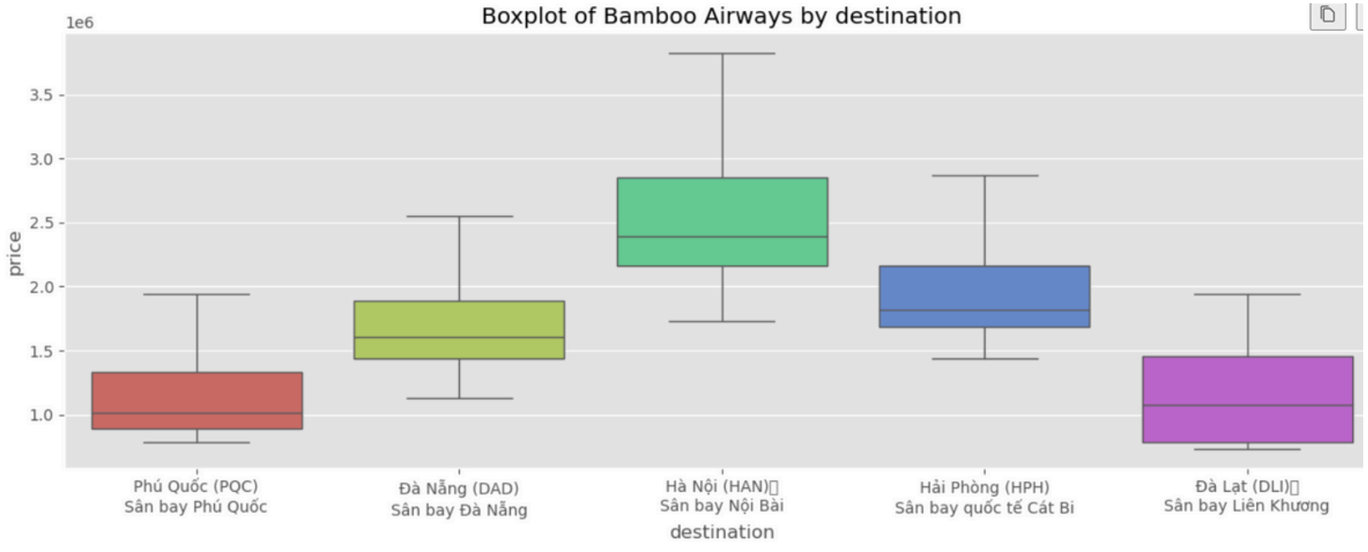



```python
def threshold_outlier(des, data):
    df = data[data['destination'] == des]
    q1 = df['price'].quantile(0.25)
    q3 = df['price'].quantile(0.75)
    iqr = q3 - q1
    lower_bound = q1 - 1.5 * iqr
    upper_bound = q3 + 1.5 * iqr
    return lower_bound, upper_bound
```

[448]

```python
def process_outliers(data):
    for des in data['destination'].unique():
        lower, upper = threshold_outlier(des, data)
        condition = (
                (data['destination'] == des) &
                (data['price'] > upper)
        )
        data.loc[condition, 'price'] = upper
    return data
```

[449]



Xây dựng mô hình ⟶

	id	price	destination	hand_luggage	checked_baggage	trip_mins	is_holiday	days_left
0	PQC0002	1245267.0	Phú Quốc (PQC)\nSân bay Phú Quốc	7	0	60	2	1
1	PQC0003	869953.0	Phú Quốc (PQC)\nSân bay Phú Quốc	7	0	60	2	2
2	PQC0004	1013756.0	Phú Quốc (PQC)\nSân bay Phú Quốc	7	0	60	2	3

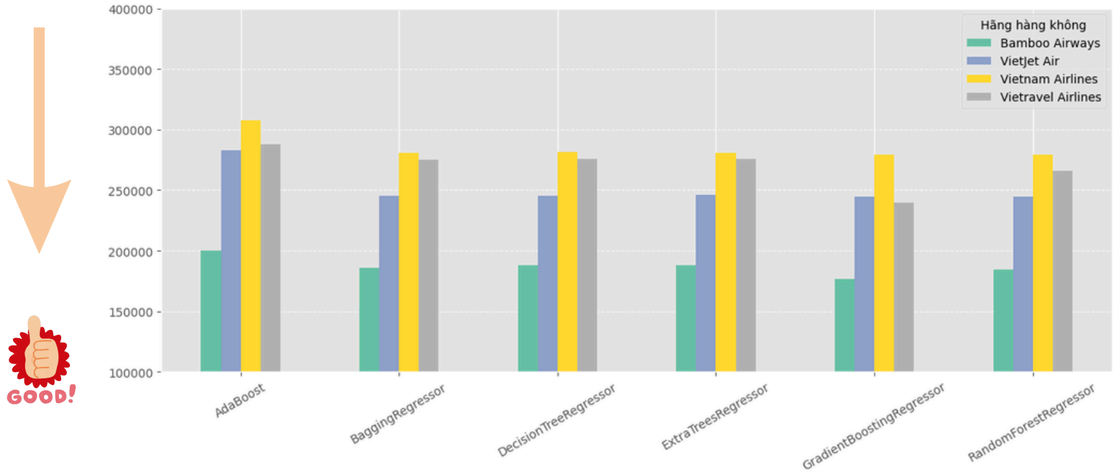
→ Gold layer: 4 tập dữ liệu theo từng brand với 8 thuộc tính

