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

DỰ ĐOÁN GIÁ XE CŨ ĐƯỢC THU THẬP TỪ TRANG WEBSITE CHỢ TỐT

Sinh viên thực hiện			
STT	Họ tên	MSSV	Ngành
1	Lê Hoàng Huy	20521392	CNTT
2	Cao Hải Hà	20521267	CNTT
3	Nguyễn Huy Hoàng	20521343	CNTT

TP. HỒ CHÍ MINH – 12/2022

GIỚI THIỆU

Đề tài này nhằm mục đích thu thập và phân tích các dữ liệu liên quan đến giá xe máy cũ được niêm yết trên trang web Chợ Tốt (https://xe.chotot.com/mua-ban-xe-may). Mục tiêu của dề tài là dự đoán giá xe máy cũ trên chợ tốt có độ chính xác cao.

Để thực hiện đề tài này, sử dụng các công cụ và kỹ thuật phân tích dữ liệu như web scraping để thu thập dữ liệu từ trang web, xử lý và làm sạch dữ liệu sử dụng ngôn ngữ lập trình và thư viện phân tích dữ liệu như Python và Pandas, và áp dụng các thuật toán máy học để xây dựng mô hình dự đoán giá.

Kết quả đạt được là bộ dữ liệu gồm 190 dòng bao gồm các thông tin liên quan đến thị trường xe máy cũ được niêm yết trên Chợ Tốt

Bộ dữ liệu phân tích được tự thu thập từ trang web Chợ Tốt và không phụ thuộc vào bộ dữ liệu nào khác. Ngoài ra, nhóm còn tham khảo và áp dụng các dự án mẫu và tài liệu tham khảo liên quan đến phân tích dữ liệu và máy học để hoàn thiện đề tài này.

Nhóm tự lên ý tưởng là thực hiện thu thập dữ liệu trên Chợ Tốt bằng các công cụ hỗ trợ như selenium và không sử dụng hoặc tổng hợp từ các dữ liệu có sẵn hoặc dựa trên đề tài khác

MÔ TẢ BỘ DỮ LỆU

Dữ liệu của đồ án là một tập dữ liệu thu thập về các xe máy cũ được rao bán trên trang Chợ Tốt. Dữ liệu được thu thập vào ngày 09 tháng 12 năm 2023, tại thành phố Hồ Chí Minh.Các thông tin được thu thập bao gồm:190 bản ghi, mỗi bản ghi đại diện cho một chiếc xe máy cũ. Các thông tin được thu thập bao gồm:

Cột	Ý nghĩa	Kiểu dữ liệu
Tên đăng bán	Tên của chiếc xe máy được đăng bán, thường bao gồm tên hãng sản xuất, tên loại xe và tên dòng xe.	Object
Giá bán	Giá bán của chiếc xe máy được đăng bán, được tính bằng đồng Việt Nam (VND).	Float 64
Hãng sản xuất	Hãng sản xuất của chiếc xe máy.	Object
Năm đăng kí	Năm sản xuất của chiếc xe máy.	Int 64
Số km đã đi	Số km đã đi của chiếc xe máy.	Float 64
Tỉnh thành	Tỉnh thành nơi chiếc xe máy được đăng bán.	Object
Loại xe	Loại xe của chiếc xe máy, bao gồm xe số, xe tay ga và xe côn tay.	Object
Dung tích xe	Dung tích động cơ của chiếc xe máy, được tính bằng phân khối (cc).	Object

Xuất xứ	Xuất xứ của chiếc xe máy, bao gồm Việt Nam, Nhật Bản, Thái Lan và Trung Quốc.	Object
Chính sách bảo hành	Chính sách bảo hành của chiếc xe máy, được tính bằng tháng.	Object
Trọng lượng xe	Trọng lượng của chiếc xe máy, được tính bằng kg.	Object

Thống kê	Giá trị
Số cột	12
Số dòng	32
Biến phân loại	9
Biến số	3
Số lượng khuyết	55
Giá bán trung bình	28.800.000 VND
Giá bán cao nhất	388.000.000 VND
Giá bán thấp nhất	5.500.000 VND
Năm sản xuất trung bình	2018
Số km đã đi trung bình	9.000 km
Trọng lượng xe trung bình	100 kg

Phương pháp thu thập dữ liệu

Chúng tôi đã tiến hành thu thập dữ liệu về giá xe máy cũ từ trang web chotot.vn. Để thực hiện việc này, chúng tôi đã sử dụng Selenium, một công cụ mạnh mẽ cho việc tự động hóa trình duyệt web.

