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

PHÂN TÍCH BỘ DỮ LIỆU VÀ XÂY DỰNG MÔ HÌNH DỰ ĐOÁN GIÁ XE Ô TÔ CỔ ĐIỂN CŨ TRÊN THỊ TRƯỜNG QUỐC TẾ

Sinh viên thực hiện:			
STT	Họ tên	MSSV	Ngành
1	Trương Khánh Long	21521750	CNTT
2	Nguyễn Chí Thi	21522614	CNTT
3	Bùi Lê Trọng Đức	21520725	CNTT
4	Hoàng Hải Anh	21521819	CNTT
5	Nguyễn Tiến Thịnh	21521472	CNTT

TP. HỒ CHÍ MINH – 12/2023

1. GIỚI THIỆU:

Trong đề tài này, chúng tôi tập trung vào việc phân tích bộ dữ liệu giá xe hơi cổ điển và lựa chọn mô hình thích hợp nhất để dự đoán giá thành của các loại ô tô đã qua sử dụng và có nhu cầu rao bán. Do thời điểm kinh tế hiện nay cũng đang gặp nhiều khó khăn, việc sở hữu một chiếc xe hơi mới trở thành một vấn đề vô cùng lớn đối với phần lớn các gia đình. Vì vậy, xu hướng mọi người sẽ tìm kiếm các ô tô đã qua sử dụng để có thể vừa tiết kiệm chi phí vừa đáp ứng được nhu cầu di chuyển hằng ngày. Việc nắm rõ các thông tin để dự đoán giá và so sánh chúng với giá hiện tại đang được niêm yết tại cửa hàng hoặc giá xe ô tô mới có thể mang lại nhiều lợi ích cho người tiêu dùng trong việc lựa chọn sản phẩm phù hợp, đặc biệt là lợi ích về mặt kinh tế.

Để hiện thực được toàn bộ đề tài, nhóm chúng tôi sử dụng ngôn ngữ lập trình chính và đóng vai trò nền tảng cho toàn bộ quy trình là Python. Python - một ngôn ngữ gần gũi đối với các lập trình viên một phần cú pháp dễ dàng nắm bắt nhưng điểm quan trọng nhất để nhóm quyết định chọn sử dụng chính là có nhiều thư viện hỗ trợ lập trình thuật toán, trực quan hóa dữ liệu, các mô hình máy học, các phương pháp tiền xử lý dữ liệu,… Một số thư viện được nhóm sử dụng trong đề tài như: Numpy, Pandas, Matplotlib, Seaborn, Pyplot, Sklearn, Random, Scipy,… . Phương pháp chúng tôi lựa chọn để xây dựng mô hình dự đoán giá chính là áp dụng các thuật toán mày học, bao gồm các mô hình hồi quy đơn biến, đa biến, đa thức và mô hình mạng nơ-ron để dựa trên các đặc trưng được truyền vào có thể đưa ra giá cả hợp lý nhất.

Nội dung bài báo cáo của chúng tôi bao gồm quá trình tự thu thập dữ liệu có sẵn trên website, tiến hành tiền xử lý và trực quan hóa từng cột thuộc tính trong bộ dữ liệu. Sau đó, nhóm chúng tôi tiền hành thăm dò dữ liệu để tìm ra các biến phụ thuộc có ảnh hưởng đến biến mục tiêu (giá xe dự đoán). Kế đến, nhóm tiến hành thử nghiệm trên nhiều mô hình máy học và chọn ra mô hình có độ chính xác cao nhất trong việc dự đoán giá xe ô tô.

Để đảm bảo tính khách quan và giá trị của bộ dữ liệu khi phân tích, nhóm đã thảo luận thống nhất chọn một trang web uy tín và nổi tiếng trên thế giới về xe ô tô cổ điển là ClassicCar. Trang web có thể hỗ trợ người dùng trong việc tìm kiếm, mua bán và trao đổi các loại xe hơi cổ điển và hiếm có từ khắp nơi trên thế giới.

2. MÔ TẢ BỘ DỮ LIỆU

Bộ dữ liệu được sử dụng trong đề tài được nhóm tự tham khảo và thu thập tại [1] vào ngày 31/10/2023. Để có thể thu thập một lượng dữ liệu tương đối lớn từ trang web, nhóm chúng em đã sử dụng các thư viện Beautiful Soup [2] và Requests [3] hỗ trợ kết hợp với ngôn ngữ lập trình Python.

Bộ dữ liệu khi vừa mới thu thập trực tiếp từ trang web bao gồm 25 cột thuộc tính trong đó có duy nhất một biến liên tục (Year of manufacture) và 24 biến phân loại với 11100 điểm dữ liệu.

Thông tin chi tiết bộ dữ liệu được mô tả chi tiết như sau:

STT	Tên Thuộc Tính	Kiểu dữ liệu	Mô tả
1	Make	Object	Tên của nhà sản xuất hoặc hãng sản xuất ô tô.

