



TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA CÔNG NGHỆ THÔNG TIN



TIÊU LUẬN CUỐI KỲ
HỌC PHẦN: KHOA HỌC DỮ LIỆU

DỰ ĐOÁN GIÁ VÀ PHÂN CỤM XE Ô TÔ CŨ

Nhóm	3
Họ Và Tên Sinh Viên	Lớp Học Phân
Phan Thanh Tâm	
Võ Đình Huy	21.10
Huỳnh Văn Lộc	

ĐÀ NẴNG, 05/2024

TÓM TẮT

Ngành công nghiệp xe ô cũ luôn sôi động và việc định giá chính xác cho xe cũ là một thách thức. Khoa học dữ liệu có thể đóng vai trò quan trọng trong việc giải quyết vấn đề này bằng cách dự đoán giá xe và phân loại chúng thành các nhóm có đặc điểm tương đồng. Đề tài này giới thiệu ứng dụng của khoa học dữ liệu trong việc dự đoán giá và phân cụm xe ô cũ dựa trên các thông số kỹ thuật, tình trạng xe và các yếu tố khác.

- Vấn đề cần giải quyết:
 - Xác định các yếu tố ảnh hưởng đến giá xe ô tô cũ.
 - Dự đoán giá bán xe ô tô cũ một cách chính xác.
 - Phân chia xe ô tô cũ thành các nhóm có đặc điểm tương đồng.
- Phương pháp giải quyết:
 - Thu thập dữ liệu về xe ô tô cũ bao gồm thông số kỹ thuật, giá bán, hãng xe, năm sản xuất,...
 - Sử dụng các thuật toán học máy như hồi quy tuyến tính, cây quyết định để dự đoán giá xe.
 - Áp dụng các phương pháp phân cụm như K-means, DBSCAN để phân chia xe ô tô cũ thành các nhóm.
- Kết quả đạt được:
 - Xây dựng mô hình dự đoán giá xe ô tô cũ với độ chính xác cao.
 - Phân chia xe ô tô cũ thành các nhóm có đặc điểm rõ ràng, giúp người mua dễ dàng lựa chọn xe phù hợp.
 - Đề xuất một số giải pháp để cải thiện hiệu quả mô hình dự đoán và phân cụm.

MỤC LỤC

DỰ ĐOÁN GIÁ VÀ PHÂN CỤM XE Ô TÔ CŨ	1
1. Giới thiệu	6
2. Thu thập và mô tả dữ liệu.....	7
2.1. Thu thập dữ liệu	7
2.2. Mô tả và trực quan hóa dữ liệu	9
3. Trích xuất đặc trưng	11
3.1. Làm sạch, chuẩn hóa dữ liệu.	11
3.1.2 Xử lý dữ liệu:.....	11
3.1.2 Giảm chiều dữ liệu.	11
3.2. Lựa chọn đặc trưng.	12
3.3. Trực quan hóa.	12
4. Mô hình hóa dữ liệu.....	15
4.1. Dự đoán giá xe	15
4.1.1 Các thông số đánh giá mô hình	15
4.1.2 Phân chia dữ liệu	16
4.1.3 Linear Regression.....	16
4.1.4 Gradient Boosting Regression.....	18
4.1.5 So sánh 2 mô hình	19
4.2. Phân cụm.....	20
4.2.1 Mô hình K-means.	20
4.2.1 Mô hình GMM	23
5. Kết luận.....	27
5.1. Kết luận:.....	27
5.1.1 Bài toán dự đoán:.....	27
5.1.2 Bài toán phân cụm:	27
5.2. Hướng phát triển:	27
6. Tài liệu tham khảo	29

BẢNG PHÂN CÔNG NHIỆM VỤ

Sinh viên thực hiện	Các nhiệm vụ	Tự đánh giá theo 3 mức (Đã hoàn thành/Chưa hoàn thành/Không triển khai)
Phan Thanh Tâm	Thu thập dữ liệu Trích xuất đặc trưng Mô hình hoá dữ liệu: <ul style="list-style-type: none"> ○ Linear Regresstion ○ Gradient Boosting Nhận xét biểu đồ Viết báo cáo, làm slide	Đã hoàn thành
Huỳnh Văn Lộc	Xử lí dữ liệu Trích xuất đặc trưng Mô hình hoá dữ liệu: <ul style="list-style-type: none"> ○ GMM (Gaussian Mixture Model) ○ K-means Trực quan hoá dữ liệu Nhận xét biểu đồ Viết báo cáo, làm slide	Đã hoàn thành
Võ Đình Huy	Thu thập dữ liệu Mô hình hoá dữ liệu: <ul style="list-style-type: none"> ○ K-means ○ GMM (Gaussian Mixture Model) Trực quan hoá dữ liệu Nhận xét biểu đồ Viết báo cáo, làm slide	Đã hoàn thành