Selenium cho phép chúng tôi tự động điều hướng trang web, tương tác với các thành phần trên trang và thu thập dữ liệu mà chúng tôi cần. Chúng tôi đã chạy mã Selenium trên Google Colab, một môi trường Jupyter notebook đám mây miễn phí được cung cấp bởi Google.

PHƯƠNG PHÁP PHÂN TÍCH

Hình 1. Quy trình PTDL.

3.1 Xác định vấn đề nghiên cứu:

Vấn đề nghiên cứu chính của chúng tôi tập trung vào việc phân tích giá xe máy cũ trên trang web chotot.vn. Chúng tôi đã tiến hành thu thập dữ liệu từ trang web này và sử dụng các phương pháp phân tích dữ liệu để hiểu rõ hơn về các yếu tố ảnh hưởng đến giá xe máy cũ.

3.2 Thu thập dữ liệu

Quy trình thu thập dữ liệu của chúng tôi bao gồm các bước sau:

1. Điều hướng đến trang web chotot.vn.
2. Sử dụng Selenium để tìm kiếm thông tin về xe máy cũ.
3. Thu thập dữ liệu từ các kết quả tìm kiếm, bao gồm thông tin về giá cả, hãng xe, dòng xe, và các thông tin khác.
4. Lưu trữ dữ liệu thu thập được vào một DataFrame pandas để tiện cho việc phân tích sau này.

3.3 Xử lý dữ liệu

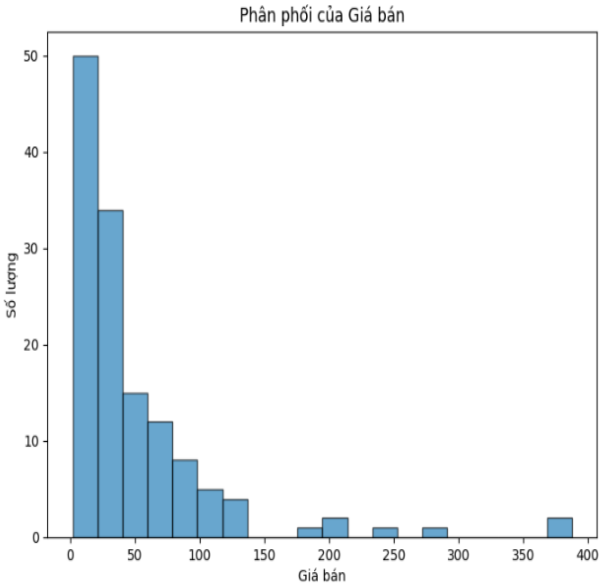
Sau khi thu thập dữ liệu từ trang web chotot.vn, chúng tôi đã tiến hành một số bước xử lý dữ liệu để chuẩn bị cho việc phân tích. Các bước xử lý dữ liệu bao gồm:

1. Chuẩn hóa tên cột: Chúng tôi đã đảm bảo rằng tất cả các tên cột đều tuân theo một định dạng nhất quán, giúp dễ dàng thao tác với dữ liệu hơn.
2. Xử lý giá trị bị khuyết: Chúng tôi đã sử dụng phương pháp dropna để loại bỏ các hàng có giá trị bị khuyết. Điều này giúp đảm bảo rằng mô hình của chúng tôi không bị ảnh hưởng bởi các giá trị không xác định.
3. Định dạng lại kiểu dữ liệu: Chúng tôi đã chuyển đổi các cột dữ liệu thành kiểu dữ liệu phù hợp, giúp cho việc phân tích và mô hình hóa dữ liệu được chính xác hơn.
4. Chuẩn hóa cột giá bán: Chúng tôi đã chuẩn hóa cột giá bán để đảm bảo rằng tất cả các giá trị đều nằm trong cùng một phạm vi, giúp cho việc huấn luyện mô hình hồi quy tuyến tính trở nên dễ dàng hơn.

Qua quá trình xử lý dữ liệu này, chúng tôi đã tạo ra một bộ dữ liệu sạch, chuẩn và sẵn sàng để được sử dụng trong các phân tích và mô hình hóa dữ liệu tiếp theo.