2	Series	Object	Loại xe hoặc dòng xe cụ thể thuộc nhà sản xuất. Ví dụ: Dòng xe Delta có các mô hình như "Delta HF 4WD", "Delta HF Integrale 16V" …
3	Model	Object	Tên mô hình cụ thể của xe.
4	Manufacturer code	Object	Mã nhà sản xuất hoặc mã ô tô.
5	First registration date	Object	Thời điểm đầu tiên ô tô được đăng ký và có thể lưu hành.
6	Chassis number	Object	Mã số khung giúp định danh ô tô.
7	Engine number	Object	Mã số động cơ giúp định danh ô tô.
8	Gearbox number	Object	Mã số hộp số giúp định danh ô tô.
9	Matching numbers	Object	Kiểm tra sự đồng nhất các mã số khung, mã số động cơ, mã số hộp số của ô tô. Có hai loại giá trị: Yes và No.
10	Body style	Object	Kiểu dáng bên ngoài của xe.
11	Steering	Object	Vị trí của hệ thống lái tùy theo quy định của quốc gia. Có hai loại giá trị: Left (LHD) và Right (RHD).
12	Gearbox	Object	Loại hộp số được sử dụng trong ô tô. Có ba loại giá trị: Manual, Automatic và Semi-Automatic.
13	Transmission	Object	Loại truyền động của ô tô. Có ba giá trị: Front, Rear, 4WD. (1)
14	Front brakes	Object	Loại hệ thống phanh trước được trang bị trên ô tô. Có hai loại giá trị: Disc và Drum. (2)
15	Rear brakes	Object	Loại hệ thống phanh sau được trang bị trên ô tô. Có hai loại giá trị: Disc và Drum.
16	Fuel type	Object	Loại nhiên liệu ô tô sử dụng để hoạt động. Có ba loại giá trị: Petrol, Diesel và Electric.
17	Previous Owner	Object	Số lượng chủ sở hữu trước của ô tô. Có bảy loại giá trị: 1,2,3,4,5,6 và > 6

18	Mileage (read)	Object	Số dặm ô tô đã chạy được dựa trên đồng hồ đo dặm của xe. Đơn vị tính là dặm và km. (3)
19	Power (kW/hp)	Object	Công suất thực tế và công suất khi sản xuất trong nhà máy (nếu có) của động cơ ô tô.
20	Cubic capacity (ccm)	Object	Dung tích xi-lanh thực tế và khi sản xuất trong nhà máy (nếu có) của động cơ ô tô.
21	Cylinders	Object	Số lượng xi-lanh trong động cơ ô tô. Miền giá trị: [1;16]
22	Doors	Object	Số lượng cửa của ô tô. Miền giá trị: [0,6].
23	Gears	Object	Số lượng bánh răng trong hộp số của ô tô. Miền giá trị: [1;9]
24	Price (£)	Object	Giá xe ô tô được niêm yết trên website. Đơn vị tính là Pound (Bảng Anh).
25	Year of manufacture	Int64	Năm sản xuất của ô tô. Miền giá trị: [1885;2023]

Chú thích thuật ngữ:

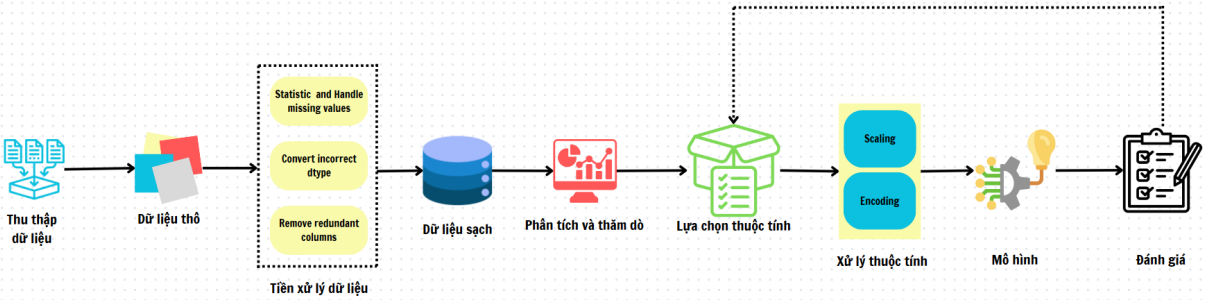
(1): Truyền động là cách động cơ chuyển đổi và truyền tải năng lượng đến bánh xe để tạo ra chuyển động. Front: chỉ truyền đến bánh trước; Rear: chỉ truyền đến bánh sau; 4WD: Tất cả bốn bánh đều được truyền động.

(2): Disc: hệ thống sử dụng đĩa kim loại xoay chuyển đồng thời với bánh xe, thường được sử dụng trong các xe ô tô đòi hỏi hiệu suất phanh cao, như xe thể thao và xe hạng sang; Drum: hệ thống sử dụng một hộp kim loại tròn xoay chuyển với bánh xe, thường được sử dụng ở các bánh xe sau.

(3): Theo quy chuẩn quốc tế, 1 dặm = 1.609344 km hay 1 km = 0.621371192 dặm.

