**ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**



BÁO CÁO BÀI TẬP LỚN

NHẬP MÔN KHOA HỌC DỮ LIỆU

**Đề tài: DỰ ĐOÁN GIÁ XE CŨ**

**Giảng viên hướng dẫn:** PGS. TS. Thân Quang Khoát

|  |  |
| --- | --- |
| **Sinh viên thực hiện:** | Phạm Nhật Anh 20214987 |
|  |
|  |
|  |
|  |

Hà Nội, 05/2025

[I. Giới thiệu bài toán 3](#_Toc707246245)

[1.1. Mô tả bài toán 4](#_Toc1718200874)

[1.2. Mục tiêu 4](#_Toc461713743)

[1.3. Yêu cầu 4](#_Toc652441005)

[1.4. Đối tượng sử dụng 4](#_Toc1311491656)

[II. Tiền xử lí dữ liệu 4](#_Toc1111350715)

[2.1. Thu thập dữ liệu 5](#_Toc148331607)

[2.2. Làm sạch dữ liệu 6](#_Toc1881938393)

[2.3. Xử lí nhiễu 7](#_Toc604596136)

[2.4. Chuẩn hóa và mã hóa dữ liệu 8](#_Toc2048094460)

[III. Xây dựng và đánh giá mô hình 8](#_Toc680477525)

[3.1. Hồi quy tuyến tính (Linear Regression): 9](#_Toc1202463896)

[3.2. Rừng cây ngẫu nhiên (Random Forest): 13](#_Toc309491969)

[3.3. XGBoost: 17](#_Toc1097886367)

[IV. Chức năng và cách sử dụng 21](#_Toc1833736106)

[V. Cấu trúc chương trình 24](#_Toc2088352411)

[VI. Khó khăn, thách thức và hướng giải quyết 26](#_Toc1422133897)

[VII. Kết luận 26](#_Toc34430357)

[VIII. Tài liệu tham khảo 26](#_Toc1771928287)

|  |  |  |  |
| --- | --- | --- | --- |
| MSSV | Họ và tên | Nhiệm vụ | Phần trăm  đánh giá |
| 20214987 | Phạm Nhật Anh | Triển khai model Linear Regression, thiết kế CSDL | 20% |
|  |  | Thiết kế Frontend, Backend dự đoán giá xe |  |
|  |  | Xây dựng hệ thống crawl dữ liệu, triển khai kết nối CSDL |  |
|  |  | Xử lý dữ liệu, triển khai model XGBoost |  |
|  |  | Triển khai model Random Forest, viết báo cáo |  |

# Giới thiệu bài toán

* 1. **Mô tả bài toán**

Xe hơi là một tài sản có giá trị lớn và đóng vai trò quan trọng trong đời sống hiện đại, không chỉ là phương tiện di chuyển mà còn gắn với nhu cầu cá nhân và công việc. Trong thị trường xe cũ, việc định giá một chiếc xe là điều không hề dễ dàng do bị ảnh hưởng bởi nhiều yếu tố như năm sản xuất, số km đã đi, hãng xe, tình trạng sử dụng, v.v. Điều này gây khó khăn cho cả người bán lẫn người mua trong việc xác định mức giá hợp lý. Do đó, bài toán đặt ra là xây dựng một mô hình dự đoán giá xe cũ dựa trên dữ liệu lịch sử và các đặc điểm kỹ thuật của xe, nhằm hỗ trợ quá trình định giá trở nên chính xác và minh bạch hơn.

* 1. **Mục tiêu**

Mục tiêu của bài toán là phát triển một hệ thống dự đoán giá xe ô tô cũ một cách chính xác bằng cách áp dụng các thuật toán Machine Learning. Hệ thống sẽ sử dụng các đặc điểm quan trọng của xe như năm sản xuất, hãng xe, dòng xe, quãng đường đã đi, loại nhiên liệu, số chỗ ngồi và các thông tin liên quan khác để ước lượng giá bán phù hợp theo thời gian thực, giúp người dùng đưa ra quyết định mua bán dễ dàng và hiệu quả hơn.

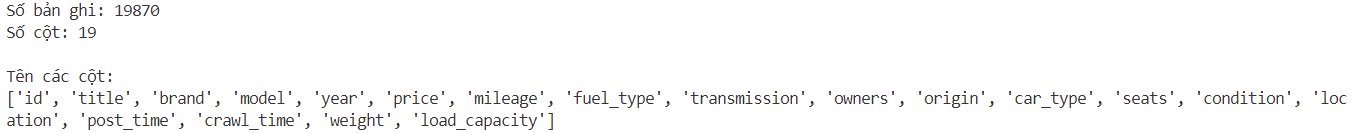
* 1. **Yêu cầu**

Ứng dụng dự đoán giá xe cũ một cách chính xác nhất dựa trên dữ liệu lịch sử, giúp người dùng dễ dàng lựa chọn những chiếc xe phù hợp với nhu cầu và mức giá mong muốn của họ.

* 1. **Đối tượng sử dụng**
* Người mua, người bán xe
* Đại lý xe cũ

# Tiền xử lí dữ liệu

* 1. **Thu thập dữ liệu**
* Dữ liệu được crawl từ trang https://xe.chotot.com/mua-ban-oto.
* Dữ liệu sau khi thu thập có 19870 dòng.



* Được lưu vào file: raw.csv
* Gồm 19 trường như sau:

+ id: đánh dấu các dữ liệu đã thu thập được

+ title: tên bài đăng

+ brand: hãng xe

+ model: dòng xe

+ year: năm sản xuất

+ price: giá bán

+ mileage: quãng đường xe đã đi

+ fuel\_type: loại nhiên liệu

+ transmission: loại hộp số

+ owners: số lượng chủ cũ

+ origin: nguồn gốc

+ car\_type: loại xe

+ seats: số lượng ghế ngồi

+ condition: tình trạng xe

+ location: địa điểm

+ post\_time: thời gian đăng

+ crawl\_time: thời gian crawl

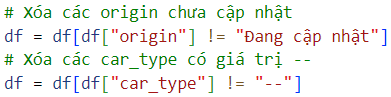
+ weight: trọng lượng

+ load\_capacity: sức chứa

* 1. **Làm sạch dữ liệu**
* Đầu tiên, loại bỏ những bộ dữ liệu hoặc những thuộc tính không có giá trị:
* Loại bỏ các cột id, title, post\_time, crawl\_time do không có ảnh hưởng đến giá xe.
* Xoá các bản ghi mà year < 2000 do có ít bản ghi.



* Xóa các bản ghi không rõ ràng (origin = “chưa cập nhật” hoặc car\_type = "--").



* Xoá cột owners do chỉ có 1 giá trị là 1 và nhiều bản ghi bị null.



* Loại bỏ các bản ghi có ít nhất 1 giá trị null.

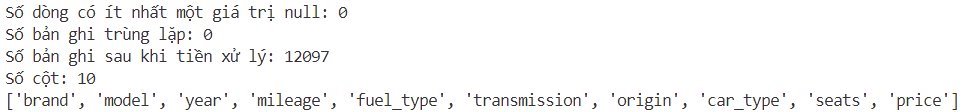


* Xóa những bản ghi trùng lặp.



* Loại bỏ các giá trị trong các cột phân loại nếu có ít hơn 10 bản ghi.
* Sau khi xử lý thì cột condition chỉ còn 1 giá trị nên cũng sẽ loại bỏ.





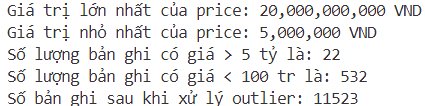
* 1. **Xử lí nhiễu**

Qua tìm hiểu, hầu hết ô tô đã qua sử dụng trên thị trường đều có mức giá khoảng vài trăm