DANH MỤC HÌNH VẼ

H.1 Sơ đồ khái niệm dự đoán.....	6
H.2 Sơ đồ khái niệm bài toán phân cụm	6
H.3 Ảnh phân tích html của trang web.....	7
H.5. Dữ liệu detail của xe	8
H.6 Phân tích html của trang detail	8
H.7 Dữ liệu đầy đủ của xe sau khi cào	9
H.8. Các đặc trưng của dữ liệu	9
H.9. Số mẫu trống của dữ liệu	9
H.10 Biểu đồ số lượng của giá	9
H.11 Đồ thị số lượng của 10 Hàng xe nổi bật của tập train và test.....	10
H.12 Đồ thị phân bố số lượng theo giá của data clean.....	10
H.13 Biểu đồ giá xe cũ trung bình 5 hàng phố biển Số Km đã đi.....	11
H.14 So sánh sự phân bố của dữ liệu của 2 phương pháp PCA và T-SNE.....	12
H.15 Biểu đồ phân bố giá theo năm sản xuất của Số tự động và Số tay	13
H.16 Biểu đồ giá xe cũ trung bình theo số Km đã đi	13
H.17 Biểu đồ giá xe trung bình theo hàng xe	14
H.18 Biểu đồ giá trung bình theo động cơ và xuất xứ.....	14
H.19 Biểu đồ giá trung bình theo kiểu dáng và hàng xe	15
H.20 Biểu đồ phân tán giữa giá thực tế và giá dự đoán	17
H.21 Biểu đồ giá dự đoán với giá thực tế.....	18
H.22 Biểu đồ phân tán giữa giá thực tế và giá dự đoán	19
H.23 Biểu đồ phân tán giữa giá dự đoán và giá thực tế	19

1. Giới thiệu

Trong phần này, sinh viên giới thiệu các bài toán cần giải quyết và đề xuất giải pháp tổng quan về dữ liệu dưới dạng sơ đồ khối (pipeline). Pipeline này có thể dùng chung cho cả 2 bài toán mà nhóm giải quyết.

Ngành công nghiệp xe ô tô cũ luôn là một lĩnh vực sôi động và đầy thách thức. Việc định giá chính xác cho các mẫu xe cũ là một nhiệm vụ phức tạp, do sự đa dạng về độ tuổi, tình trạng sử dụng và các đặc điểm kỹ thuật của từng chiếc xe. Đề tài này giúp cho người mua lần người bán có 1 quyết định sáng suốt khi mua bán xe (người mua thì sẽ mua được xe với giá hợp lý tránh bị hớ, còn người bán thì giúp xác định một cách khách quan và chính xác giá trị thị trường của chiếc xe cũ của mình sẽ giúp họ đưa ra mức giá hợp lý và cạnh tranh).

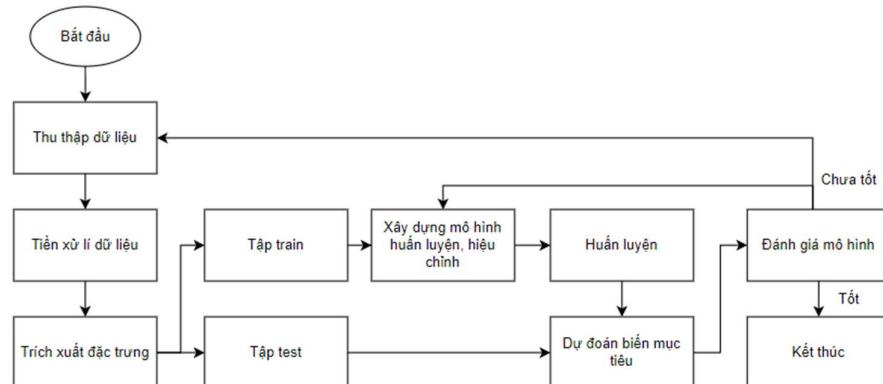
- a. Các bài toán cần giải quyết:

Dự đoán giá xe dựa trên các thông số của xe như: hãng xe, tình trạng, số km đã đi, xuất xứ, ...

Phân cụm xe trên tất cả các đặc trưng của xe từ đó rút ra ...