3. PHƯƠNG PHÁP PHÂN TÍCH:

Sau khi thu thập dữ liệu từ trang web, quá trình phân tích dữ liệu được thể hiện chi tiết qua sơ đồ bên dưới:

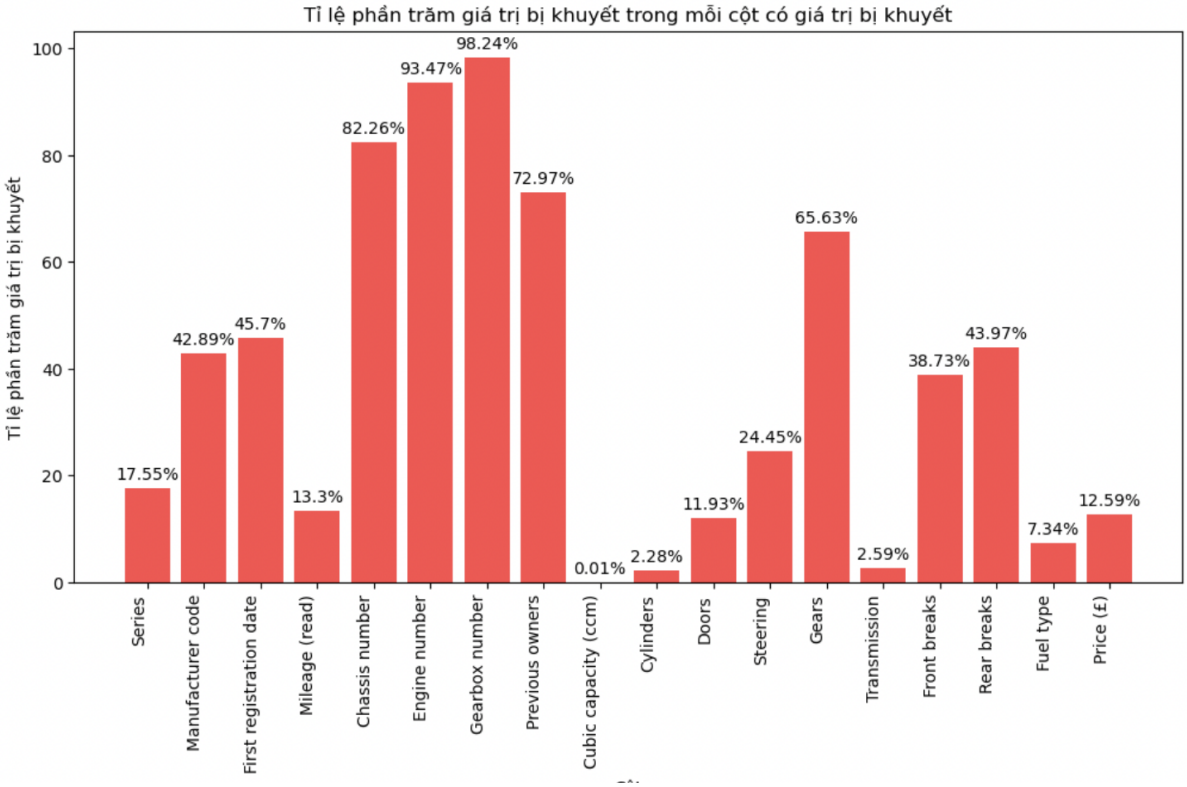


Hình 1. Sơ đồ quy trình thu thập và phân tích dữ liệu

3.1. Tiền xử lý dữ liệu:

3.1.1. Thống kê số lượng giá trị bị khuyết trong từng biến:

Bộ dữ liệu ban đầu có những giá trị đặc biệt đại diện cho giá trị khuyết ở các biến bao gồm: "Not provided", "price on request", "Not specified". Vì vậy, để có thể thống kê được, nhóm đã thay thế các giá trị trên thành giá trị NaN. Từ đó, nhóm đã có bảng thống kê như sau:



Hình 2. Bảng thống kê số lượng giá trị bị khuyết trong từng biến

Các cột thuộc tính "Gearbox number", "Chassis number", "Engine number", "Matching number", "First registration date" có số lượng giá trị khuyết lớn và chức năng của các biến này chỉ dùng để định danh phụ kiện cho xe. Thuộc tính "Series" do có nhiều 895 giá trị phân loại khác nhau nên chúng em quyết định loại bỏ chúng để đơn giản hóa quá trình phân tích.

3.1.2. Xử lý các giá trị bị khuyết:

Các cột thuộc tính còn lại chúng em lược bỏ các giá trị khuyết bằng cách xóa đi các hàng trong bộ dữ liệu. Riêng đối với biến "Mileage (read)", chúng em điền khuyết bằng cách sử dụng thuật toán KNN (K-Neighrest Neighbor). Phương pháp giúp tìm ra những giá trị tương đồng dựa trên số lượng (k) dòng dữ liệu xung quanh nó.

3.1.3. Định dạng dữ liệu:

Trong quá trình thu thập dữ liệu, để tối ưu hóa tốc độ cũng như lượng dữ liệu thu thập được, nhóm đã để hầu hết toàn bộ kiểu dữ liệu của các biến là kiểu Object. Vì vậy,

trước khi bước vào phân tích thăm dò, dữ liệu cần phải được chuyển về đúng kiểu và đúng định dạng của nó.

