ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH YẾU TỐ ẢNH HƯỞNG ĐẾN GIÁ XE Ô TÔ TẠI VIỆT NAM

Sinh viên thực hiện: **(KHÔNG ghi GVHD)**		
STT	Họ tên	MSSV
1	Trần Hoài Bão	21520628
2		
3		

TP. HỒ CHÍ MINH – 12/2022

1. GIỚI THIỆU

Đồ án môn học "Phân tích yếu tố ảnh hưởng đến giá xe ô tô tại Việt Nam" nhằm mục đích nghiên cứu và đánh giá các yếu tố quan trọng đối với sự biến động của giá xe ô tô trên thị trường Việt Nam. Việc hiểu rõ về những yếu tố này không chỉ là quan trọng trong việc định hình chiến lược kinh doanh của các nhà sản xuất và nhà phân phối ô tô, mà còn giúp người tiêu dùng có cái nhìn sâu sắc hơn về các yếu tố ảnh hưởng đến quyết định mua sắm xe của họ.

Đồ án này sẽ tập trung vào việc phân tích những yếu tố quyết định giá xe ô tô tại Việt Nam, bao gồm nhưng không giới hạn trong các lĩnh vực như kinh tế, xã hội, công nghệ và các yếu tố khác có thể đóng góp vào quá trình hình thành giá cả thị trường ô tô. Qua quá trình phân tích này, nhóm em hy vọng sẽ mang lại cái nhìn toàn diện và sâu sắc về cách mà những biến động trong môi trường nội và ngoại vi của Việt Nam có thể ảnh hưởng đến giá cả xe ô tô.

Phương pháp nghiên cứu sẽ bao gồm việc thu thập dữ liệu từ các nguồn tin cậy, sử dụng các mô hình phân tích thống kê và các phương pháp máy học để xử lý và đánh giá dữ liệu. Kết quả từ đồ án có thể cung cấp thông tin quý báu cho các doanh nghiệp ô tô, chính phủ, cũng như người tiêu dùng, để họ có thể hiểu rõ hơn về cơ cấu giá cả thị trường và đưa ra các quyết định chiến lược phù hợp.

2. MÔ TẢ BỘ DỮ LỆU

1. Bộ dữ liệu ban đầu

Thuộc tính	Mô tả	Ví dụ
Title	Tiêu đề của bài đăng ở trang web	Xe Toyota Veloz Cross Top 1.5 CVT
Date	Thông tin về ngày đăng bán	6/12/2023
Location	Địa điểm bán xe	Đại Việt Car Điện thoại: 0982 348 912 Địa chỉ: 561 Đ. Nguyễn Văn Linh, TT. Sài Đồng, Long Biên Hà

		Nội Website: daivietcar.bonbanh.com
Year	Năm sản xuất của xe	2023
Used	Trạng thái sử dụng của xe (đã sử dụng hoặc xe mới)	Xe đã dùng
Kms	Số kilomet đã đi của xe	21,000 Km
Import	Nguồn gốc nhập khẩu của xe (Nhập khẩu hoặc lắp ráp trong nước)	Nhập khẩu
Style	Kiểu dáng của xe	SUV
gearBox	Loại hộp số	Số tự động
fuelEngine	Chứa thông tin gồm loại nhiên liệu sử dụng và dung tích xy lanh	Xăng 1.5 L
Exterior	Màu sắc bên ngoài	Đen
Interior	Màu sắc nội thất	Ghi
Seats	Số chỗ ngồi	7 chỗ
Doors	Số cửa	5 cửa
Motivated	Hệ thống dẫn động	FWD – Dẫn động cầu trước
URL	Đường link đến tin bán xe tại trang web	

- Bộ dữ liệu ban đầu gồm 15,000 dòng tương ứng với 15,000 tin đăng bán xe ô tô được thu thập từ trang web bonbanh.com. Đây là một trang web đặc biệt dành cho cộng đồng đam mê và quan tâm đến lĩnh vực ô tô hoặc với những cá nhân có nhu cầu mua bán ô tô nói chung tại Việt Nam. Với sứ mệnh kết nối người mua và người bán xe ô tô, bonbanh.com đã trở thành một nền tảng trực

tuyến rất phổ biến, cung cấp thông tin chi tiết và đáng tin cậy về thị trường xe hơi trong nước.