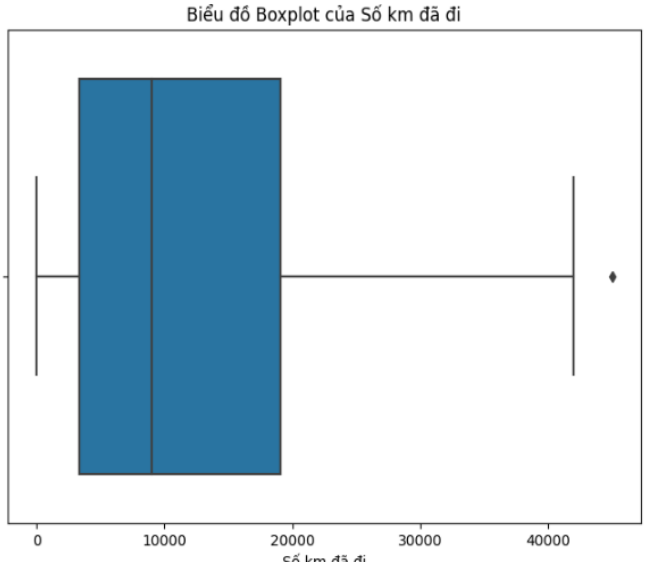
PHÂN TÍCH THĂM DÒ/SƠ BỘ

Tập trình bày các phát hiện chính, không bắt buộc trình bày hết kết quả phân tích thăm dò các biến. Trình bày các biến quan trọng đã chọn lọc lại.



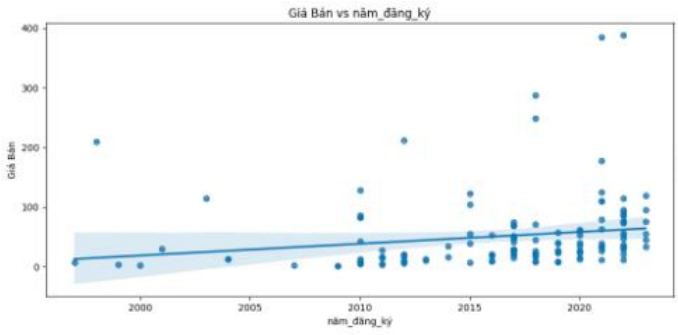
Nhận xét:

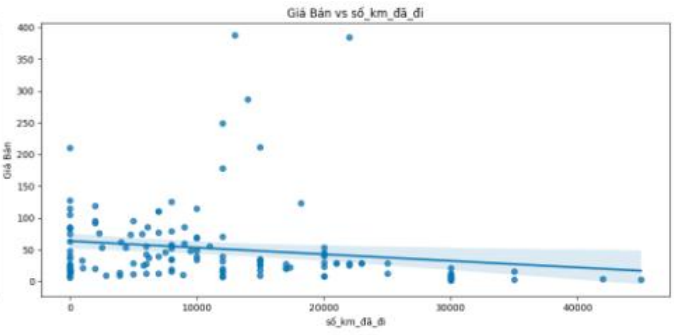
- Giá bán của xe máy cũ ở Việt Nam chủ yếu tập trung trong khoảng từ hơn 0 triệu đồng đến khoảng 100 triệu đồng.
- Số lượng xe giảm dần khi giá tăng dần.



Nhận xét:

- Hầu hết các xe máy cũ ở Việt Nam đã đi được dưới 10.000 km.
- Số km đã đi trung bình của xe máy cũ ở Việt Nam là 5.000 km.
- Có một số xe máy cũ đã đi được hơn 50.000 km.





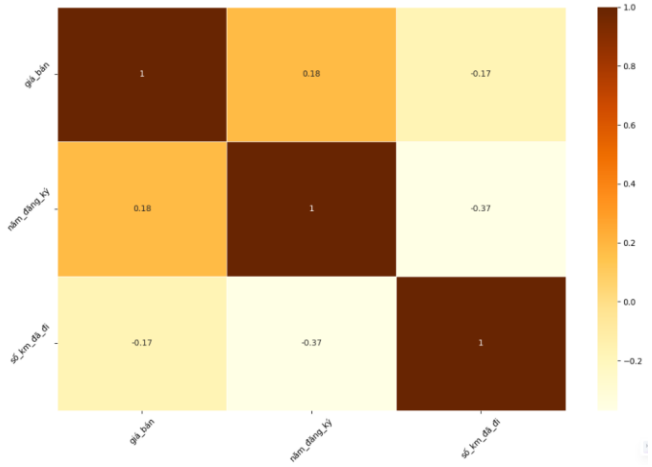
-Biểu đồ 1: Giá Bán vs Năm Đăng Ký

-Biểu đồ 2: Giá Bán vs Số Km Đã Đi

- Giá bán xe thường giảm dần theo năm đăng ký và số km đã đi.
- Xe càng mới và chạy ít thì giá bán càng cao.