- "Mileage (read)": có kiểu dữ liệu là object và kèm theo đơn vị bao gồm km và mls. Trước khi có thể chuyển về kiểu dữ liệu số phục vụ cho việc tính toán, nhóm đã loại bỏ toàn bộ các đơn vị phía sau. Nếu hàng đó có đơn vị là km, nhóm sẽ đổi về dặm dựa theo (3) và lưu trở ngược lại vào trong bộ dữ liệu và chuyển về kiểu dữ liệu số nguyên.
- "Power (kW/hp)": lấy giá trị mã lực (hp) thực tế và lưu vào bộ dữ liệu với tên cột với là "Horsepower", kiểu dữ liệu số nguyên.
- "Cubic capacity (ccm)": lấy giá trị dung tích xi lanh thực tế và đổi kiểu dữ liệu của cột về số nguyên.
- "Price (£)": Xóa bỏ dấu phẩy ngăn cách giữa các dữ liệu và đổi kiểu dữ liệu của cột về số nguyên nhờ vào sự hỗ trợ của thư viện re.

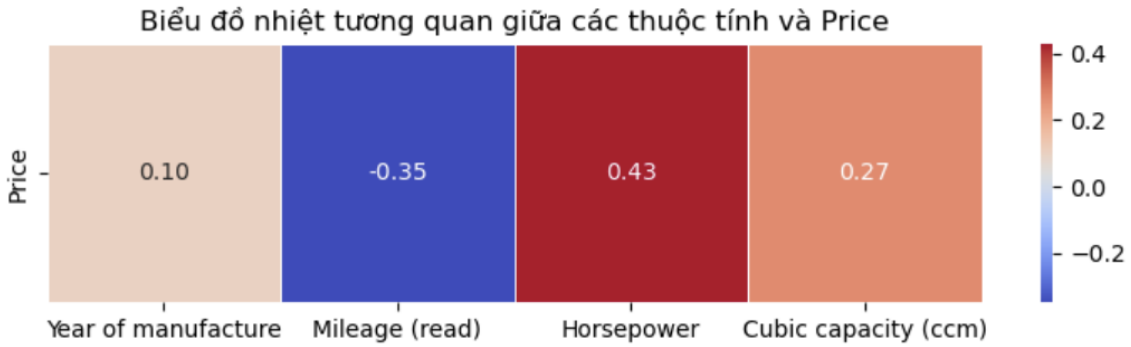
3.2. Phân tích và thăm dò dữ liệu:

Bộ dữ liệu sau khi tiền xử lý còn lại 18 cột thuộc tính trong đó có năm cột thuộc tính kiểu số, với 601 điểm dữ liệu không chứa bất kỳ giá trị khuyết nào. Thông qua việc trực quan hóa các biến dữ liệu kiểu số kết hợp với việc sử dụng IQR (Interquartile Range), nhóm đã tiến hành phát hiện ra một số điểm dữ liệu ngoại lệ trong bộ dữ liệu và tiến hành xóa chúng ra khỏi bộ dữ liệu để đơn giản hơn trong quá trình phân tích.

Để đánh giá sự ảnh hưởng của các biến độc lập tới giá trị của biến phụ thuộc, nhóm đã tiến hành chia bộ dữ liệu ra làm hai loại là bộ dữ liệu kiểu số (gồm 4 thuộc tính) và bộ dữ liệu kiểu phân loại (gồm 12 thuộc tính) không bao gồm biến 'Model'. Nhóm có xử lý một vài giá trị tại biến phân loại bằng phương pháp gom nhóm trước khi tiến hành phân tích:

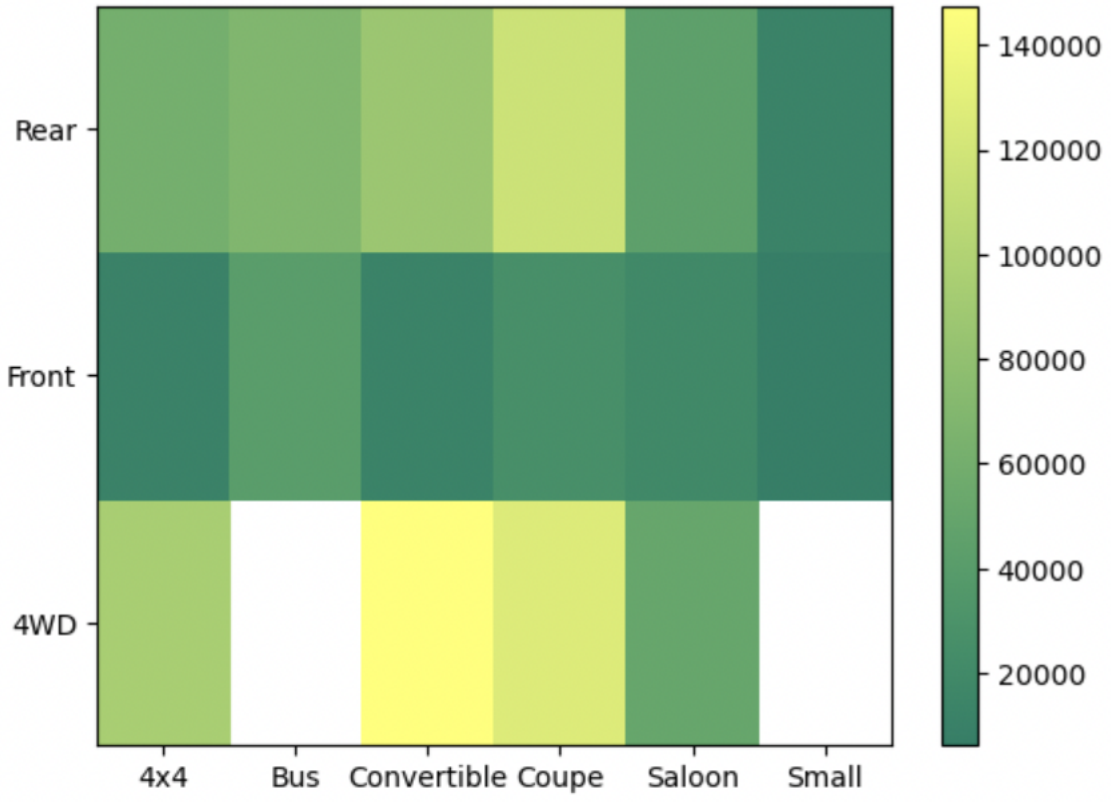
- "Make": gom nhóm các điểm dữ liệu thành ba loại: "Luxury", "Sports", "Compact" lưu vào cột "Type" và đưa vào bộ dữ liệu để phục vụ phân tích thay thế cho cột "Make".
- "Body Style": gom nhóm các điểm dữ liệu có từ bắt đầu giống nhau về cùng một loại. Kết quả thu được 11 nhóm dữ liệu so với 34 nhóm dữ liệu trước đó.
- Tiến hành xóa một vài điểm dữ liệu có số lần xuất hiện ít trong từng cột dữ liệu.