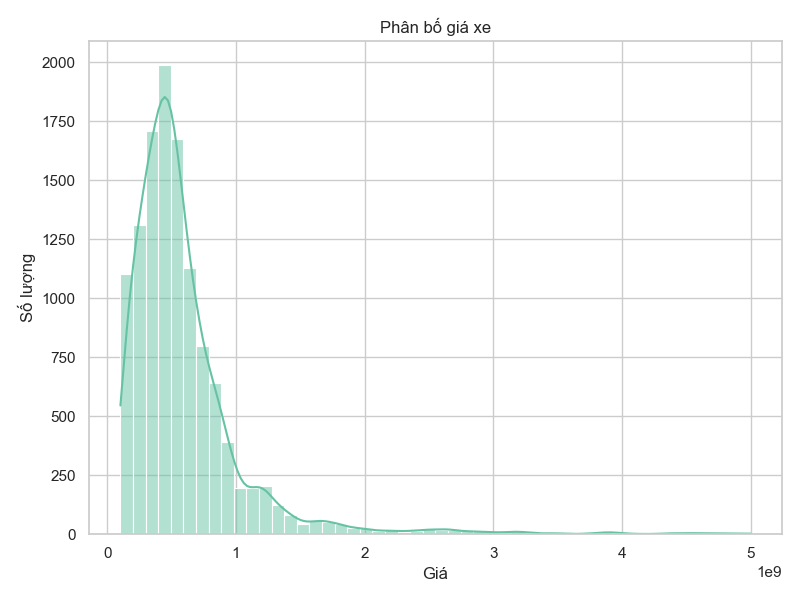
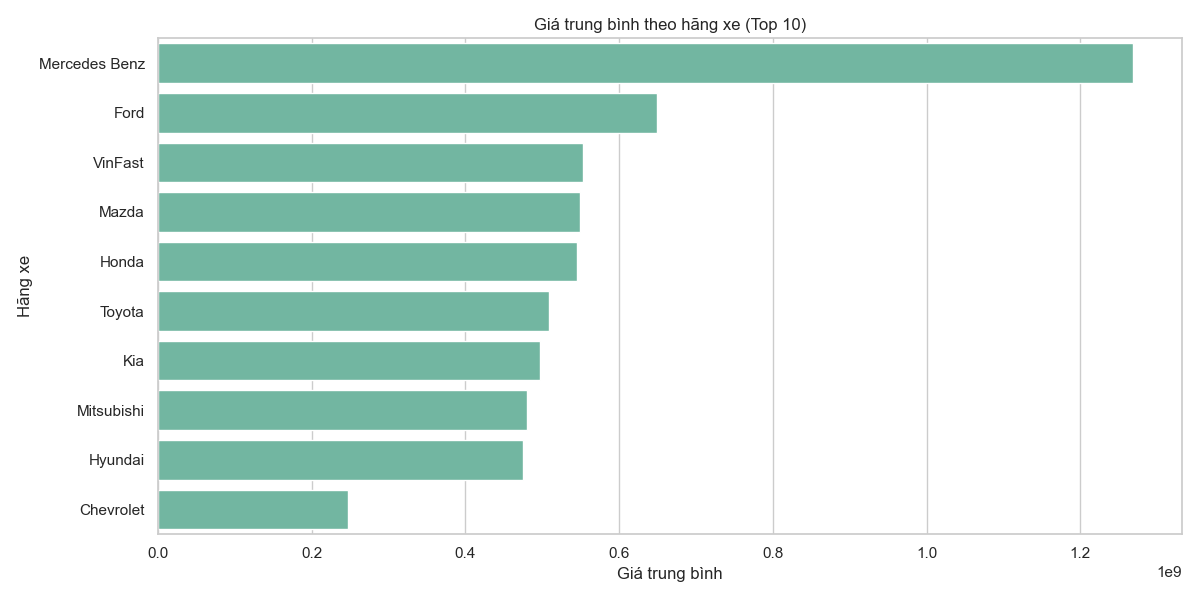
triệu đến vài tỷ đồng. Do đó loại bỏ các bản ghi có giá xe quá cao (> 5 tỷ đồng) hoặc quá thấp (< 100 triệu đồng).

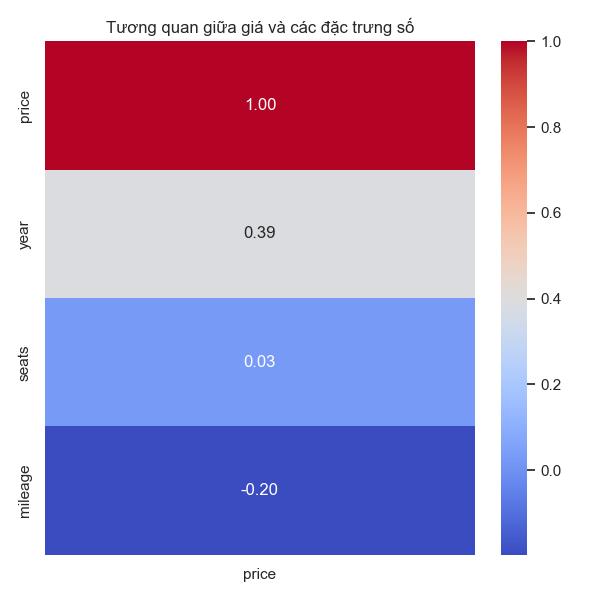
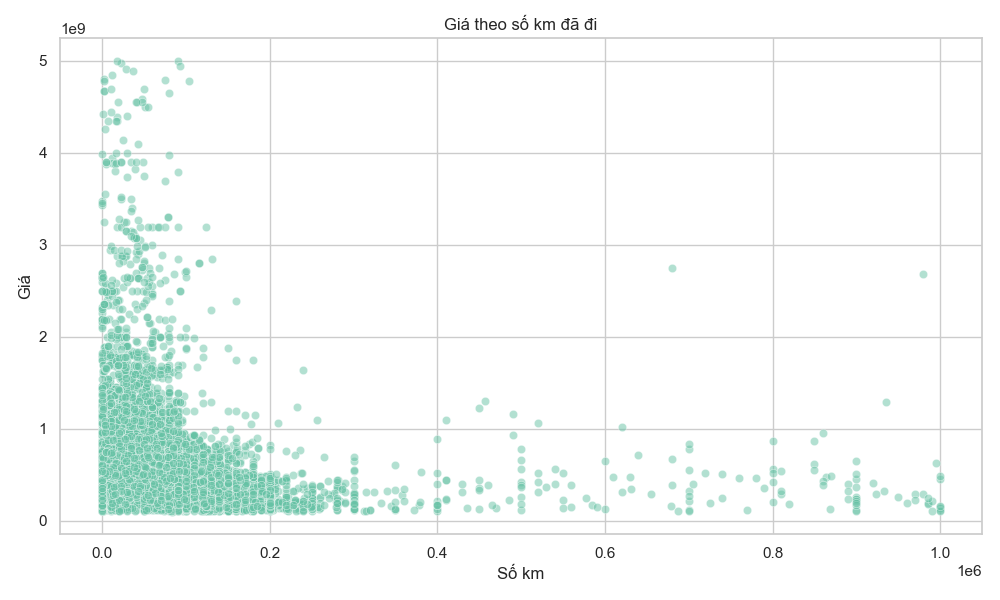
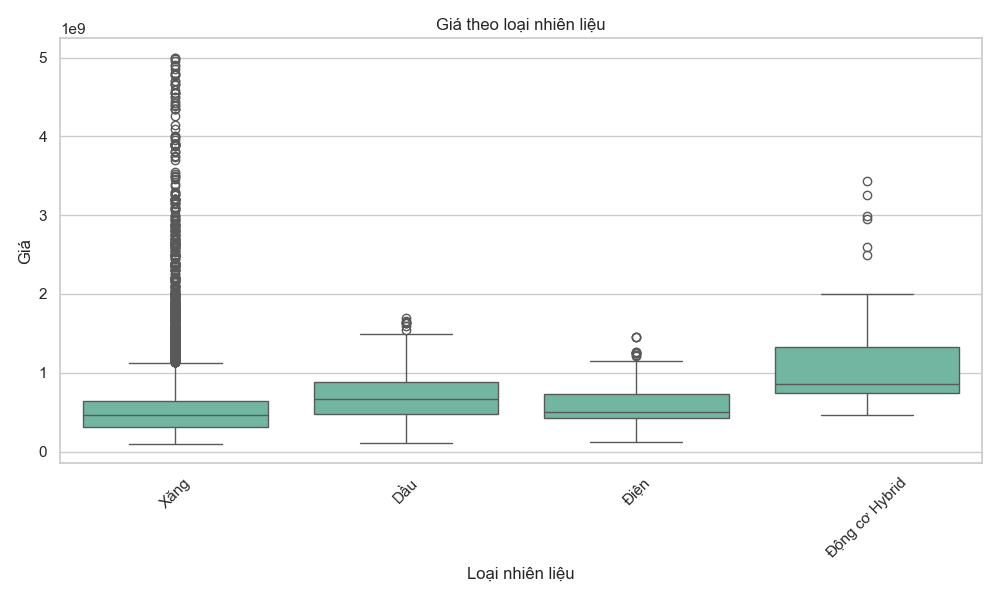
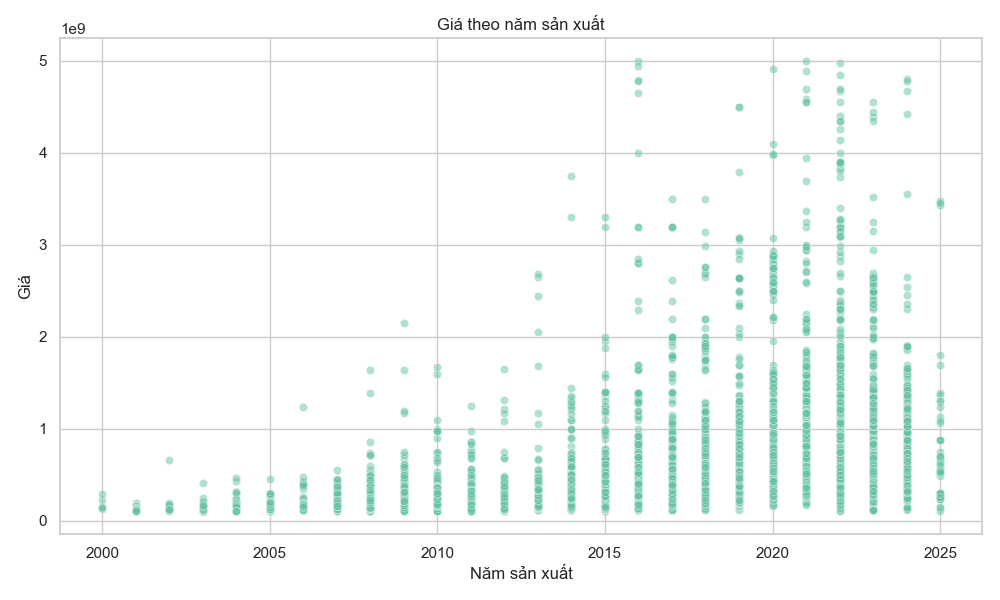