- Đặc điểm nổi bậc của bonbanh.com:
    - **Dữ Liệu Phong Phú:** Bonbanh.com là nơi tập trung cung cấp thông tin đa dạng về các loại xe ô tô, từ xe mới đến xe đã qua sử dụng. Người dùng có thể dễ dàng tìm kiếm, so sánh và lựa chọn xe theo nhu cầu của mình.
    - **Cộng Đồng Sôi Động:** Với lượng người truy cập đông đảo, Bonbanh.com là nơi tương tác và chia sẻ thông tin giữa các độc giả, đồng thời cung cấp không gian để cộng đồng ô tô Việt Nam trao đổi kinh nghiệm và kiến thức.

2. Biến dummy
    - Trong đề tài này, nhóm chúng em ứng dụng biến dummy vào bộ dữ liệu để có thể dễ dàng xử lý các biến phân loại nhằm phục vụ cho việc áp dụng mô hình hồi quy tuyến tính lên bộ dữ liệu
    - Biến dummy (dummy variable) là một biến nhị phân được sử dụng trong các phương pháp thống kê và mô hình hóa để biểu diễn thông tin chủ yếu dưới dạng "có" hoặc "không,", biến dummy có 2 giá trị 1 hoặc 0. Biến dummy thường được tạo ra từ biến phân loại (categorical variable), còn được gọi là biến rời rạc, mà không có thứ bậc tự nhiên giữa các giá trị của nó.
    - Ví dụ trong bộ dữ liệu hiện tại có biến ‘motivated’ là một biến phân loại có 4 giá trị là “4WD – Dẫn động 4 bánh”, “AWD – 4 bánh toàn thời gian”, “FWD – Dẫn động cầu trước” và “RFD – Dẫn động cầu sau”. Để thuận tiện cho việc xử lý và phân tích, nhóm em đã thay biến “motivated” thành 4 biến dummy. Ví dụ:

Motivated
4WD – Dẫn động 4 bánh

FWD – Dẫn động cầu trước

Sau khi chuyển đổi

4WD – Dẫn động 4 bánh	AWD – 4 bánh toàn thời gian	FWD – Dẫn động cầu trước	RFD – Dẫn động cầu sau
1	0	0	0
0	0	1	0

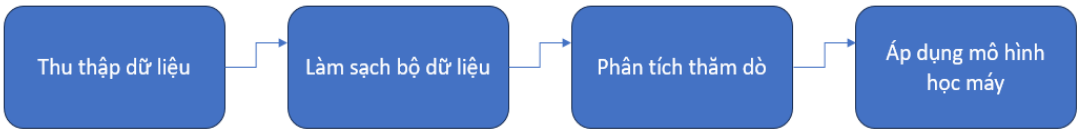
3. Bộ dữ liệu sau khi được tiền xử lý
    - Bộ dữ liệu mới sau bước tiền xử lý đã có nhiều sự thay đổi. Bộ dữ liệu mới gồm 14,994 dòng dữ liệu, với 38 dòng, trong đó gồm 37 độc lập và 1 biến phụ thuộc là giá trị "price". Sau đây là bảng mô tả tóm tắt về bộ dữ liệu:

Tên cột	Mô tả	Kiểu dữ liệu
Year	Năm sản xuất	Số nguyên
Kms	Số kilomet đã sử dụng	Số thực
Brand_over10Bs, brand_5to10Bs, brand_2to5Bs, brand_1to2Bs, brand_under1B	Lần lượt là các biến dummy thể hiện hãng xe thuộc phân khúc giá trung bình trên 10 tỷ, từ 5 đến 10 tỷ, từ 2 đến 5 tỷ, từ 1 đến 2 tỷ hoặc dưới 1 tỷ (Việt Nam Đồng)	0 hoặc 1
Nhập khẩu, Lắp ráp trong nước	Là các biến dummy thể hiện nguồn gốc của xe	0 hoặc 1
Số tay, số hỗn hợp, số tự động	Các biến dummy thể hiện sự phân loại xe về hộp số	0 hoặc 1

FWD – Dẫn động cầu trước, 4WD – Dẫn động 4 bánh, RWD – Dẫn động cầu sau, AWD – 4 bánh toàn thời gian	Các biến dummy thể hiện sự phân loại của xe về hệ thống dẫn động	0 hoặc 1
Xe đã dùng, Xe mới	Các biến dummy thể hiện tình trạng sử dụng của xe	0 hoặc 1
Van/Minivan, SUV, Hatchback, Hatchback, Bán tải/ Pickup, Sedan, Crossover, Coupe, Convertible/ Cabriolet, Truck, Wagon	Các biến dummy thể hiện phân loại xe về kiểu dáng	0 hoặc 1
numberOfDoors	Số cửa	Số nguyên
numberOfSeats	Số chỗ ngồi	Số nguyên