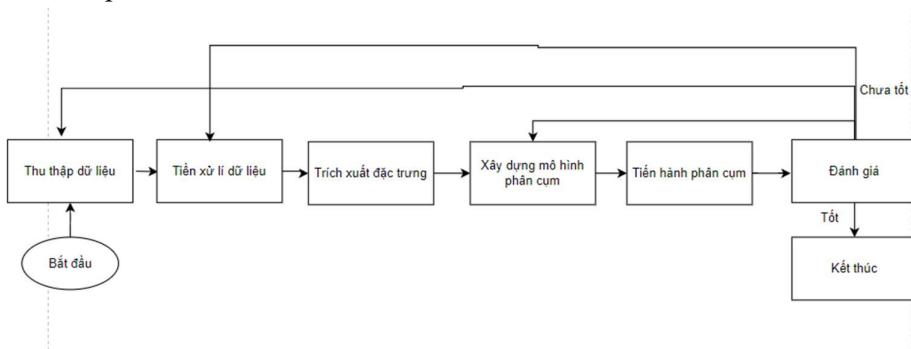
- b. Giải pháp tổng quan

- Bài toán dự đoán:



H.11 Sơ đồ khái niệm bài toán dự đoán

- Bài toán phân cụm:



H.22 Sơ đồ khái niệm bài toán phân cụm

2. Thu thập và mô tả dữ liệu

2.1. Thu thập dữ liệu

SV mô tả mô tả giải pháp thu thập dữ liệu gồm nguồn dữ liệu, công cụ thu thập, cách thức sử dụng công cụ, đầu vào và đầu ra của quá trình thu thập, và cho ví dụ minh họa.

- Nguồn thu thập dữ liệu được lấy từ website: <https://bonbanh.com>

- Công cụ thu thập: **request** và sử dụng thư viện **Beautiful Soup** để trích lọc các thông tin cần thiết từ 1 trang HTML.

- Các thức thu thập:

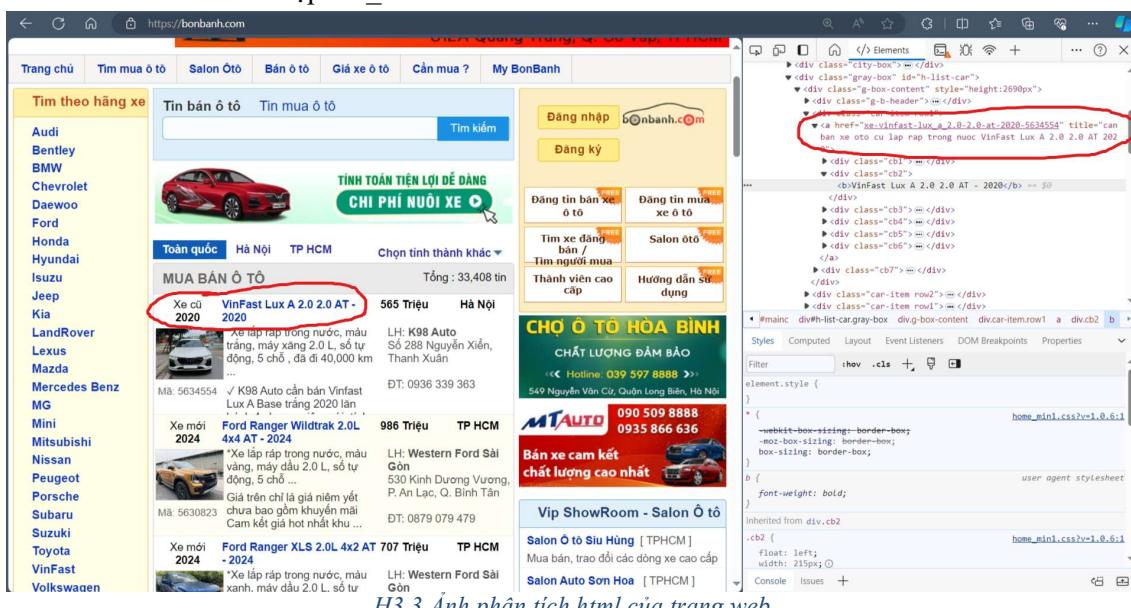
+ Bước 1: thu thập car_id của mỗi trang (mỗi trang gồm 20 car_id) bằng cách dùng Beautiful Soup để trích xuất thông tin trong thẻ **<a>** và lưu lại vào file car_ids.txt.

+ Bước 2: dùng **request** của Python để trang web trên cung cấp trang HTML của xe theo mã id vừa lưu.