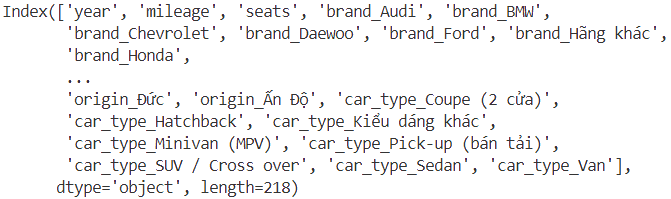
* 1. **Chuẩn hóa và mã hóa dữ liệu**
* **Chuẩn hóa:** Thực hiện bằng phương pháp **Z-score (Standardization)** nhằm đưa các đặc trưng về cùng một thang đo với giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1. Quá trình này được thực hiện thông qua công cụ StandardScaler trong thư viện scikit-learn, giúp mô hình học máy hoạt động hiệu quả hơn.
* **Mã hóa:** sử dụng One-Hot Encoding để chuyển các giá trị phân loại (categorical variables) thành các đặc trưng số học (numeric features), giúp mô hình học máy có thể hiểu và xử lý các thông tin phân loại một cách chính xác hơn, đồng thời tránh việc tạo ra mối quan hệ giả giữa các nhãn phân loại mà phương pháp gán số thứ tự (label encoding) có thể gây ra.
  1. **Trực quan hoá dữ liệu**

 Phần lớn các xe có mức giá thấp hơn trung bình, tập trung nhiều ở phía bên trái trục giá. Phân bố này bị lệch phải, biểu thị sự hiện diện của một số ít xe có giá rất cao (outliers), thường là xe sang hoặc xe mới, kéo dài phần đuôi phải của biểu đồ. Đây là đặc trưng phổ biến trong dữ liệu giá xe đã qua sử dụng.

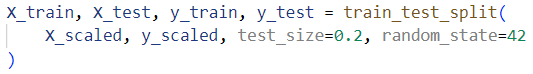
Biểu đồ thanh cho thấy sự chênh lệch rõ rệt về giá trung bình giữa các hãng xe phổ biến nhất. Những hãng như Mercedes-Benz và BMW có giá trung bình cao hơn đáng kể so với các hãng như Kia hoặc Hyundai. Điều này phản ánh sự khác biệt về phân khúc thị trường, trong đó các hãng xe sang thường có giá cao hơn nhiều so với các hãng xe phổ thông.

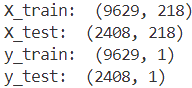
# Xây dựng và đánh giá mô hình

* **Đầu vào***:* Sau khi chuẩn hóa và mã hóa dữ liệu đầu vào gồm 218 trường



* ***Kích thước dữ liệu:*** 80% là dữ liệu train và 20% là dữ liệu test

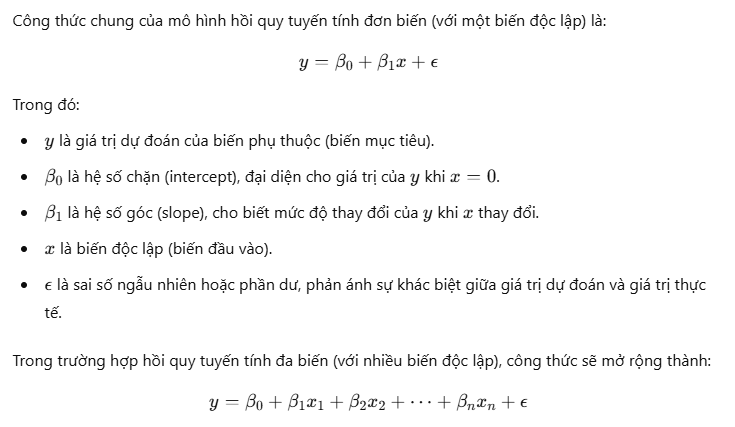




* 1. **Hồi quy tuyến tính (Linear Regression)**:

- ***Lý do chọn mô hình:*** Mô hình đơn giản và dễ hiểu, phù hợp khi mối quan hệ giữa các yếu tố và giá xe có tính chất tuyến tính. Mô hình này dễ dàng triển khai, tính toán nhanh và yêu cầu ít tài nguyên, đặc biệt khi dữ liệu không quá phức tạp. Hồi quy tuyến tính cũng cung cấp khả năng giải thích rõ ràng về sự ảnh hưởng của từng yếu tố lên giá xe, giúp dễ dàng kiểm tra và tối ưu mô hình.

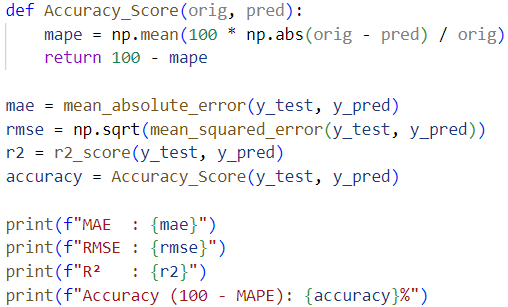
***- Định nghĩa:*** Mô hình hồi quy tuyến tính là một phương pháp phân tích dữ liệu nhằm xác định mối quan hệ tuyến tính giữa một hoặc nhiều biến độc lập (biến đầu vào) và một biến phụ thuộc (biến mục tiêu). Mục tiêu của mô hình là tìm ra một phương trình đường thẳng (hoặc siêu phẳng trong trường hợp có nhiều biến độc lập) sao cho sự chênh lệch giữa giá trị dự đoán và giá trị thực tế là nhỏ nhất.