-Nhận xét chung

Cả hai đồ thị đều cho thấy mối quan hệ tiêu cực giữa giá bán và năm đăng ký, số km đã đi. Điều này có nghĩa là giá bán xe thường giảm dần theo năm đăng ký và số km đã đi.

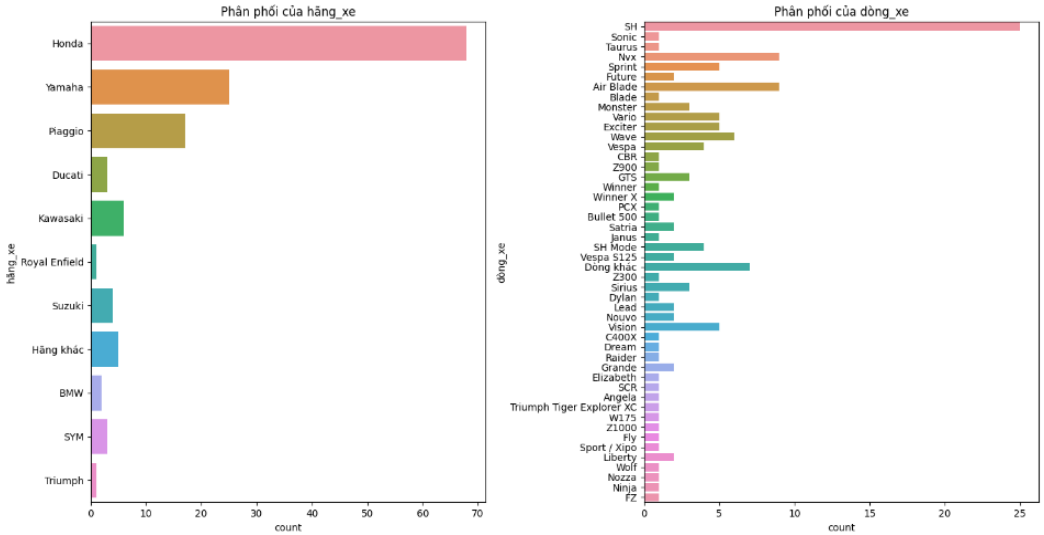


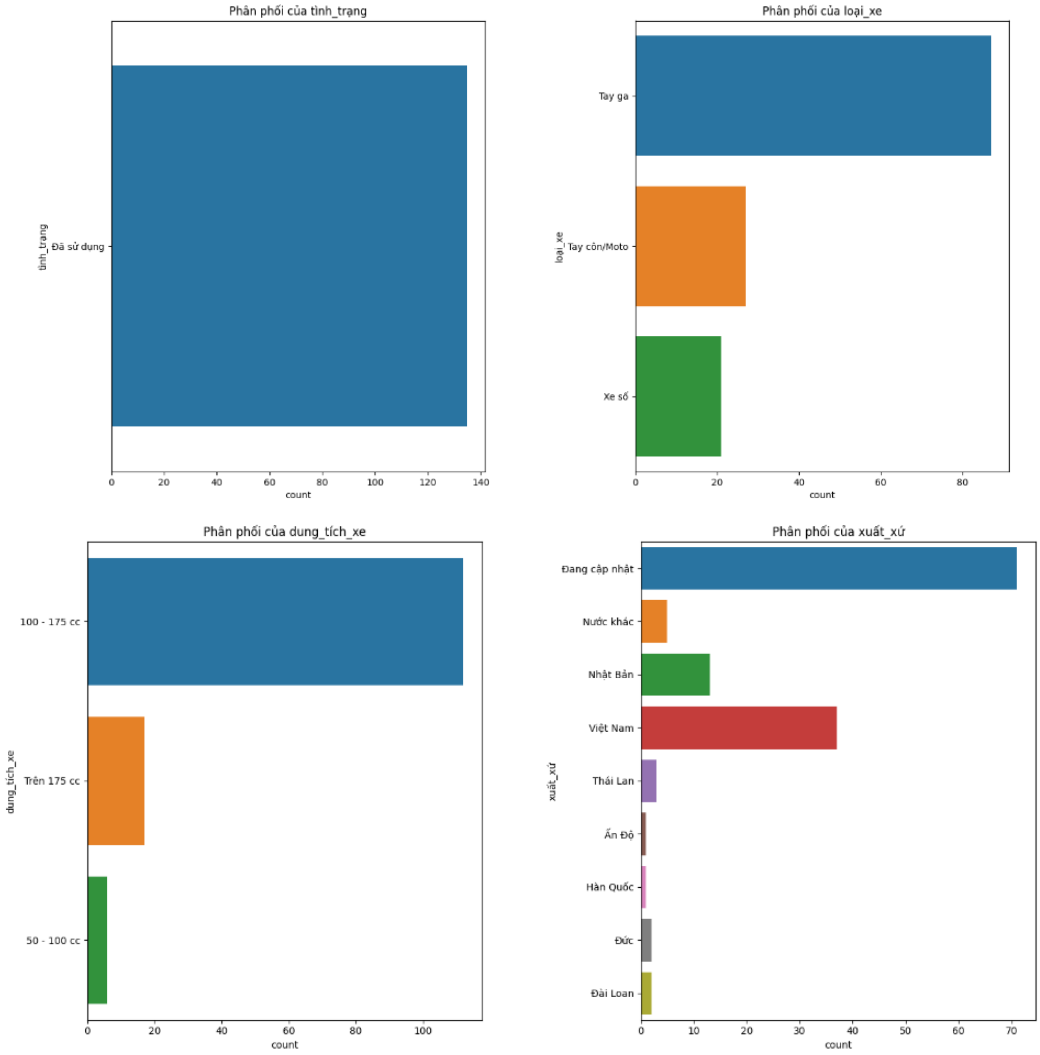
Nhận xét: **giá_bán và số_km_đã_đi:**

- Mối tương quan giữa hai biến số này là âm. Điều này có nghĩa là khi số km đã đi của ô tô tăng lên, giá bán của ô tô có xu hướng giảm xuống.
- Mức độ tương quan giữa hai biến số này là yếu. Giá trị tương quan Pearson là -0.17.

Phân tích đơn biến

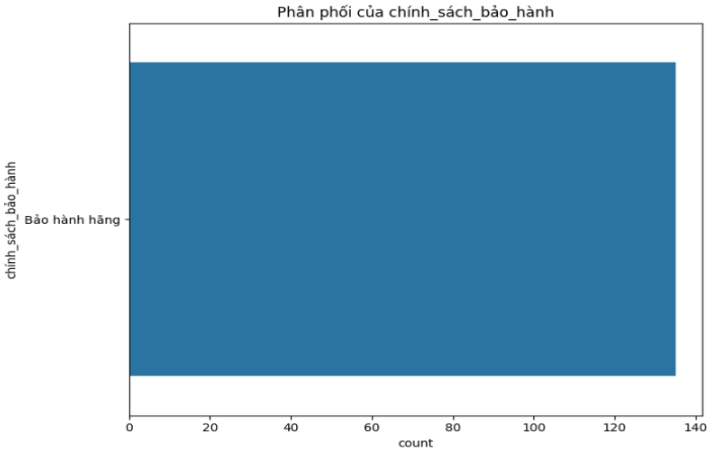
Biểu đồ cột dữ liệu của từng biến phân loại.





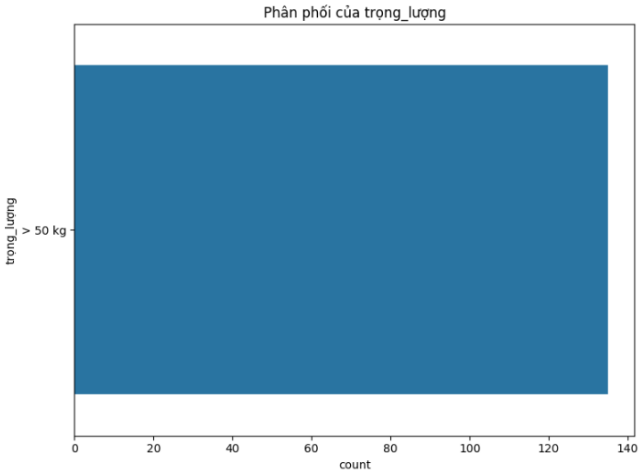
Nhận xét:

- Phân phối của hãng xe:Biểu đồ đầu tiên cho thấy Honda và Yamaha là hai hãng xe phổ biến nhất trong dữ liệu, với số lượng xe Honda cao hơn đáng kể so với các hãng khác.
- Phân phối của dòng xe: Sh, Nvx và Air Blade là ba dòng xe phổ biến nhất. Có sự đa dạng trong các dòng xe khác nhưng với số lượng ít hơn.
- Phân phối của tình trạng : Biểu đồ cho thấy toàn bộ là xe cũ đã qua sử dụng.
- Phân phối của loại xe : Biểu đồ cho thấy xe tay ga có số lượng nhiều nhất.
- Phân phối của dung tích xe : Nhiều nhất là các xe có dung tích từ 100-175cc
- Phân phối của xuất xứ : Xe chưa được cập nhật xuất xứ và Việt Nam chiếm số lượng lớn trong dữ liệu, với xe chưa được cập nhật xuất xứ là cao nhất.