3. PHÂN TÍCH

Sơ đồ các bước thực hiện



3.1. Thu thập dữ liệu

- Nhóm chúng em đã sử dụng thư viện Requests, BeautifulSoup của Python để thu thập dữ liệu từ API của trang web bonbanh.com
- Quá trình thu thập dữ liệu diễn ra từ ngày 1/10/2023 – 2/11/2023. Do ban đầu tần suất cào dữ liệu quá nhanh dẫn đến việc trang web đã chặn địa chỉ IP của máy tính mà nhóm sử dụng, do đó nhóm đã sửu dụng thêm một phần

mềm VPN bên thứ ba để có thể truy cập trang web từ địa chỉ IP từ các khu vực quốc gia khác. Một phần quá trình thu thập dữ liệu diễn ra khá mất thời gian là do thiết bị phần cứng không đảm bảo, bên cạnh đó là có nhiều sự thay đổi ở kết cấu trang web bonbanh.com xảy ra trong quá trình thu thập, dẫn đến nhiều lần nhóm phải cập nhật đoạn mã Python dùng để thu thập dữ liệu.

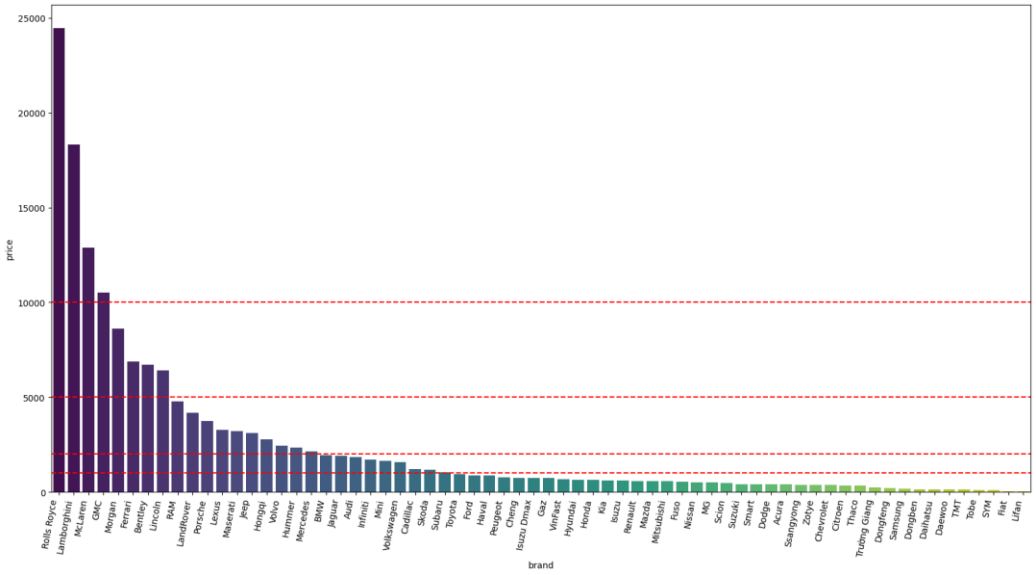
3.2. Làm sạch bộ dữ liệu