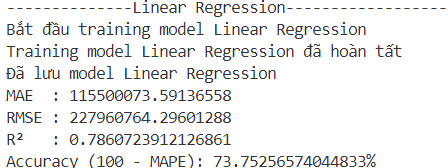


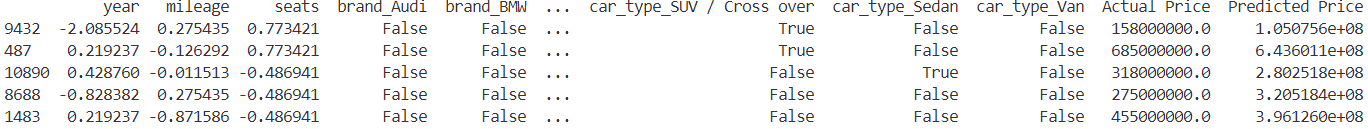
* ***Cài đặt:***



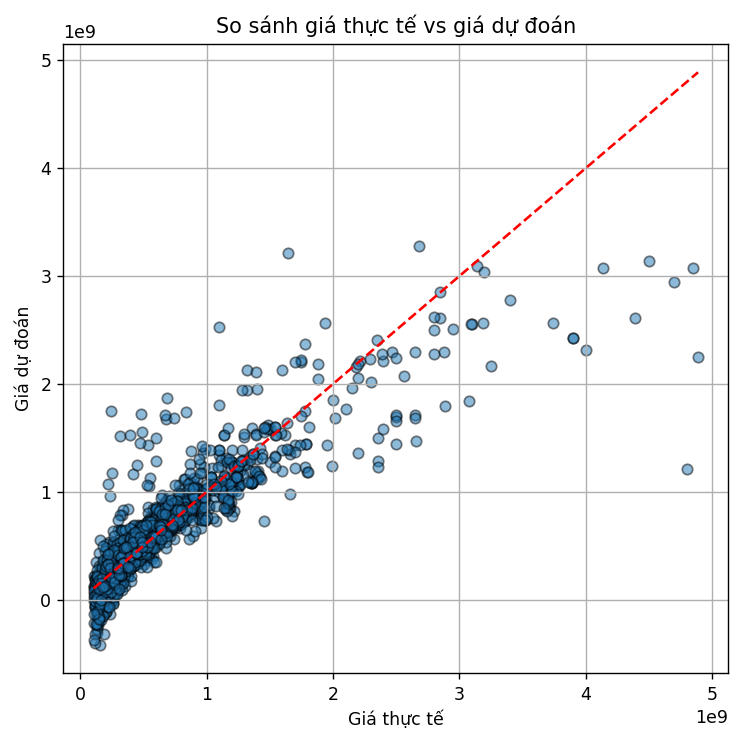
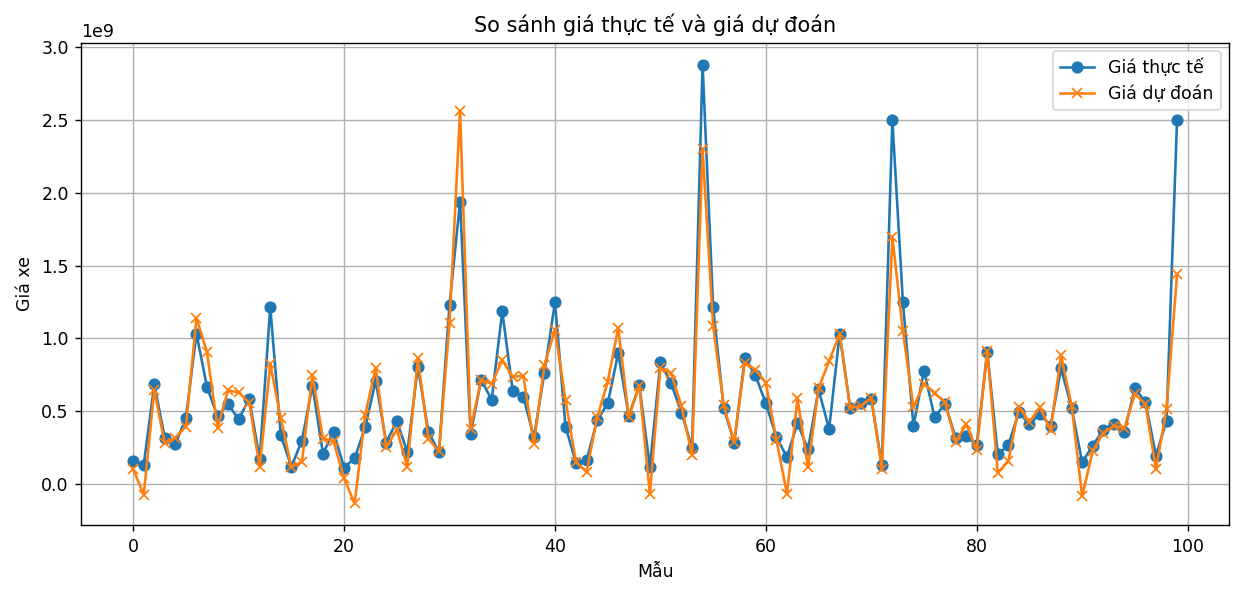
* ***Kết quả và đánh giá:***







* ***Nhận xét:***
* **MAE (Mean Absolute Error)** cho biết sai số tuyệt đối trung bình giữa giá trị thực tế và giá trị dự đoán. Trong trường hợp này, MAE ≈ 115.5 triệu đồng nghĩa là, trung bình, mỗi lần dự đoán giá xe, mô hình sai lệch khoảng 115 triệu đồng so với giá thực tế. Đây là mức sai số có thể chấp nhận được trong bối cảnh dữ liệu giá trị lớn (như giá ô tô).
* **RMSE (Root Mean Squared Error)** là căn bậc hai của MSE. Với RMSE ≈ 227.96 triệu đồng, mô hình có một số trường hợp dự đoán sai lệch khá lớn. RMSE cao hơn MAE cho thấy tồn tại những điểm dữ liệu có sai số cao vượt trội (outliers).
* **R² (Hệ số xác định)** đo lường mức độ phù hợp của mô hình với dữ liệu. Với R² ≈ 0.786, tức là khoảng 78.6% sự biến thiên của giá xe có thể được giải thích bởi các đặc trưng đầu vào. Đây là một kết quả tương đối tốt đối với bài toán hồi quy với dữ liệu thực tế có nhiễu.
* **Độ chính xác tương đối (100 - MAPE)** là khoảng 73.75%, nghĩa là trung bình mô hình dự đoán giá xe đúng khoảng 74% so với giá trị thực tế, một mức độ chấp nhận được.
* ***Biểu đồ so sánh:***



* 1. **Rừng cây ngẫu nhiên (Random Forest)**:

- ***Lý do chọn mô hình:*** Random Forest là một mô hình học máy mạnh mẽ và ổn định, đặc biệt hiệu quả khi làm việc với dữ liệu có nhiều nhiễu hoặc đặc trưng phức tạp. Với khả năng xử lý cả đặc trưng dạng số và dạng phân loại, Random Forest dễ sử dụng, ít cần điều chỉnh tham số và cho phép đánh giá tầm quan trọng của từng đặc trưng một cách trực quan và dễ hiểu.

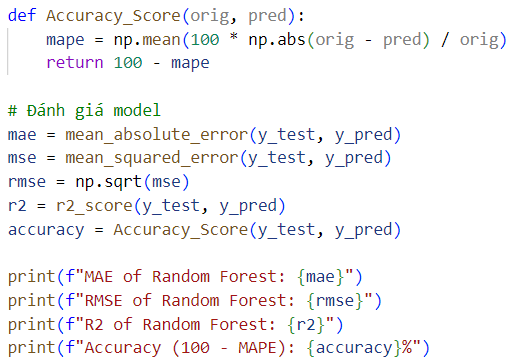
***- Định nghĩa:*** Random Forest là một thuật toán học máy thuộc nhóm mô hình học dựa trên cây quyết định (decision tree), sử dụng kỹ thuật **bagging** (Bootstrap Aggregating) để cải thiện độ chính xác và giảm thiểu overfitting. Random Forest xây dựng một tập hợp các cây quyết định, mỗi cây được huấn luyện trên một mẫu ngẫu nhiên của dữ liệu huấn luyện, và kết quả dự đoán của mô hình được tính bằng cách lấy trung bình hoặc bỏ phiếu (với các bài toán phân loại) từ tất cả các cây trong rừng.

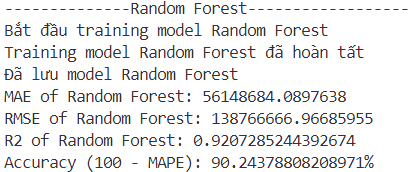
Các thành phần chính của Random Forest:

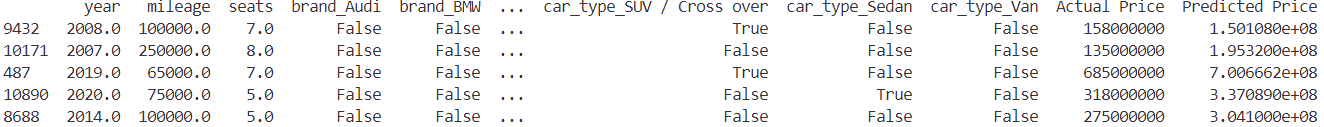
* **Bootstrap Sampling**: Random Forest sử dụng phương pháp **sampling** (lấy mẫu) ngẫu nhiên với thay thế từ dữ liệu huấn luyện, nghĩa là mỗi cây trong rừng được huấn luyện trên một tập con khác nhau của dữ liệu gốc. Điều này giúp tạo ra sự đa dạng trong các cây quyết định và làm giảm độ phương sai của mô hình.
* **Các cây quyết định (Decision Trees)**: Random Forest xây dựng nhiều cây quyết định, mỗi cây chỉ sử dụng một tập con ngẫu nhiên của các đặc trưng (features) tại mỗi bước phân chia, điều này giúp giảm sự phụ thuộc vào các đặc trưng có sự tương quan cao và cải thiện tính tổng quát của mô hình.
* **Kết hợp kết quả (Ensemble Learning)**: Đối với bài toán phân loại, kết quả của Random Forest được đưa ra bằng cách lấy kết quả bỏ phiếu từ tất cả các cây trong rừng. Đối với bài toán hồi quy, kết quả dự đoán được tính là trung bình cộng các dự đoán từ các cây.
* ***Cài đặt:***