CHUẨN HÓA DỮ LIỆU → **Standard Scaler**

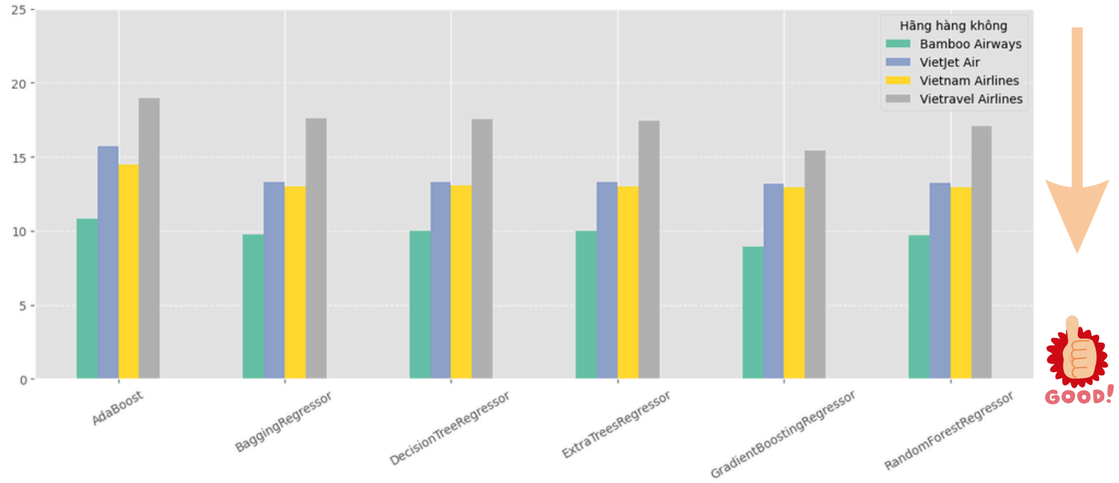
ONE-HOT ENCODING → **destination**

Xây dựng mô hình

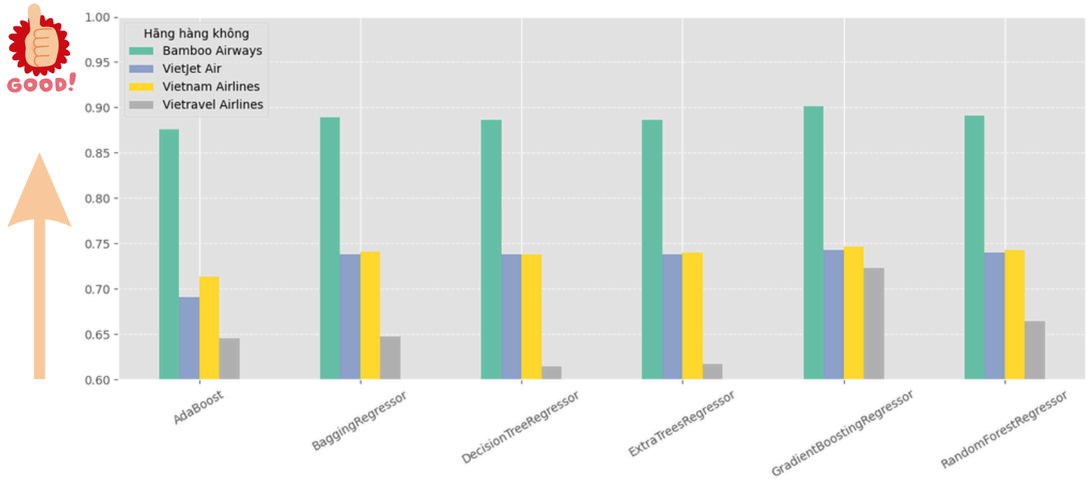
MAE



MAPE

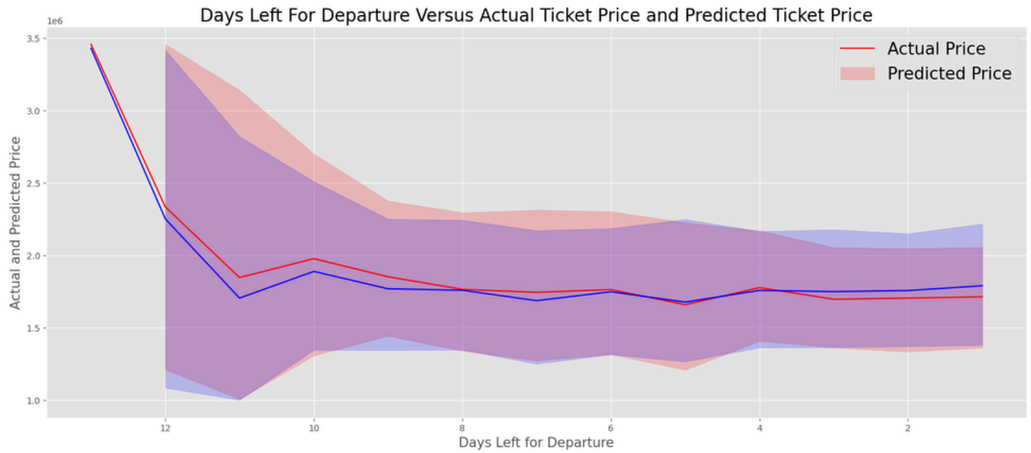


R^2 Score

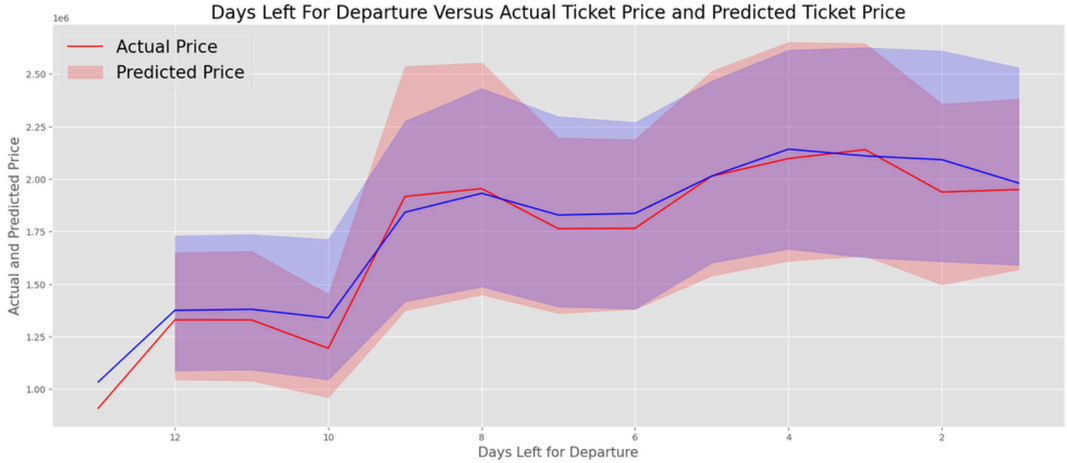


Thực nghiệm & Đánh giá

Bamboo Airways

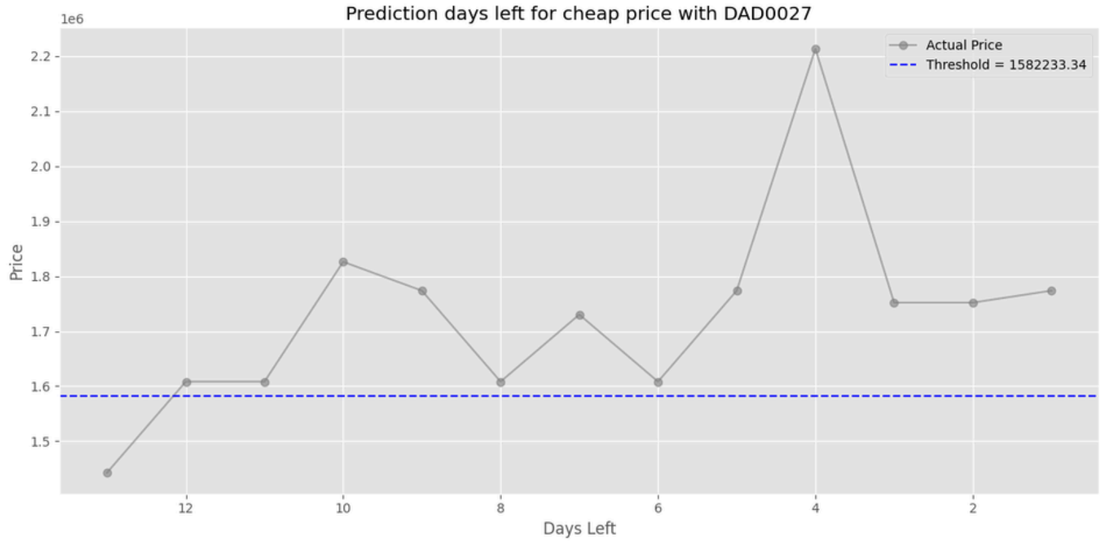


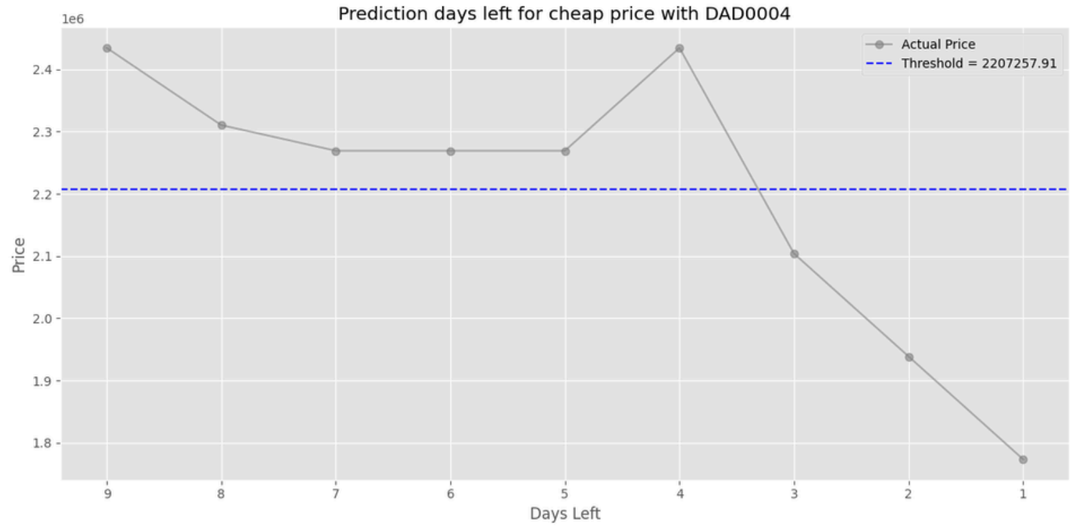
Vietnam Airlines

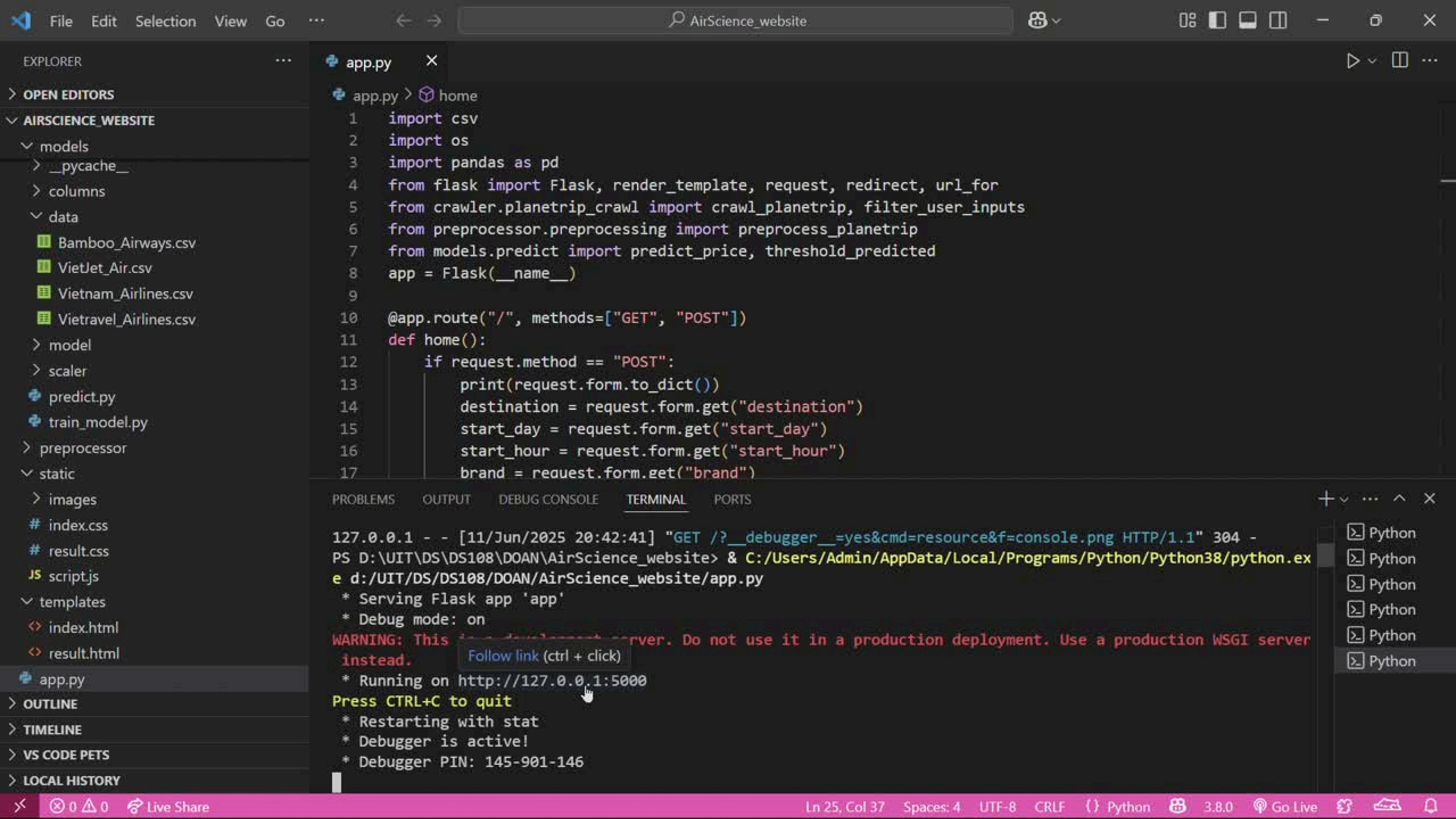


Mô hình
Gradient Boosting Regressor

Thực nghiệm & Đánh giá






File Edit Selection View Go
AirScience_website
EXPLORER
app.py
OPEN EDITORS
AIRSCIENCE_WEBSITE