3.2.1. Loại bỏ dữ liệu dư thừa

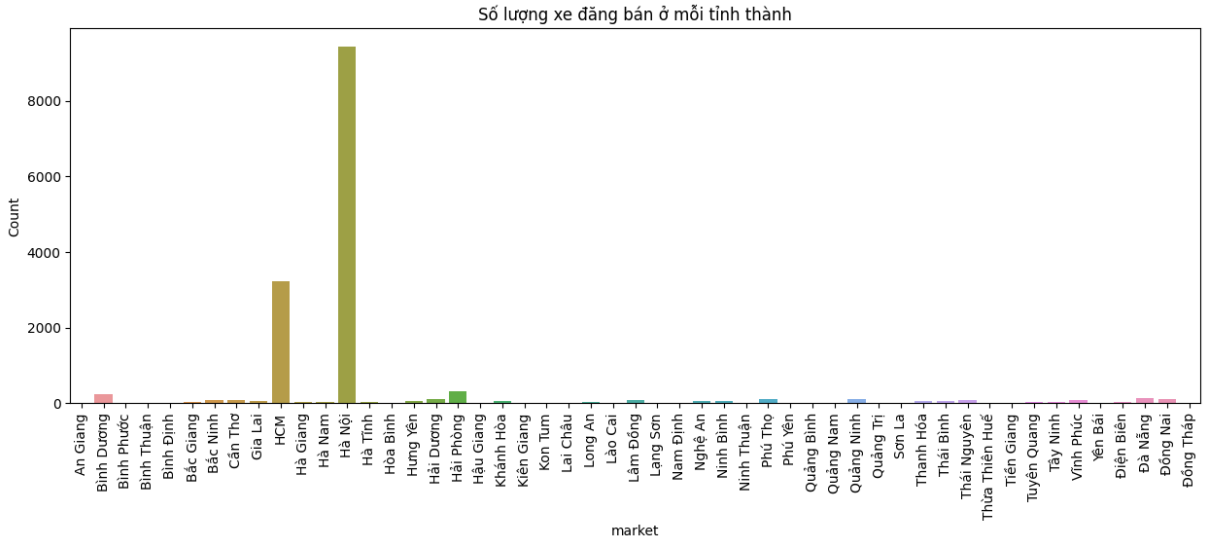
- Bộ dữ liệu ban đầu có nhiều biến không có giá trị sử dụng cao, ví dụ như biến "date" thể hiện ngày đăng tin, nhưng tại thời điểm thực hiện cào dữ liệu, biến "date" chỉ có đúng 1 giá trị (ít nhất là trong 15,000 dữ liệu được cào thành công). Bên cạnh đó, biến "exterior" và "interior" qua xem xét được nhóm em nhận thấy có độ chính xác kém (so sánh giữa giá trị 2 biến và hình ảnh xe từ trang web, lấy từ 100 mẫu dữ liệu). Do đó, nhóm em đã quyết định loại bỏ 3 cột này khỏi bảng dữ liệu ban đầu.
- Bên cạnh đó, trong 15,000 mẫu dữ liệu ban đầu có 6 mẫu bị lỗi (tất cả các biến đều không có giá trị). Do số lượng mẫu giá trị này là không đáng kể so với kích thước của bộ dữ liệu, nên nhóm đã quyết định xóa những mẫu này đi, do vậy bộ dữ liệu cuối cùng có 14,994 mẫu dữ liệu.