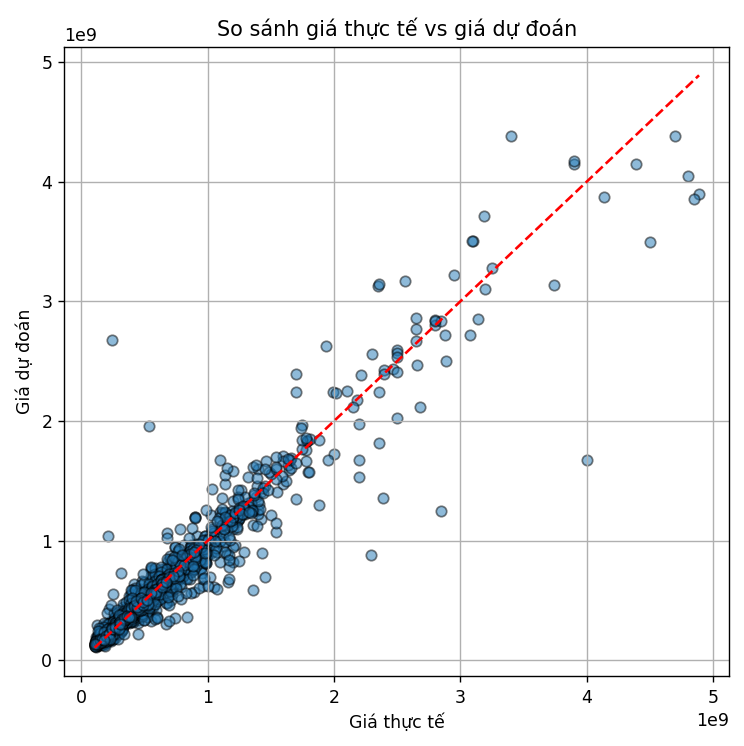
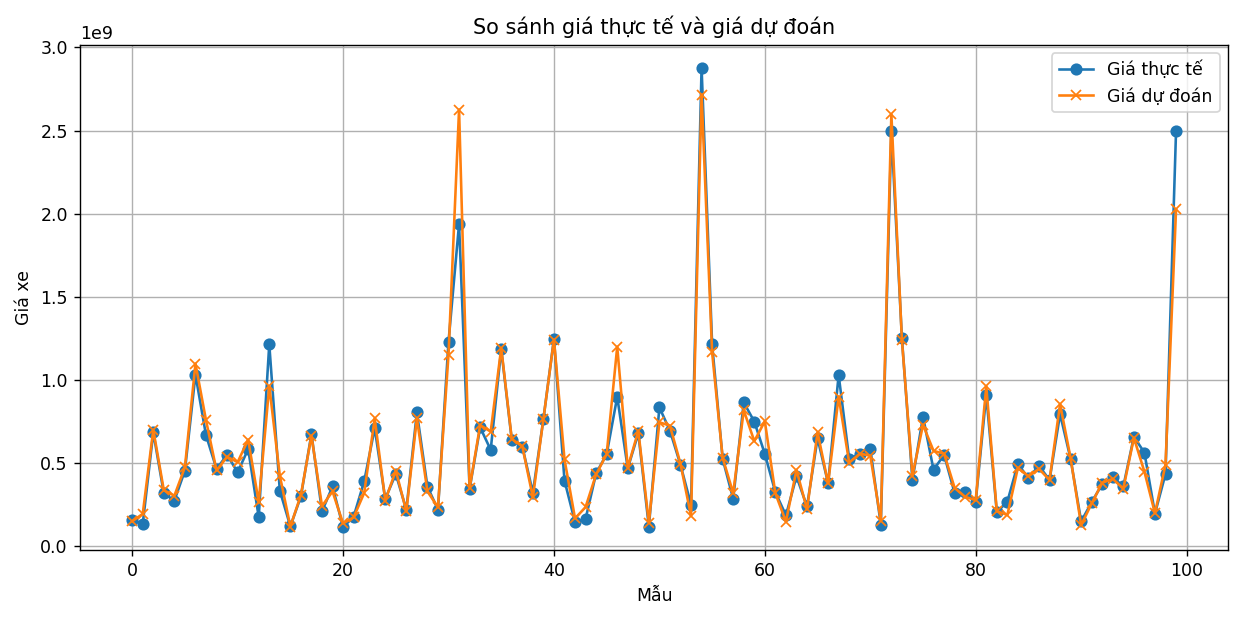
* ***Kết quả và đánh giá:***



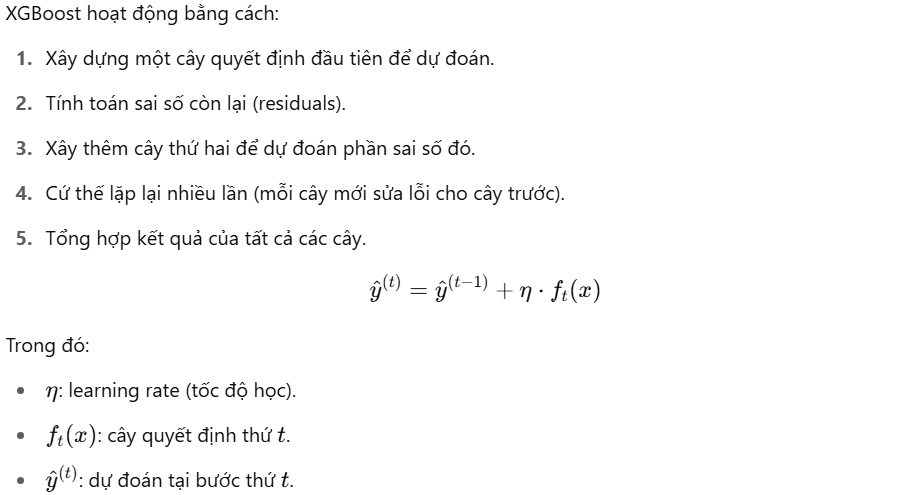




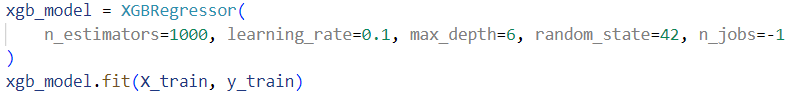
* ***Nhận xét:***
  + **MAE (Mean Absolute Error)** đo lường sai số tuyệt đối trung bình giữa giá trị dự đoán và giá trị thực tế. Với MAE ≈ 56.15 triệu đồng, điều này cho thấy, trung bình, mỗi lần mô hình Random Forest dự đoán giá xe, nó sai lệch khoảng 56 triệu đồng so với giá thực tế. Đây là một mức sai số khá thấp, thể hiện hiệu suất tốt của mô hình.
  + **RMSE (Root Mean Squared Error)** là căn bậc hai của sai số bình phương trung bình. Với RMSE ≈ 138.77 triệu đồng, giá trị này cao hơn MAE cho thấy vẫn có một số điểm dữ liệu gây sai số lớn (outliers), nhưng nhìn chung vẫn trong mức kiểm soát.
  + **R² (Hệ số xác định)** cho biết mức độ mô hình giải thích được sự biến thiên trong dữ liệu. Giá trị R² ≈ 0.921 tức là 92.1% sự thay đổi trong giá xe có thể được giải thích bằng các đặc trưng đầu vào. Đây là một kết quả rất tốt, cho thấy mô hình Random Forest có khả năng học được mối quan hệ giữa các biến.
  + **Độ chính xác tương đối (100 - MAPE)** đạt khoảng 90.24%, thể hiện mô hình có độ chính xác cao trong việc dự đoán giá xe, đặc biệt là so với các mô hình tuyến tính đơn giản như Linear Regression.
* ***Biểu đồ so sánh:***



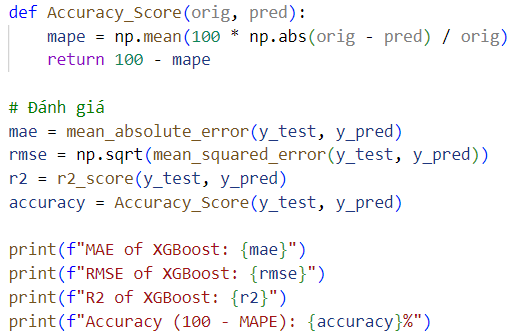
* 1. **XGBoost:**
* ***Lý do chọn mô hình:***Mô hình XGBoost mang lại hiệu suất dự đoán cao nhờ vào cơ chế học tăng cường (boosting) phù hợp với các bài toán phức tạp có nhiều đặc trưng và mối quan hệ phi tuyến giữa dữ liệu đầu vào và đầu ra.
* ***Định nghĩa*:** Extreme Gradient Boosting (XGBoost) làmột thuật toán học máy có giám sát (supervised learning) rất mạnh, dựa trên kỹ thuật boosting, được thiết kế để đạt hiệu suất cao và tốc độ tối ưu. Đây là một ensemble model xây dựng từ nhiều cây quyết định (decision trees) và được cải tiến từng bước để giảm lỗi dự đoán.