Biểu đồ chính sách bảo hành:

Nhận xét : Biểu đồ cho thấy toàn bộ đều được chính sách bảo hành hãng.

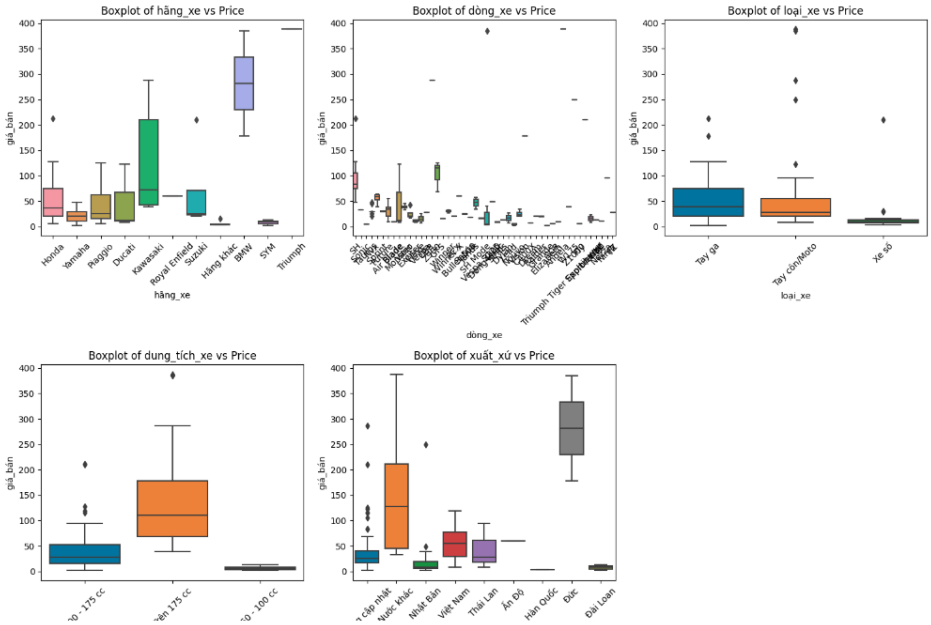


Biểu đồ trọng lượng :

Nhận xét : Biểu đồ cho thấy toàn bộ đều được trọng lượng.

Phân tích đa biến

Biểu đồ Boxplot cho mỗi biến phân loại so với giá xe:

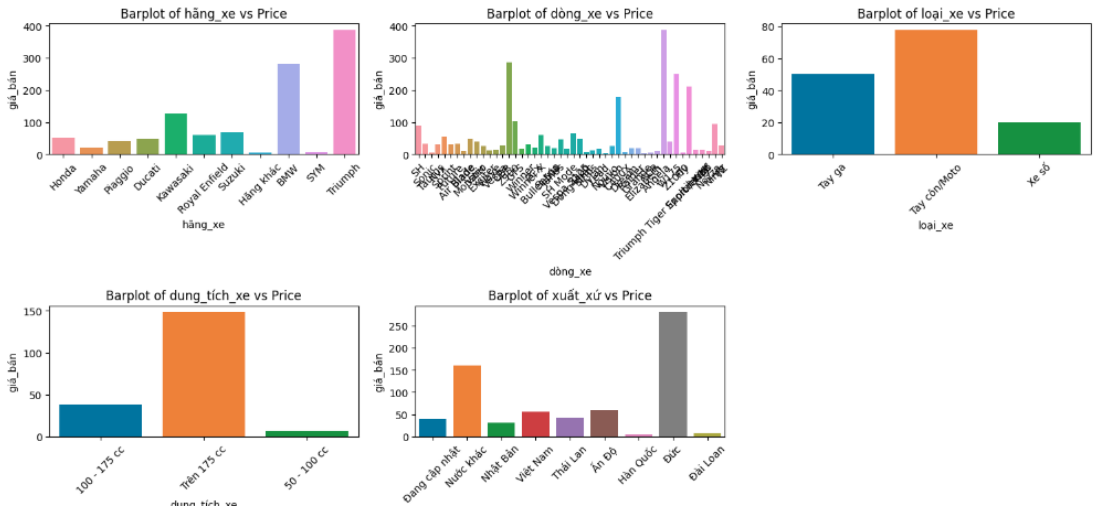


Nhận xét :