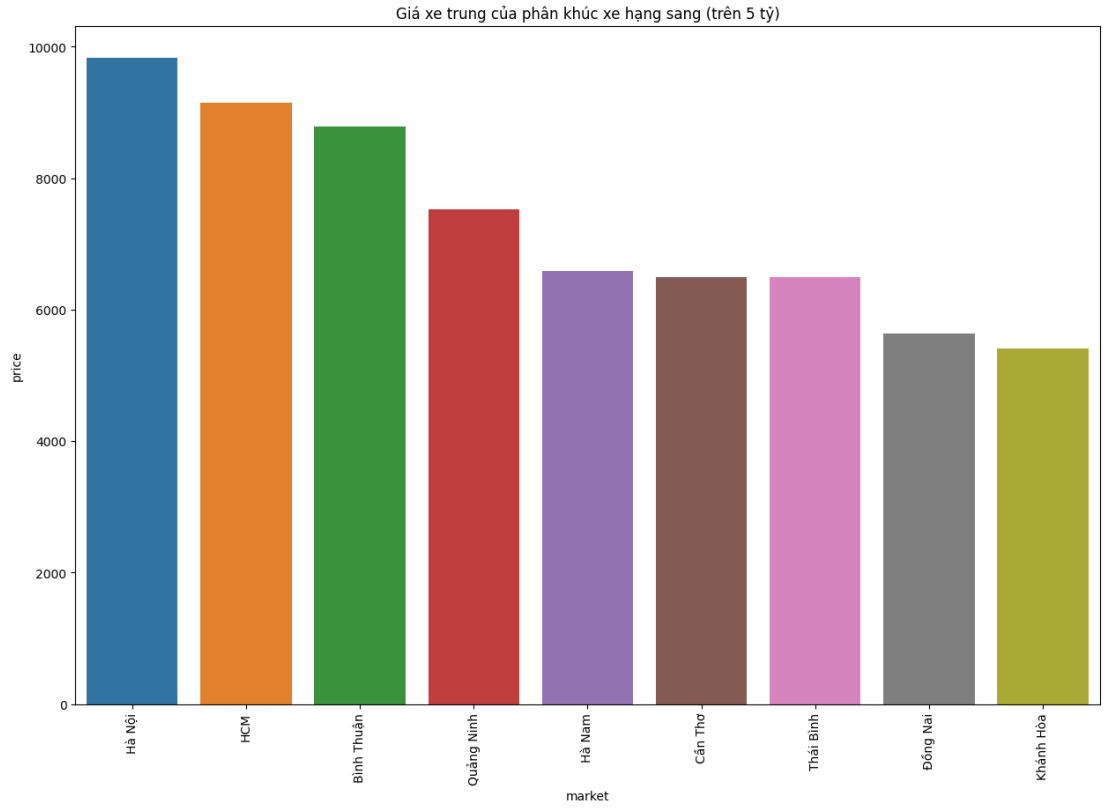
3.2.2. Chuẩn hóa dữ liệu

- Với mục tiêu áp dụng mô hình học máy lên bộ dữ liệu, nhóm em nhận thấy cần phải chuyển đổi kiểu của các biến về dạng số. Tất cả các biến có ý nghĩa tuyến tính đều được chuyển về dạng số, tất cả các b iến có ý nghĩa phân loại được chuyển về dạng các biến dummy[1].
- Có 2 trường hợp đặc biệt là ở biến "brand" và biến "location".
    - Biến "brand": thể hiện nhà sản xuất của xe. Qua thống kê, nhóm quyết định chia các giá trị của biến này thành 5 lớp theo giá xe trung bình: dưới 1 tỷ, từ 1 đến 2 tỷ, từ 2 đến 5 tỷ, từ 5 đến 10 tỷ và trên 10 tỷ.



- Biến “location”: thể hiện địa điểm (cụ thể là tỉnh/thành phố bán xe). Qua thống kê, nhóm nhận thấy có 2 thị trường có lượng xe bán cao vượt trội so với các khu vực còn lại, đó là Hà Nội và TP.HCM. Mặt khác, 2 thị trường này cũng có lượng xe sang (giá cao trên 5 tỷ VND) nhiều hơn so với các khu vực còn lại. Do đó nhóm đã tiến hành chia giá trị của biến này vào 2 nhóm thuộc thị trường Hà Nội/TP.HCM hoặc thuộc các thị trường còn lại





4. MÔ HÌNH HỒI QUY TUYẾN TÍNH

4.1. Cài đặt thư viện

- Scikit-learn[2] là một thư viện mã nguồn mở trong ngôn ngữ lập trình Python được thiết kế cho học máy. Nó cung cấp nhiều công cụ hiệu quả cho việc xử lý và phân tích dữ liệu, xây dựng và đánh giá mô hình machine learning.
- Nhóm em đã sử dụng lớp LinearRegression() của thư viện sklearn để tạo một mô hình hồi quy tuyến tính. Mô hình này đã được đào tạo trên tập dữ liệu của chúng ta.