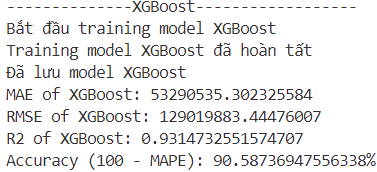


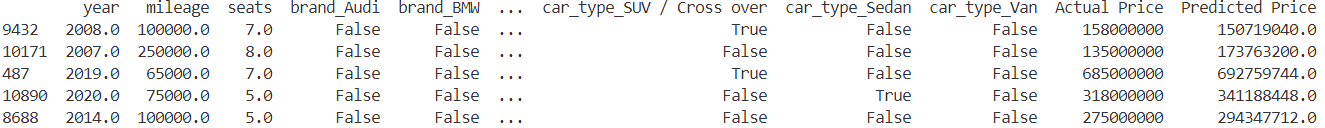
* ***Cài đặt:***



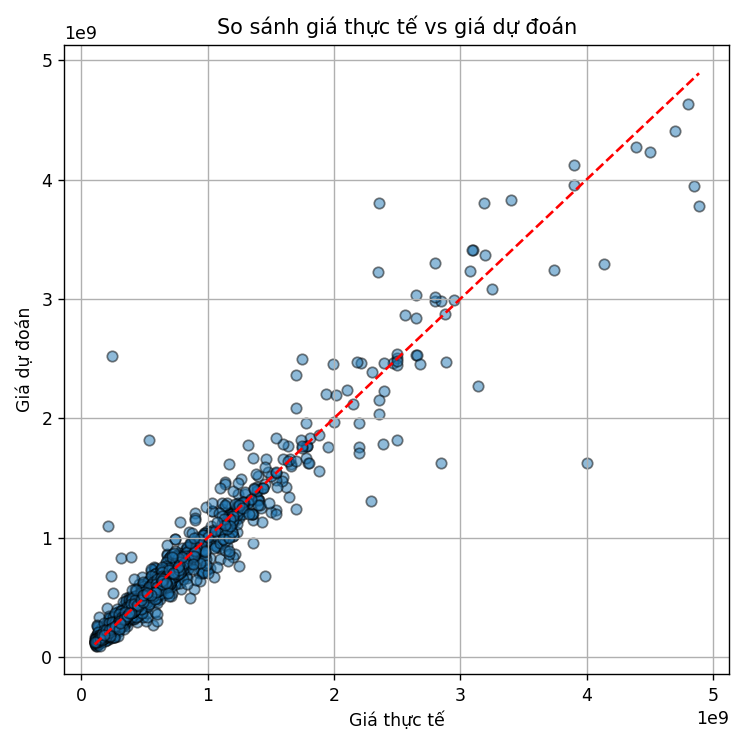
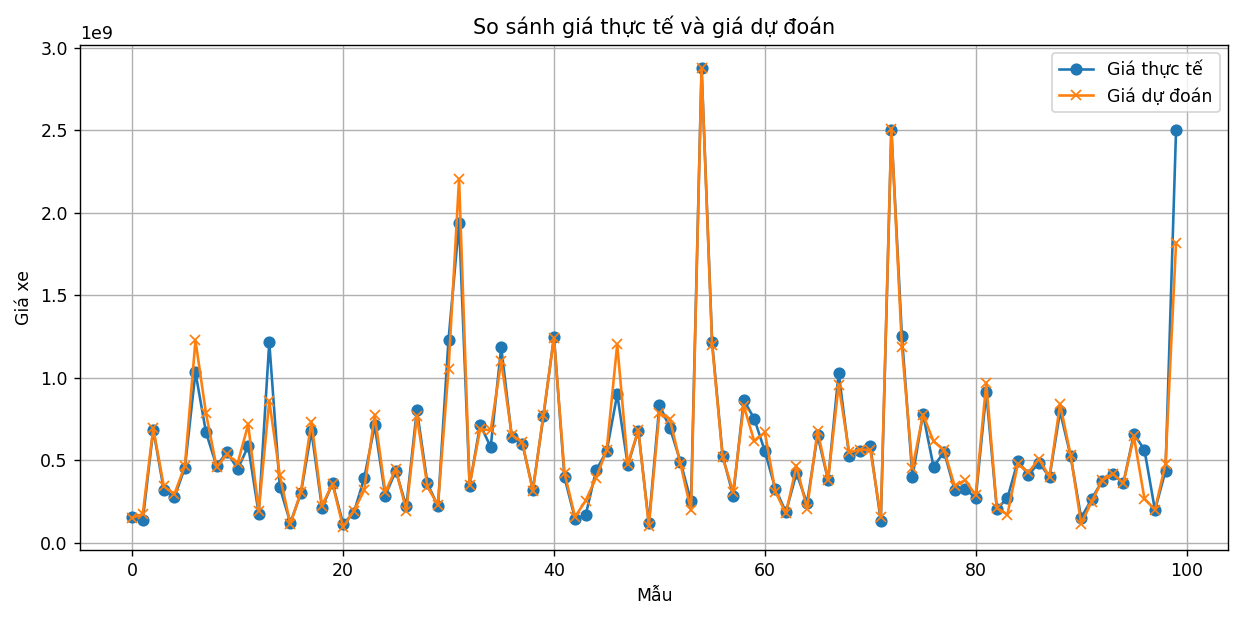
* ***Kết quả và đánh giá:***





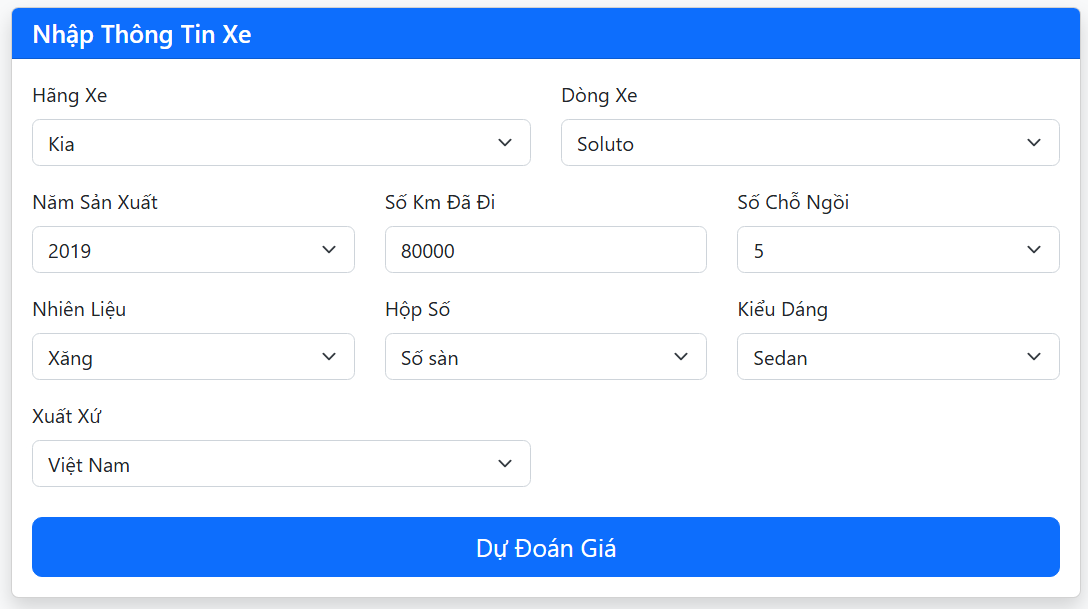


* ***Nhận xét:***
  + **MAE (Mean Absolute Error)** cho thấy sai số tuyệt đối trung bình giữa giá trị thực tế và dự đoán. Với MAE ≈ 53.29 triệu đồng, điều này có nghĩa là trung bình mỗi lần dự đoán, mô hình XGBoost sai lệch khoảng 53 triệu đồng so với giá xe thực tế. Đây là một mức sai số thấp, cho thấy mô hình có hiệu quả tốt.
  + **RMSE (Root Mean Squared Error)** bằng ≈ 129.02 triệu đồng, phản ánh mức độ sai lệch lớn hơn khi có những điểm ngoại lai. Tuy nhiên, so với MAE và các mô hình khác, RMSE của XGBoost vẫn nằm trong ngưỡng tốt và thấp hơn Random Forest, cho thấy mô hình này xử lý các outlier hiệu quả hơn.
  + **R² (Hệ số xác định)** đạt ≈ 0.931, tức là 93.1% sự biến động của giá xe có thể được giải thích bởi các biến đầu vào. Đây là một chỉ số rất cao, phản ánh khả năng học tốt của mô hình XGBoost với dữ liệu hiện tại.
  + **Độ chính xác tương đối (100 - MAPE)** đạt ≈ 90.59%, nghĩa là mô hình dự đoán giá xe đúng khoảng 90.6% so với giá trị thực tế. Đây là độ chính xác rất cao, cho thấy XGBoost là một mô hình mạnh, đáng tin cậy trong bài toán dự đoán giá xe.
* ***Biểu đồ so sánh:***