- Boxplot của hàng_xe so với Price: Giá cả có sự biến thiên lớn giữa các loại hàng_xe khác nhau. Điều này cho thấy rằng hãng xe có thể là một yếu tố quan trọng ảnh hưởng đến giá cả.
- Boxplot của dòng_xe so với Price: Một số dòng xe có phạm vi giá rộng, trong khi một số khác lại tập trung ở mức giá thấp. Điều này cho thấy rằng dòng xe cũng có thể ảnh hưởng đến giá cả.
- Boxplot của loại_xe so với Price: Loại xe 'loai_3' có phạm vi giá cả rộng nhất và median cao nhất. Điều này cho thấy rằng loại xe có thể ảnh hưởng đến giá cả.
- Boxplot của dung_tich_xe so với Price: Dung tích xe từ 175-200cc có median và phạm vi giá cao nhất. Điều này cho thấy rằng dung tích xe có thể ảnh hưởng đến giá cả.
- Boxplot của xuất_xứ so với Price: Xe nhập khẩu từ Mỹ và Đức có median và phạm vi giá cao hơn so với các nước khác. Điều này cho thấy rằng xuất xứ của xe có thể ảnh hưởng đến giá cả.

Nhìn chung, các biểu đồ boxplot cho thấy rằng các yếu tố như hãng xe, dòng xe, loại xe, dung tích xe và xuất xứ xe đều có thể ảnh hưởng đến giá cả của xe.

Biểu đồ Barplot cho mỗi biến phân loại so với giá xe:



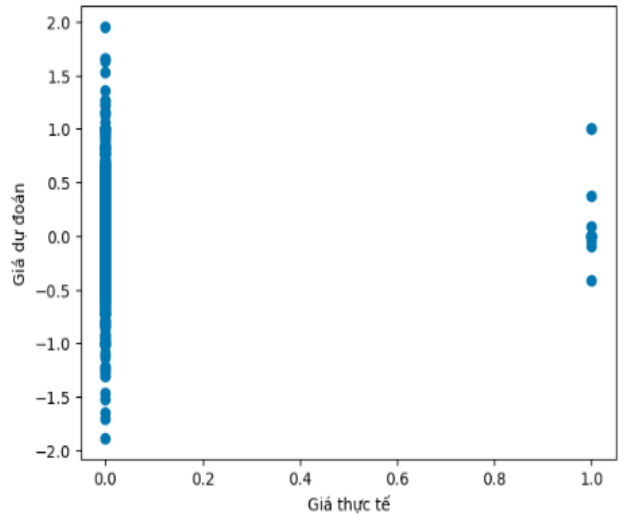
Nhận xét

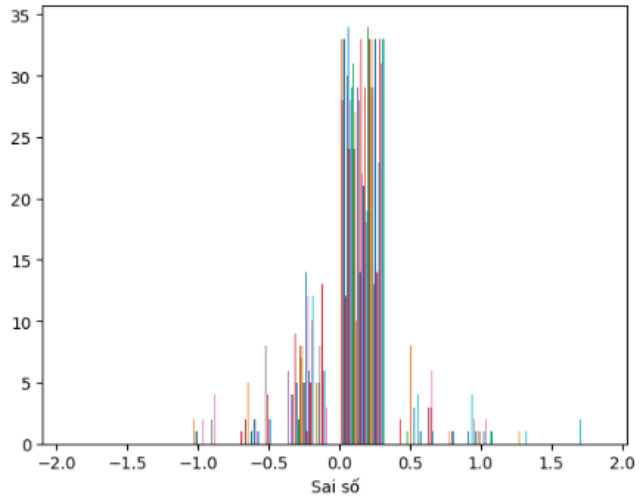
Các biến phân loại có ảnh hưởng đáng kể đối với giá xe là:

- hãng_xe
- dòng_xe
- loại_xe
- dung_tích_xe
- xuất_xứ

KẾT QUẢ PHÂN TÍCH

Sau khi xây dựng mô hình hồi quy truyến tính thì chúng tôi thu được các kết quả như sau:





MSE: 0.06339497461490633

R2: -0.14464251732828098

Nhận xét :

Dựa trên biểu đồ giá thực tế và giá dự đoán:

- Phân bố dữ liệu: Có hai cụm dữ liệu rõ ràng trong biểu đồ. Một cụm là một đường thẳng dọc tại giá trị x xấp xỉ 0, cho thấy rằng có nhiều giá dự đoán khác nhau tương ứng với giá thực tế khoảng 0. Cụm thứ hai bao gồm một số điểm phân tán xung quanh giá trị 1 trên trục x, cho thấy một số biến động trong giá dự đoán cho giá thực tế khoảng 1.
- Không có mối tương quan rõ ràng: Không có một mối tương quan hoặc mô hình rõ ràng giữa giá dự đoán và giá thực tế như được chỉ ra bởi các cụm này. Điều này có thể cho thấy rằng mô hình dự đoán của bạn không đang mô phỏng chính xác mối tương quan giữa các biến đầu vào và giá thực tế.
- Giá dự đoán chủ yếu tập trung vào một điểm cụ thể: Trong khi giá thực tế có sự phân tán rộng lớn, giá dự đoán chủ yếu tập trung vào một điểm cụ thể. Điều này cho thấy mô hình dự đoán có thể không chính xác hoặc không đáng tin cậy trong việc tái hiện sự biến thiên của giá thực tế.

Dựa trên kết quả MSE và R2 ở trên, cùng với biểu đồ sai số:

- Giá trị MSE (Mean Squared Error) là 0.0634: Đây là một giá trị sai số trung bình khá thấp, cho thấy mô hình của bạn có độ chính xác tương đối. Tuy nhiên, giá trị này vẫn cho thấy có sự sai lệch giữa giá trị dự đoán và giá trị thực tế.

- Giá trị R2 là -0.1446: Giá trị R2 âm cho thấy mô hình của bạn không phù hợp với dữ liệu. Một mô hình tốt sẽ có giá trị R2 gần với 1. Trong trường hợp này, mô hình của bạn đang hoạt động kém hơn so với một đường ngang biểu diễn trung bình của tất cả các điểm dữ liệu.
- Biểu đồ sai số: Biểu đồ cho thấy có một số lượng lớn dự đoán chính xác hoặc có sai số rất nhỏ (như được thể hiện bởi cột cao ở mức sai số 0). Tuy nhiên, cũng có một số dự đoán có sai số lớn hơn, như được thể hiện bởi các cột thấp hơn ở các mức sai số khác nhau.

CHỈNH SỬA SAU BÁO CÁO

KẾT LUẬN

Quá trình phân tích dữ liệu và xây dựng mô hình dự đoán giá xe đã được thực hiện theo các bước sau: Làm sạch dữ liệu: Xác định và xử lý các giá trị null, giá trị bất thường. Sau đó , phân tích thăm dò/sơ bộ: Tìm hiểu các đặc điểm của dữ liệu, bao gồm phân phối, mối quan hệ giữa các biến. Tiếp đến là xây dựng mô hình: Sử dụng mô hình hồi quy tuyến tính để dự đoán giá xe. Cuối cùng , đánh giá mô hình: Sử dụng các chỉ số như sai số bình phương trung bình (MSE) và hệ số xác định (R2) để đánh giá hiệu suất của mô hình.

Tóm lại dựa trên kết quả phân tích dữ liệu, có thể thấy dataset mà nhóm chọn để thực hiện phục vụ cho đồ án cuối kỳ có bộ dữ liệu có chất lượng ổn, các giá trị null hoặc giá trị bất thường ở mức trung bình trên có thể thực hiện được , các biến số có phân phối tương đối bình thường.

Kết quả xây dựng mô hình dự đoán LinearRegression với hệ số xác định R2 mang giá trị âm và gần với 1. Trong trường hợp này, mô hình đang hoạt động kém hơn so với một đường ngang biểu diễn trung bình của tất cả các điểm dữ liệu.

TÀI LIỆU THAM KHẢO

https://viblo.asia/p/a-tutorial-on-eda-and-feature-engineering-WAyK8drmKxX (5/12/2023)

https://csc.edu.vn/tin-tuc/san-pham-mon-hoc/Data-Analysis-Quy-trinh-Phan-tich-du-lieu-tu-A-Z-8264 (7/12 /2023)

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Cao Hải Hà	• Lên ý tưởng , thu thập dữ liệu và chọn lọc bộ dataset. • Thực hiện xây dựng code python trên Colab phần " Làm sạch dữ liệu" và "Phân tích thăm dò/sơ bộ". • Hổ trợ xây dựng , sửa chửa và tổng kết word. • Hỗ trợ xây dựng PowerPoint.
2	Lê Hoàng Huy	• Thực hiện xây dựng code python trên Colab phần "Phân tích thăm dò/sơ bộ" và "Xây dựng mô hình". • Thực thiện xây dựng PowerPoint của nhóm. • Hổ trợ xây dựng Word.
3	Nguyễn Huy Hoàng	• Thực hiện code python trên Colab phần "Xây dựng mô hình" và " Đánh giá mô hình". • Thực hiện xây dựng Word của nhóm. • Hỗ trợ xây dựng PowerPoint.