Với bộ dữ liệu kiểu số, nhóm sử dụng phương pháp hệ số tương quan Pearson để phân tích mức độ ảnh hưởng của từng biến đến biến mục tiêu (Price (£)). Độ tương quan được thể hiện trong hình bên dưới. Biến có mức độ tương quan cao nhất là biến Horsepower.



Hình 3. Mức độ tương quan giữa các biến dữ liệu kiểu số với giá tiền

Với bộ dữ liệu phân loại, ta có thể đánh giá mức độ tương quan rõ ràng hơn thông qua hình bên dưới. Khi gom nhóm dữ liệu theo Body style (kiểu dáng – trục hoành), Transmission (loại truyền động – trục tung), phân bố giá thành có sự đa dạng thể hiện qua độ đậm nhạt của màu sắc. Ta có thể cân nhắc lựa chọn các biến này làm dữ liệu đầu vào để xây dựng và phát triển mô hình.



Hình 4. Mức độ tương quan giữa các biến dữ liệu kiểu phân loại với giá tiền

3.3. Lựa chọn và chuẩn hóa thuộc tính:

Qua quá trình phân tích, thăm dò dữ liệu bằng phương pháp sử dụng tương quan Pearson (đối với các biến kiểu số) và phân tích ANOVA (đối với các biến kiểu phân loại), nhóm đã tìm ra được các thuộc tính có mức độ ảnh hưởng cao đến biến mục tiêu bao gồm: “Horsepower”, “Mileage (read)”, “Type”, “Previous owners”, “Body style”,

"Cylinders", "Doors", "Gears", "Transmission", "Rear breaks". Bộ dữ liệu sau khi được loại bỏ một vài giá trị đến thời điểm này còn 602 dòng dữ liệu được sử dụng.

Đối với các biến kiểu số, để đảm bảo tính đúng đắn của giá trị dự đoán, chúng em đã sử dụng phương pháp Standard Scaler [4] để đưa chúng về cùng một giá trị trước khi đưa vào mô hình. Còn đối với các biến phân loại, nhóm tiến hành sử dụng phương pháp Label Encoding [4] toàn bộ các giá trị chuỗi thành số để mô hình có thể hiểu được.

4. KẾT QUẢ PHÂN TÍCH:

Sau khi tiến hành chuẩn hóa dữ liệu và thực nghiệm nhiều lần trên các mô hình máy học cũng như các mô hình học sâu như: Linear Regression, Lasso Regression, Ridge Regression và Random Forest, nhóm đã thu thập được kết quả dự đoán giá như bảng bên dưới. Độ chính xác của mô hình được tính toán dựa trên các thang đo phổ biến: Mean Square Error, Root Mean Square Error và R-squared. Trong đó, chúng em lấy R-squared làm tiêu chí quan trọng nhất trong việc lựa chọn mô hình thích hợp.

Bảng bên dưới thể hiện độ chính xác khi dự đoán của các mô hình:

Mô hình	Tham số tối ưu	R^2	RMSE	MSE
Linear Regression	None	0.3934	34349.56	1179892288.75
Ridge Regression	Alpha = 10 Max_iter = 100	0.3941	34332.18	1178698538.6617
Lasso Regression	Alpha = 10 Max_iter = 100	0.3934	34352.19	1180072771.97
Random Forest	**Max_depth = 20** **n_estimators = 100**	**0.5473**	**29673.46**	**880514395,08**

Random Forest có hiệu suất tốt hơn so với ba mô hình khác trong bảng đánh giá. Với R^2 cao nhất (0.5473), MSE thấp nhất (29673.46), và RSS thấp nhất (880514395.08), Random Forest vượt trội trong khả năng dự đoán. Sự đa dạng của nhiều cây quyết định ngẫu nhiên giúp mô hình này giảm overfitting và làm giảm ảnh hưởng của nhiễu, trong khi khả năng xử lý tương tác phi tuyến và đặc trưng quan trọng làm tăng tính linh hoạt. Trong khi Linear Regression, Ridge Regression và Lasso Regression cũng cho kết quả tốt, Random Forest tỏ ra là lựa chọn xuất sắc với khả năng tối ưu hóa hiệu suất dự đoán.