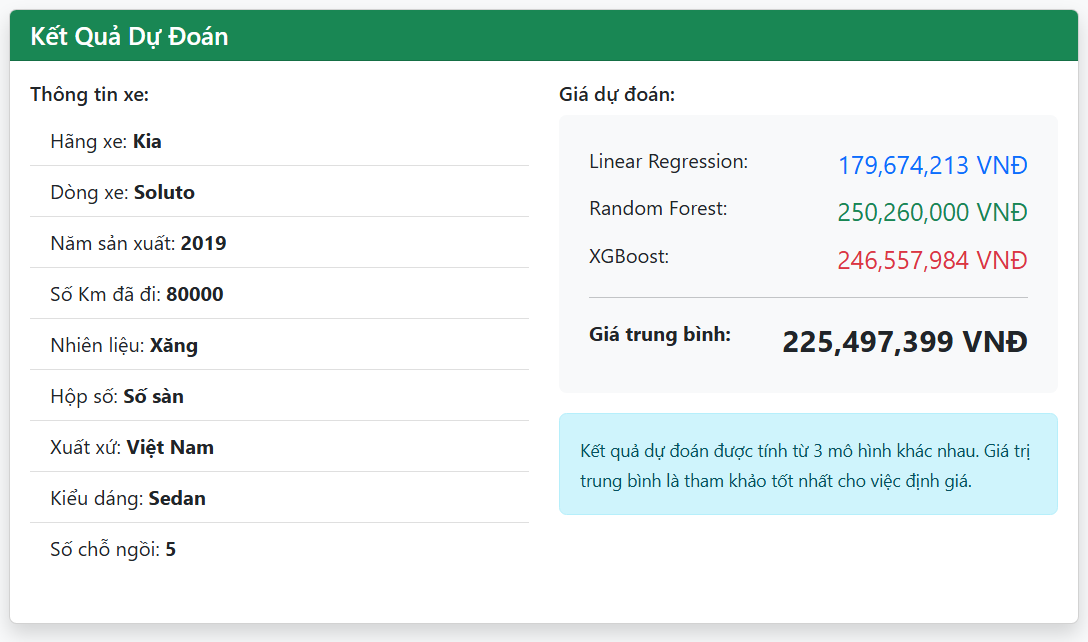


# Chức năng và cách sử dụng

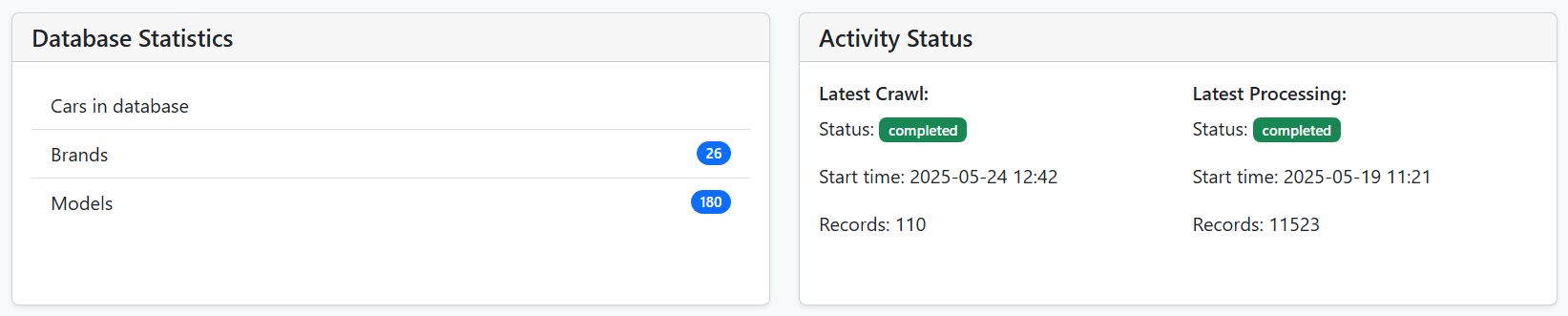
* **Chức năng:** Dự đoán giá xe cũ dựa trên các đặc điểm quan trọng của xe như năm sản xuất, hãng xe, dòng xe, quãng đường đã đi, loại nhiên liệu, số chỗ ngồi và các thông tin liên quan khác để ước lượng giá bán.
* **Cách sử dụng:**
* Người dùng
* Chọn và điền đầy đủ thông tin



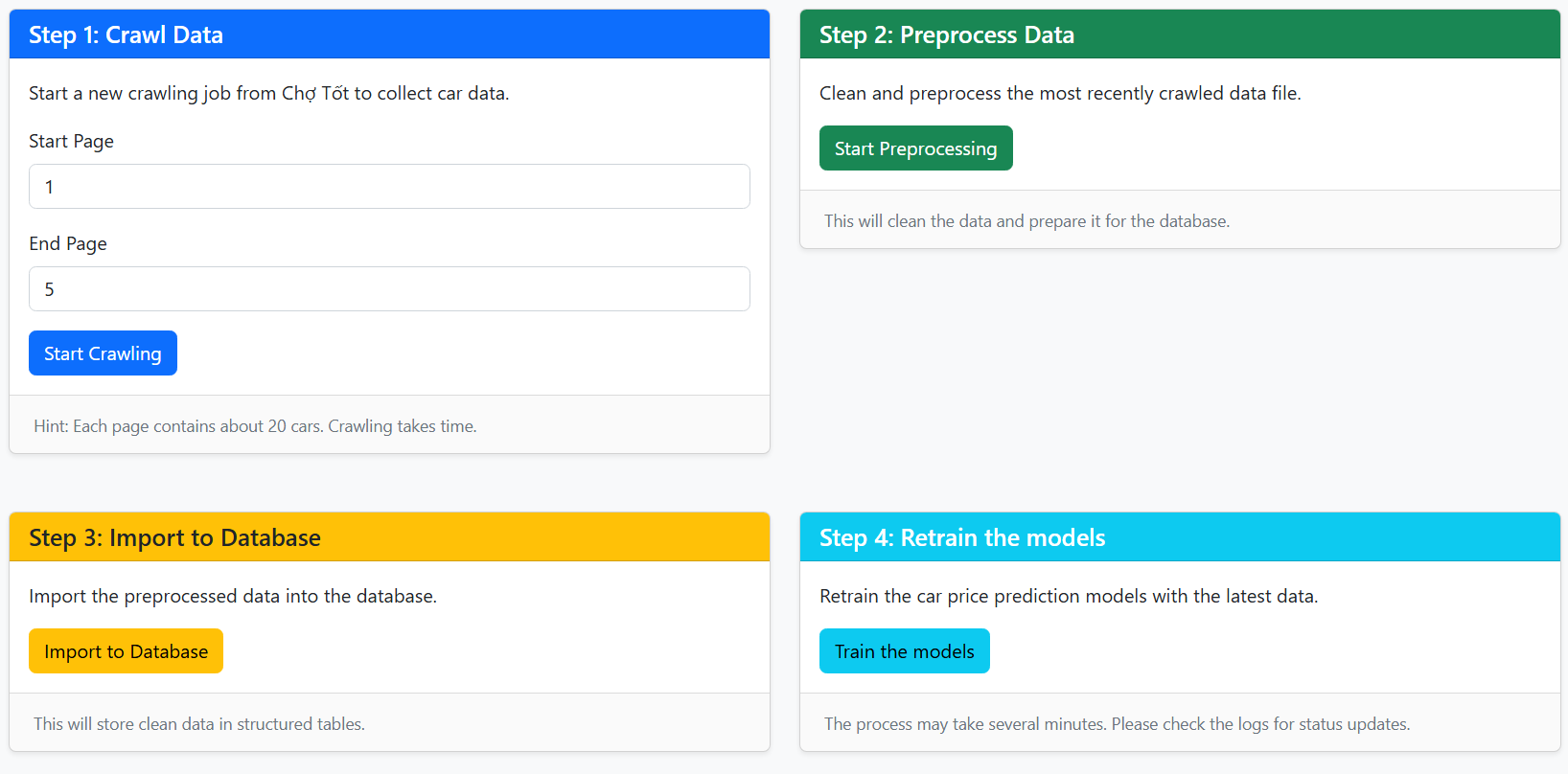
* Nhấn nút “Dự đoán giá”, kết quả dự đoán giá xe theo các mô hình sẽ hiện ra bên dưới