+ Bước 3: sử dụng **Beautiful Soup** để trích lọc từ các thẻ **<div>**, ****,... mà có thông tin cần thiết của xe và sau đó lưu vào file raw_data.csv.

- Ví dụ:

+ Bước 1: thu thập car_id và lưu vào file.



H3.3 Ảnh phân tích html của trang web

```

car_ids.txt
1 xe-suzuki-ertiga-hybrid-1.5-mt-2023-5527171
2 xe-suzuki-swift-glx-1.2-at-2023-5522881
3 xe-toyota-vios-e-1.5-mt-2024-5461269
4 xe-ford-everest-ambiente-2.0l-4x2-at-2022-5541828
5 xe-mitsubishi-xpander-premium-1.5-at-2024-5494445
6 xe-mercedes_benz-glb-200-amg-2024-4925370
7 xe-mercedes_benz-e_class-e300-amg-2023-4934866
8 xe-mercedes_benz-v_class-v250-amg-2023-5398549

```

H.4 Dữ liệu car id được cào về

H.54. Dữ liệu detail của xe

+ Bước 3: trích lọc các thông tin của xe và lưu vào file csv.

H.5 Phân tích html của trang detail

H6.7 Dữ liệu đầy đủ của xe sau khi cào

-Đầu vào: trang đầu và trang cuối của web (gồm khoảng 1500 trang).

-Đầu ra: 30 ngàn mẫu dữ liệu thô của xe từ trang web.

2.2. Mô tả và trực quan hóa dữ liệu

- Kích thước tập dữ liệu: 30 nghìn mẫu
 - Số đặc trưng của một mẫu dữ liệu: 14 đặc trưng.

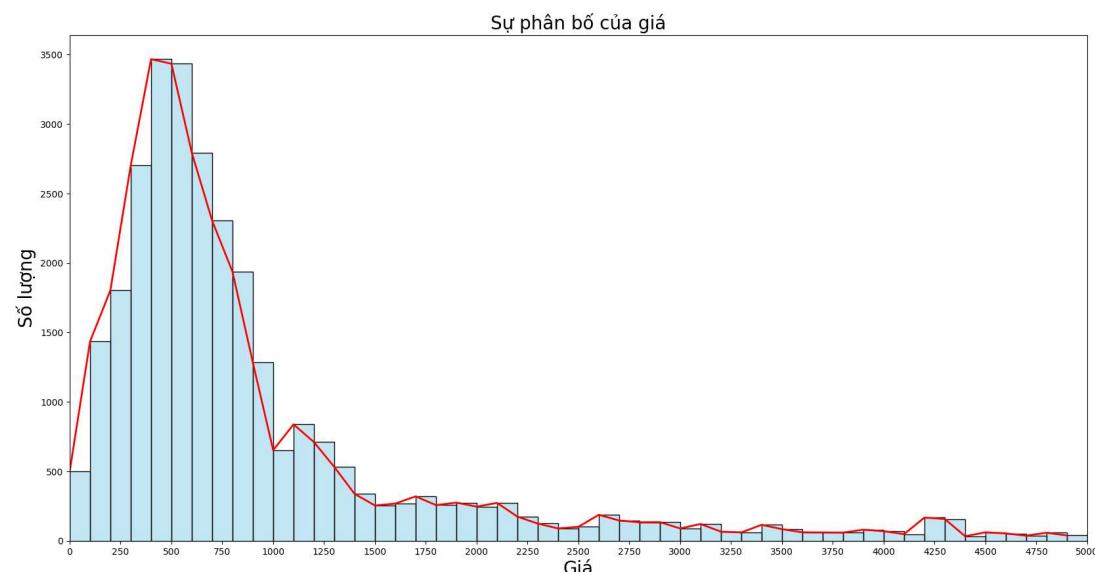
Đặc trưng	ID	Hãng xe	Năm sản xuất	Tình trạng	Số Km đã đi	Xuất xứ	Kiểu dáng	Hộp số	Động cơ	Số chỗ ngồi	Dẫn động	Ngày đăng	Địa điểm	Giá
Kiểu dữ liệu	string	string	int	string	string	string	string	string	int	string	string	string	string	float

H.7. Các đặc trưng của dữ liệu

- Số mẫu trống của dữ liệu:

H.8. Số mẫu trống của dữ liệu

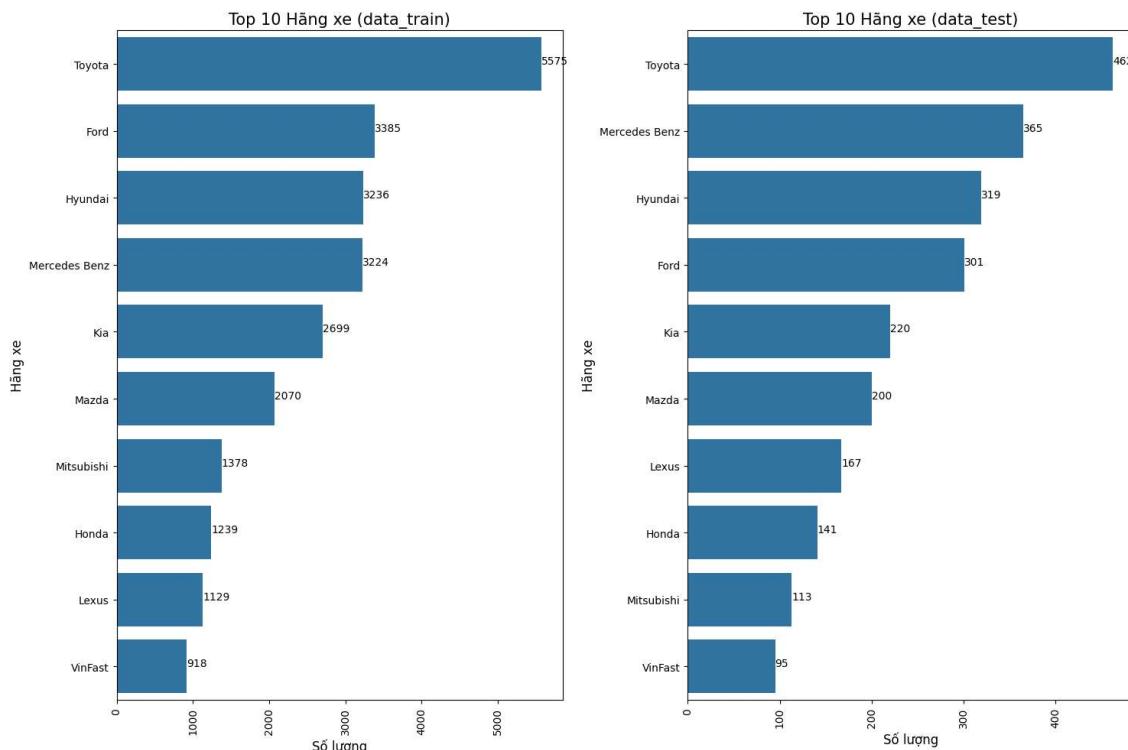
- Các thống kê mô tả về tập dữ liệu



H9.10 Biểu đồ số lượng của giá

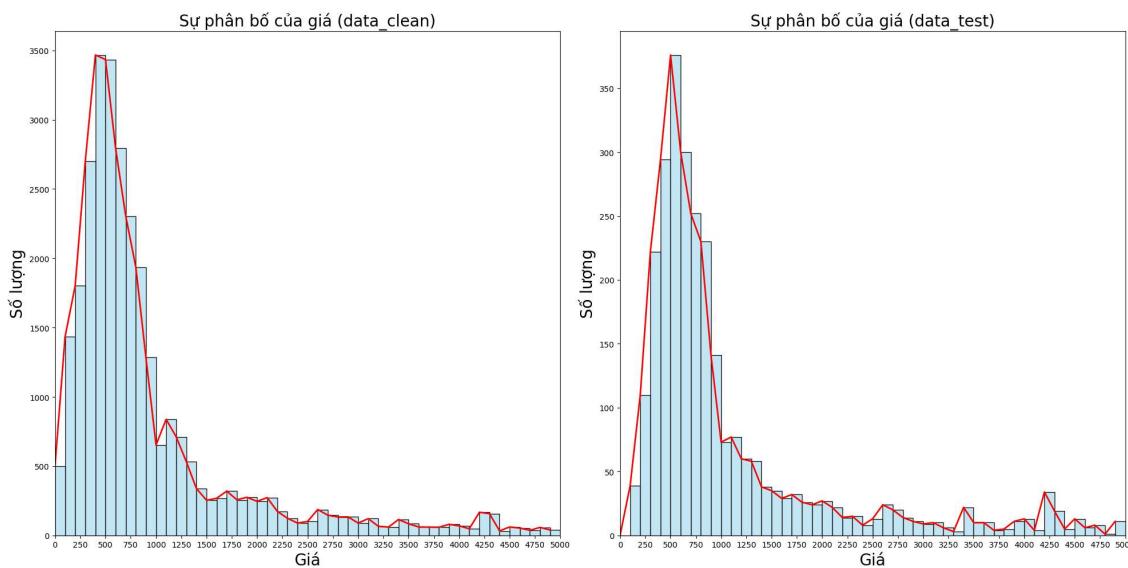
Đa số các mẫu tập trung ở phân khúc giá từ 250 đến 800 triệu.