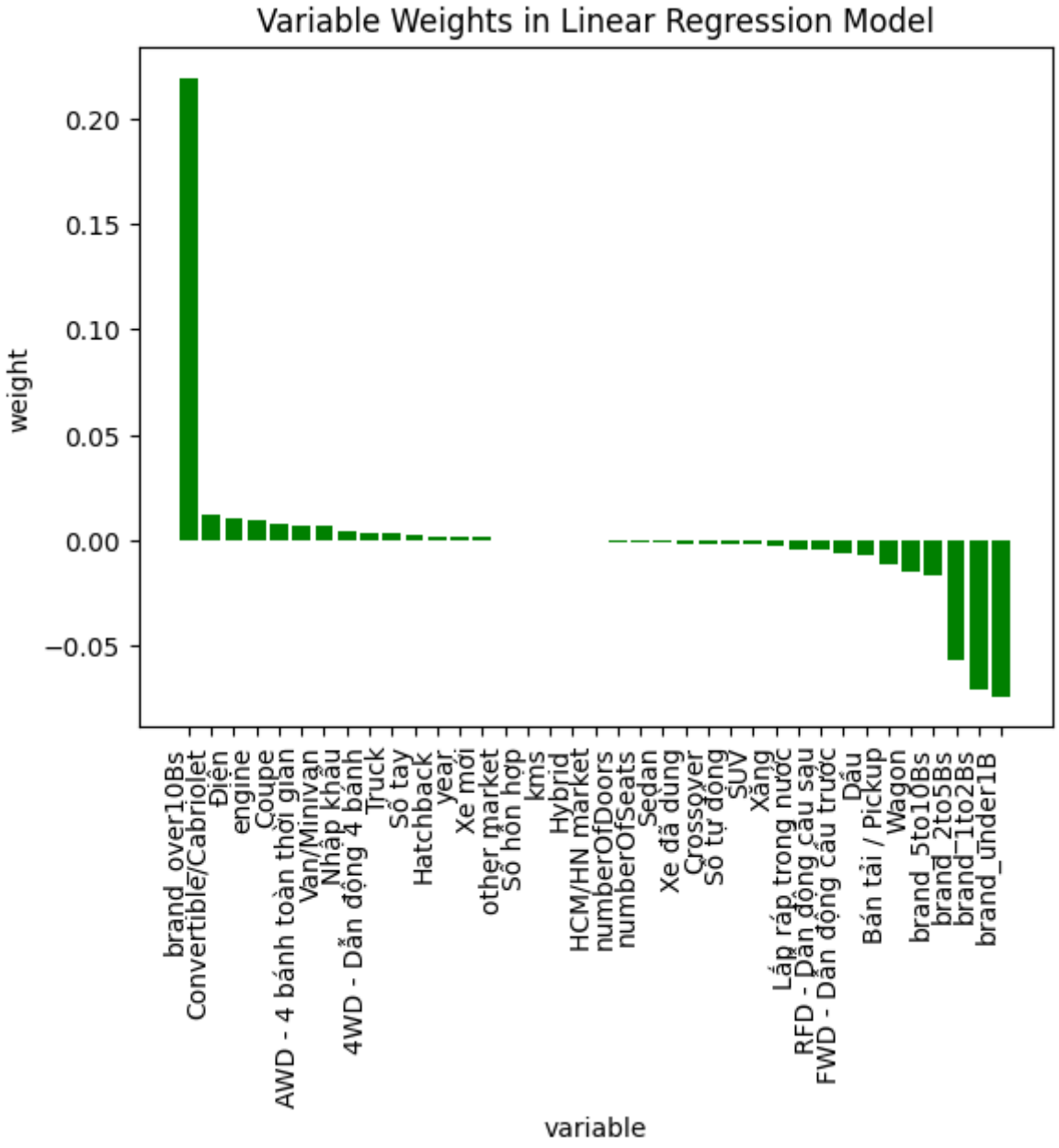
4.2. Kết quả thực nghiệm

- Kết quả sau khi áp dụng mô hình hồi quy tuyến tính (Linear regression) thì cho ra kết quả như sau:

Thang đo	Kết quả	Nhận xét
MSE	0.000568	MSE (**Mean Squared Error)** là một đo lường của sự chênh lệch giữa giá trị dự đoán và giá trị thực tế,

		được tính bằng cách lấy trung bình của bình phương các sai số. Giá trị MSE càng thấp, mô hình càng chính xác. Trong trường hợp này, MSE là 0.000568, điều này gợi ý rằng sai số trung bình giữa giá trị dự đoán và giá trị thực tế là khá nhỏ.
MAE	0.009698	MAE (**Mean Absolute Error)** là một đo lường khác của sai số trung bình giữa giá trị dự đoán và giá trị thực tế, được tính bằng cách lấy trung bình của giá trị tuyệt đối của sai số. Giá trị MAE càng thấp, mô hình càng chính xác. Trong trường hợp này, MAE là 0.009698, điều này cũng gợi ý rằng sai số trung bình là khá nhỏ.
R-squared	0.644	R-squared đo lường mức độ giải thích của mô hình đối với sự biến động của dữ liệu. Nó có giá trị từ 0 đến 1, và giá trị càng gần 1 thì mô hình càng tốt. Trong trường hợp này, R-squared

		là 0.644, điều này cho biết mô hình giải thích được khoảng 64.4% biến động của dữ liệu.



5. KẾT LUẬN

Từ những phân tích trên, nhóm em rút ra nhận xét rằng các yếu tố có ảnh hưởng nhiều đến giá xe bao gồm: Hãng xe, kiểu dáng, nhiên liệu, dung tích xy lanh và hệ thống dẫn động.

TÀI LIỆU THAM KHẢO

[1] What Are Dummy Variables And How To Use Them In A Regression Model – Time Series Analysis, Regression, and Forecasting (timeseriesreasoning.com)

[2] Getting Started — scikit-learn 1.3.2 documentation

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Trần Hoài Bão	Thu thập dữ liệu
		Làm sạch bộ dữ liệu
		Phân tích, áp dụng mô hình học máy