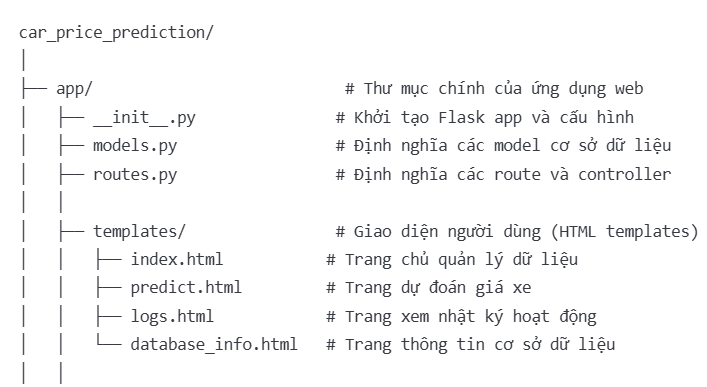
* Admin
* Thông tin dữ liệu và những lần hoạt động (crawl dữ liệu, xử lý dữ liệu)

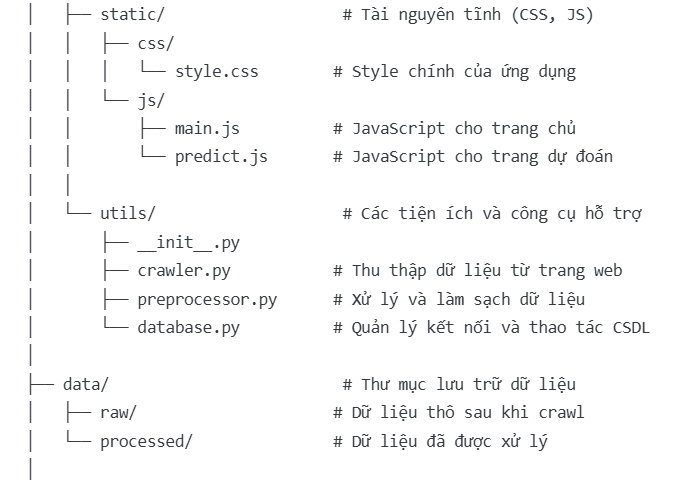


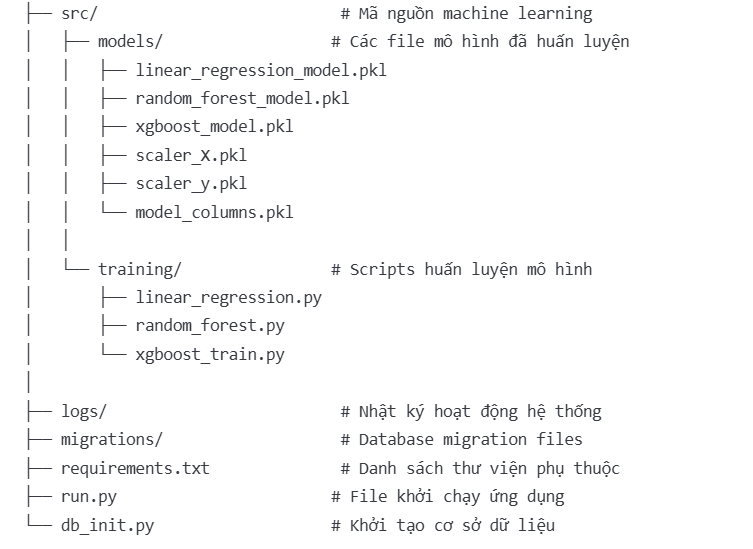
* Các bước thực hiện: crawl dữ liệu, xử lý dữ liệu, nhập dữ liệu vào database, train lại model theo dữ liệu mới nhất

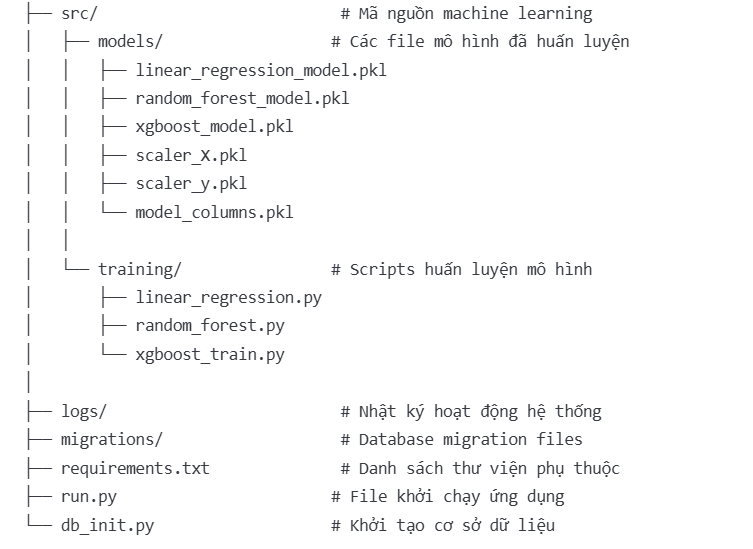


1. **Cấu trúc chương trình**









# Khó khăn, thách thức và hướng giải quyết

* Trong quá trình phân tích dữ liệu ban đầu, nhóm nhận thấy kết quả dự đoán chưa đạt độ chính xác mong muốn. Sau khi kiểm tra kỹ lưỡng, nhóm phát hiện có sự xuất hiện của các điểm ngoại lệ (outliers), cụ thể là một số ít xe có giá bán vượt trội từ 5 đến 20 tỷ đồng. Những giá trị này gây ra sự chênh lệch lớn trong phân phối dữ liệu, ảnh hưởng tiêu cực đến hiệu suất của mô hình học máy. Để khắc phục, nhóm đã tiến hành loại bỏ hoặc xử lý các giá trị ngoại lệ này nhằm giúp mô hình tập trung vào phân khúc giá phổ biến và dự đoán chính xác hơn.
* Một trong những khó khăn lớn nhất trong quá trình thực hiện dự án là việc xây dựng hệ thống crawl dữ liệu. Ban đầu, nhóm lựa chọn một trang web lớn chuyên rao bán xe cũ với kỳ vọng thu thập được nguồn dữ liệu đa dạng và phong phú. Tuy nhiên, sau khi bắt đầu thử nghiệm crawler, nhóm phát hiện trang web này đã được trang bị hệ thống chống bot rất nghiêm ngặt.

=> Sau nhiều lần thử nghiệm và đánh giá chi phí – lợi ích, nhóm quyết định chuyển hướng sang một trang web khác có nội dung tương tự nhưng có cấu trúc đơn giản và dễ tiếp cận hơn. Trang web mới không có hệ thống bảo vệ chống crawler nghiêm ngặt, cho phép nhóm dễ dàng triển khai hệ thống thu thập dữ liệu tự động. Việc thay đổi nguồn dữ liệu tuy làm tốn thêm thời gian ban đầu nhưng về lâu dài đã giúp nhóm ổn định quy trình thu thập dữ liệu, từ đó có được một tập dữ liệu đủ lớn và chất lượng để phục vụ mô hình học máy.

1. **Kết luận**

Mô hình chưa thể đạt được độ chính xác tương đối và cả nhóm cũng đã rất cố gắng để cho ra sản phẩm có đầu ra tốt nhất có thể. Qua môn học và dự án, cả nhóm đã hiểu hơn về học máy, các mô hình như hồi quy tuyến tính, rừng ngẫu nhiên, ... và cả cách ứng dụng vào trang web. Bên cạnh đó chúng em còn có thêm kiến thức về xử lý và trực quan hóa dữ liệu. Tuy rằng có rất nhiều khó khăn trong qua trình thực hiện nhưng cả nhóm đã cùng nhau vượt qua và học hỏi được rất nhiều điều từ kiến thức đến cách làm việc nhóm.

1. **Tài liệu tham khảo**
   1. <https://users.soict.hust.edu.vn/khoattq/ml-dm-course/IT3190-Tai-lieu-doc.pdf>
   2. <https://xe.chotot.com/mua-ban-oto>