Phân bố giữa biến mục tiêu và các biến quan trọng:



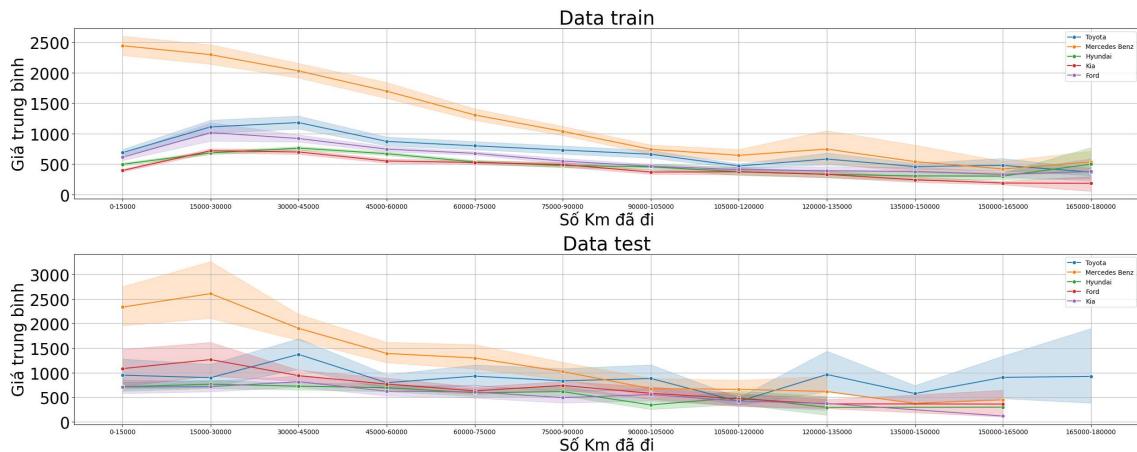
H.11 Đồ thị số lượng của 10 Hãng xe nổi bật của tập train và test10

Đồ thị cho thấy top 10 hãng xe có số lượng nhiều nhất trong tập train và tập test, tập test có số lượng bằng 10% so với tập train.



H.12 Đồ thị phân bố số lượng theo giá của data clean 11

2 đồ thị khác tương đồng nhau, nhưng số lượng xe trong khoảng 250 triệu ở tập train nhiều hơn so với tập test.



H.1312 Biểu đồ giá xe cũ trung bình 5 hãng phỏ biến Sô Km đã đi

Khoảng từ 0 đến 50.000 km và từ 120.000 trở đi biến động giá trung bình hơn nhiều so với tập train.

3. Trích xuất đặc trưng

3.1. Làm sạch, chuẩn hoá dữ liệu.

3.1.2 Xử lí dữ liệu:

- Chuẩn hoá cột “Sô Km đã đi”: loại bỏ đi các kí tự và chuyển về kiểu Int.
- Xoá những mẫu trùng hoặc tiêu nhiều hơn 3 trường dữ liệu từ file “raw data.csv”.
- Chuyển đổi cột “Giá” từ đơn vị đồng sang đơn vị triệu bằng cách chia cho 10 mũ 6 để dễ tính toán.
- Sửa các “Địa điểm” bị lỗi (có cả Huế và Hà Nội)
- Lọc dữ liệu bị trùng lặp.
- Chỉnh sửa biến “Ngày đăng” chỉ lấy tháng và năm.
- Chuẩn hoá trường “Số chỗ ngồi” từ string sang int.