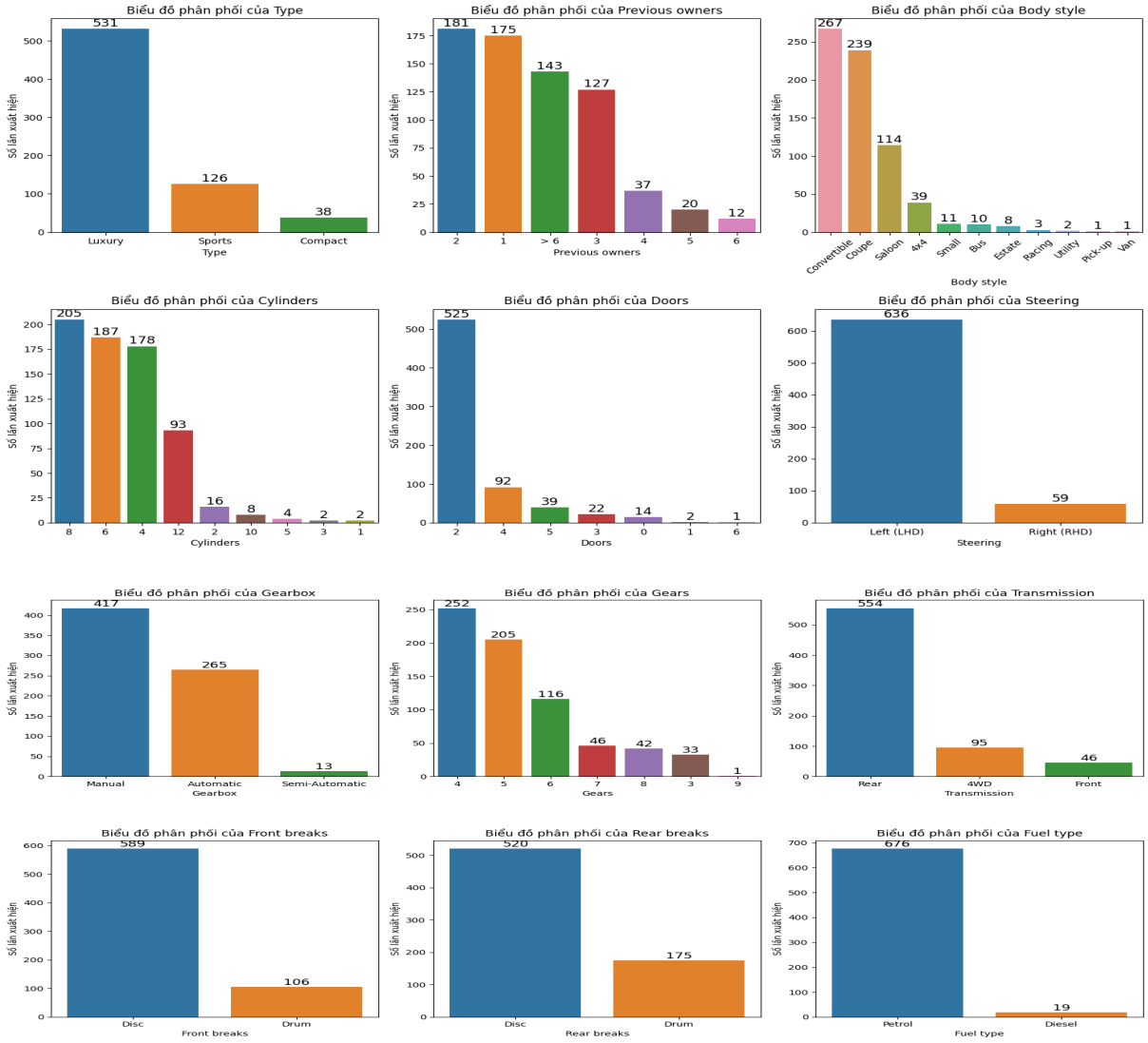
Nhóm có xem xét sử dụng mô hình hồi quy đa thức để tiền xử lý dữ liệu trước khi đưa vào mô hình hồi quy tuyến tính để dự đoán giá xe. Sau quá trình thử nghiệm qua nhiều bậc, trực quan kết quả bằng hình ảnh và số liệu, nhóm đã phát hiện ra bậc 2 là bậc

tốt nhất đối với bộ dữ liệu của nhóm. Kết quả thử nghiệm được thể hiện qua hình bên dưới:

Bậc	R^2	MSE	RMSE
1	0.2228	2293841407.3009	47894.0644
2	**0.2058**	**2344106226.2893**	**48415.9708**
3	-233493091205563 37987584	689138172374909 93006485962752	830143464935748
4	-248466299664089 36	733330527096929 50674669568	8563472000870.49 71