models
__pycache__
columns
data
Bamboo_Airways.csv
VietJet_Air.csv
Vietnam_Airlines.csv
Vietravel_Airlines.csv
model
scaler
predict.py
train_model.py
preprocessor
static
images
index.css
result.css
script.js
templates
index.html
result.html
app.py
OUTLINE
TIMELINE
VS CODE PETS
LOCAL HISTORY
app.py > home
1 import csv
2 import os
3 import pandas as pd
4 from flask import Flask, render_template, request, redirect, url_for
5 from crawler.planetrip_crawl import crawl_planetrip, filter_user_inputs
6 from preprocessor.preprocessing import preprocess_planetrip
7 from models.predict import predict_price, threshold_predicted
8 app = Flask(__name__)
9
10 @app.route("/", methods=["GET", "POST"])
11 def home():
12 if request.method == "POST":
13 print(request.form.to_dict())
14 destination = request.form.get("destination")
15 start_day = request.form.get("start_day")
16 start_hour = request.form.get("start_hour")
17 brand = request.form.get("brand")
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
127.0.0.1 - - [11/Jun/2025 20:42:41] "GET /?__debugger__=yes&cmd=resource&f=console.png HTTP/1.1" 304 -
PS D:\UIT\DS\DS108\DOAN\AirScience_website> & C:/Users/Admin/AppData/Local/Programs/Python/Python38/python.exe d:/UIT/DS/DS108/DOAN/AirScience_website/app.py
* Serving Flask app 'app'
* Debug mode: on
WARNING: This
Follow link (ctrl + click)
instead.
* Running on http://127.0.0.1:5000
Press CTRL+C to quit
* Restarting with stat
* Debugger is active!
* Debugger PIN: 145-901-146
Python
Python
Python
Python
Python
Python
Ln 25, Col 37 Spaces: 4 UTF-8 CRLF Python 3.8.0 Go Live
Live Share

KẾT LUẬN

- **DEMO**
- **HẠN CHẾ:**
  - Các chuyến bay giới hạn trong 20 ngày trước ngày bay
  - Chuyến bay chỉ xuất phát từ HCM → 6 thành phố khác
  - Không tự động hoàn toàn do reCAPTCHA
- **HƯỚNG PHÁT TRIỂN:**
  - Thời gian đợi từ nhập dữ liệu → dự đoán
  - Học tăng cường