3.1.2 Giảm chiều dữ liệu.

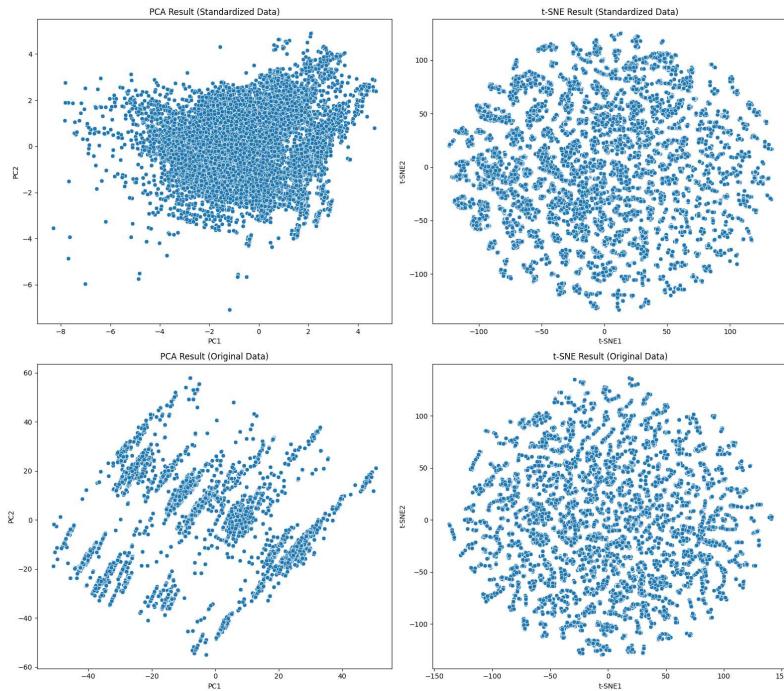
Để trực quan hóa và nhìn nhận tính chất phân cụm của dữ liệu. Ta lần lượt kiểm tra 2 phương pháp: PCA và t-sne để lựa chọn phương pháp tối ưu để giảm chiều dữ liệu.

Ta tiến hành so sánh trên 2 phương diện

- Đo thời gian thực thi của việc giảm chiều dữ liệu từ 12 đặc trưng về 2 đặc trưng đối với cả 2 phương pháp
- So sánh sự biến đổi của cấu trúc, phân bố khi có sự biến động trên tập gốc

Thời gian thực thi của PCA: 0.033 giây

Thời gian thực thi của t-SNE: 670.072 giây



H.1413 So sánh sự phân bố của dữ liệu của 2 phương pháp PCA và T-SNE

- Phương pháp PCA có thời gian thực thi cực kì nhanh hơn vượt trội so với phương pháp t-sne
- Phương pháp PCA lại bị ảnh hưởng, biến đổi lớn về cấu trúc và phân bố khi dữ liệu gốc có sự thay đổi còn t-sne vẫn còn giữ lại được, dữ liệu ít bị mất mát hơn

=> Lựa chọn t-sne để giảm chiều và trực quan hóa tuy lâu về mặc thời gian nhưng độ chính xác cao và mất mát dữ liệu thấp

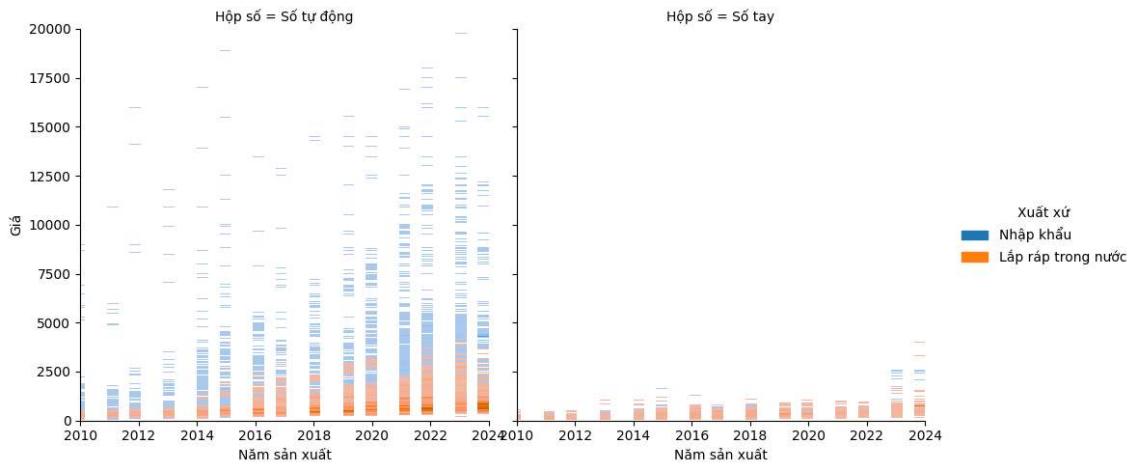
3.2. Lựa chọn đặc trưng.

Để cho việc xử lí 2 bài toán tốt hơn thì chúng ta sẽ chọn những đặc trưng phù hợp, nên chọn lựa những đặc trưng là số, nếu đặc trưng thuộc dạng category thì chúng ta sẽ sử dụng các phương pháp để chuẩn hoá mà mã hoá chúng.