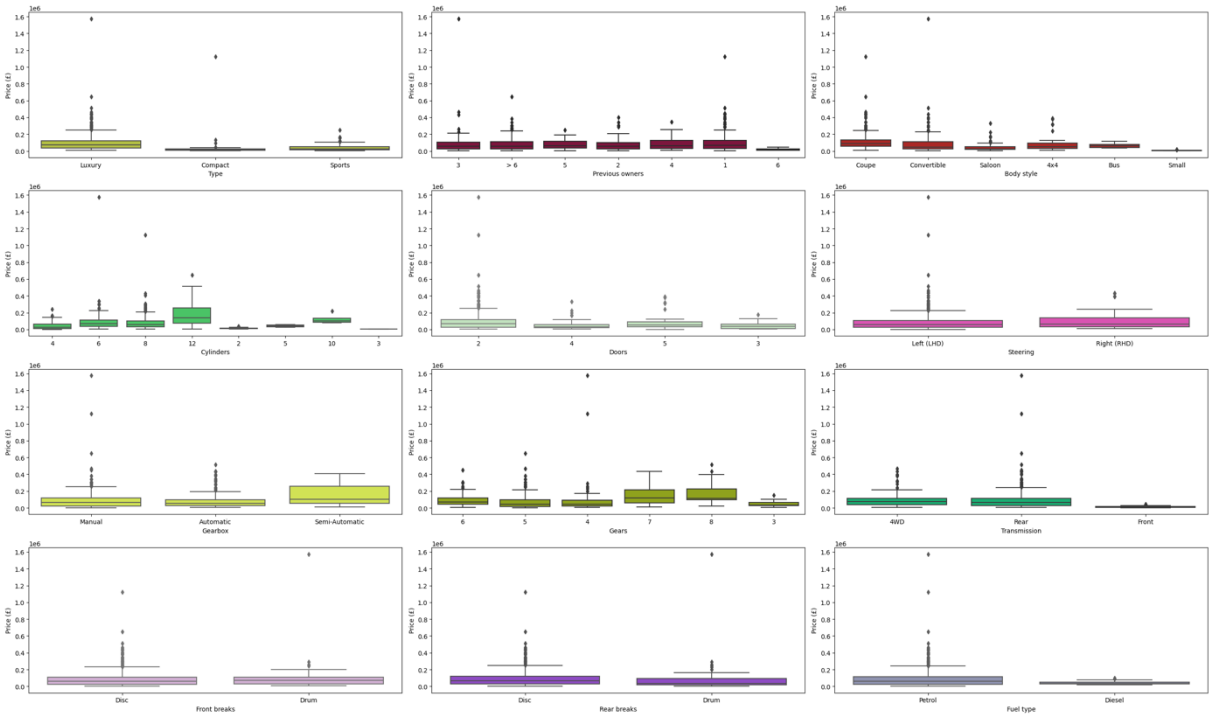
Tuy nhiên, với phương pháp xử lý dữ liệu theo hồi quy đa thức cho kết quả hiệu suất thấp hơn ở cả 3 thang đo nên chúng em sẽ lựa chọn phương pháp đầu tiên là chuẩn hóa dữ liệu và chọn mô hình Random Forest làm mô hình chính để dự đoán giá xe ô tô

5. TRỰC QUAN HÓA DỮ LIỆU



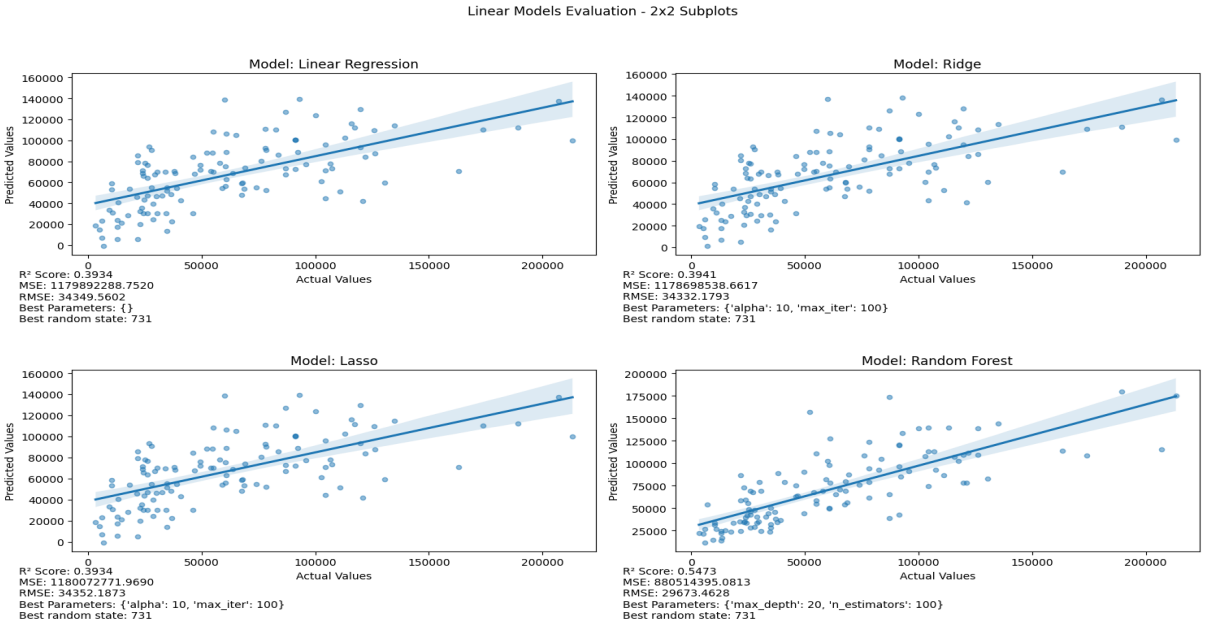
Hình 5. Trực quan một số cột dữ liệu phân loại bằng Histogram

Các biểu đồ trên thể hiện số lượng giá trị trong từng cột thuộc tính được gom nhóm lại với nhau. Nhìn tổng quan, nhóm em đã đưa ra được nhận xét rằng dữ liệu có sự phân bố rộng rãi. Tuy nhiên, một vài cột thuộc tính có nhóm giá trị chiếm phần lớn có thể kể đến như: “Front brakes”, “Fuel type”, “Steering”. Điều này có thể dẫn đến việc lệch lạc trong quá trình phân tích và chúng ta cần phải có phương pháp dự đoán và xử lý phù hợp với các cột thuộc tính này trong quá trình xây dựng mô hình.



Hình 6. Mối quan hệ giữa các thuộc tính phân loại với biến mục tiêu

Các biểu đồ trên thể hiện mối quan hệ của từng nhóm giá trị dữ liệu của các biến phân loại đối với biến mục tiêu “Price (£)”. Qua quan sát ban đầu, chúng ta có thể nhận thấy tại các vị trí gần nhau của các giá trị thuộc tính, giá thành cũng chúng có sự chênh lệch không đáng kể. Tuy nhiên, vẫn còn một vài vị trí tuy cùng giá trị thuộc tính nhưng lại có giá thành vượt trội hơn so với các điểm dữ liệu còn lại.



Hình 7. Trực quan hiệu suất của từng mô hình dùng để huấn luyện bộ dữ liệu

Bốn biểu đồ trên được sử dụng để đánh giá hiệu suất của các mô hình. Các điểm chấm chính là giá trị ban đầu trong tập test còn đường thẳng chính thể hiện mối tương quan giữa giá trị dự đoán lấy được từ mô hình và giá trị thực. Lướt qua bốn biểu đồ, ta có thể nhận xét: Tại biểu đồ thứ tư, các điểm dữ liệu có vẻ nằm sát với đường thẳng tương quan hơn do đó sai số của mô hình cũng thấp hơn. Qua kết quả tính toán trực tiếp bằng số liệu, ta cũng đã chứng minh được Random Forest là mô hình tốt nhất trong bốn mô hình được sử dụng để xây dựng.

6. KẾT LUẬN

Thông qua toàn bộ nội dung được trình bày trong đồ án, nhóm đã tự thu thập, phân tích và xây dựng mô hình định giá ô tô cổ điển. Người tiêu dùng cũng như người bán có thể xem bài báo cáo này như là một nguồn tham khảo để có thể phục vụ cho việc mua bán và kinh doanh. Thông qua việc trực quan hóa dữ liệu, nhóm cũng đã phát hiện ra được những đặc tính chung nhất, phổ biến nhất của toàn bộ xe ô tô được bán. Và sau đó dùng các thuật toán, các phương pháp để đưa ra các yếu tố có ảnh hưởng nhất đến giá thành của một chiếc xe.

Trong số các mô hình được sử dụng, Random Forest cho hiệu suất R^2 cao nhất nhưng chỉ với con số khiêm tốn là 0.5473. Có nhiều nguyên nhân dẫn đến hiện tượng này như: bộ dữ liệu thu thập còn hạn chế, sai sót trong quá trình xử lý các thuộc tính và các mô hình máy học sử dụng còn đơn giản. Trong tương lai, nhóm sẽ tiếp tục nghiên cứu và phát triển mô hình đạt hiệu suất cao nhất có thể để trở thành một nguồn tham khảo giá đáng tin cậy đối với những người đã, đang và sẽ có nhu cầu mua xe ô tô .

TÀI LIỆU THAM KHẢO

[1] .Classic Trader. Link: Classic Cars for Sale on Classic Trader | www.classic-trader.com (28/10/2023)

[2] Beautiful Soup. Link: Beautiful Soup Documentation — Beautiful Soup 4.4.0 documentation (beautiful-soup-4.readthedocs.io) (28/10/2023)

[3] Requests. Link: Requests: HTTP for Humans™ — Requests 2.31.0 documentation (28/10/2023)

[4] Scikit - learn. Link: User guide: contents — scikit-learn 1.3.2 documentation (10/12/2023)

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Trương Khánh Long	Tiền xử lý dữ liệu + Thăm dò dữ liệu
2	Nguyễn Chí Thi	Thu thập dữ liệu + Phát triển mô hình
3	Bùi Lê Trọng Đức	Thăm dò dữ liệu + Phát triển mô hình
4	Hoàng Hải Anh	Trực quan hóa dữ liệu + Mô tả bộ dữ liệu
5	Nguyễn Tiến Thịnh	Trực quan hóa dữ liệu + Mô tả bộ dữ liệu