Qua quá trình khảo sát cũng như đánh giá bằng trực quan hóa và dựa vào số liệu (sẽ được trình bày trong phần trực quan hóa) thì chúng em đã chọn được khoảng 10 đặc trưng quan trọng quyết định đến biến mục tiêu (Giá).

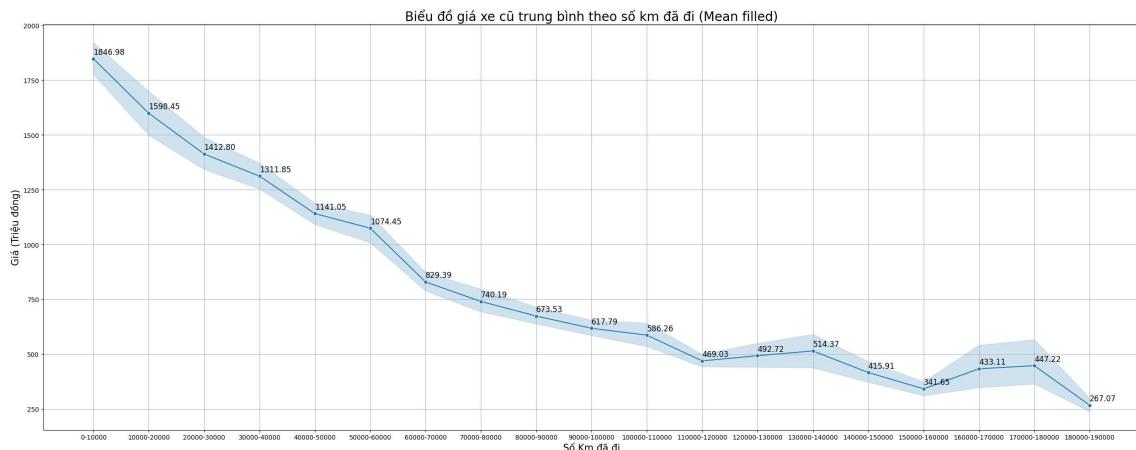
Các đặc trưng đã chọn bao gồm: Năm sản xuất,Tình trạng,Kiểu dáng,Số Km đã đi, Xuất xứ, Hãng xe, Dẫn động, Số chỗ ngồi,Địa điểm, Động cơ,Hộp số, Giá

3.3. Trực quan hóa.



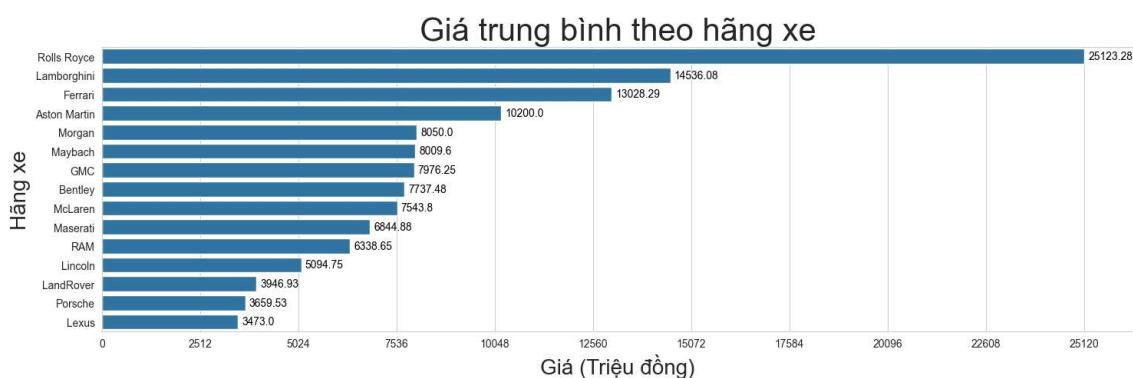
H.15 Biểu đồ phân bố giá theo năm sản xuất của Số tự động và Số tay 14

- Xe trên thị trường hiện nay chủ yếu là xe có hộp số tự động.
- Xe có xuất xứ nhập khẩu có giá lớn hơn khoảng 3 lần so với xe được lắp ráp trong nước.
- Xe có hộp số tự động nhập khẩu tăng mạnh qua các năm, còn có hộp số là số tay thì chủ yếu lắp ráp trong nước và ít quyết định đến giá xe qua các năm.



H.16 Biểu đồ giá xe cũ trung bình theo số Km đã đi 15

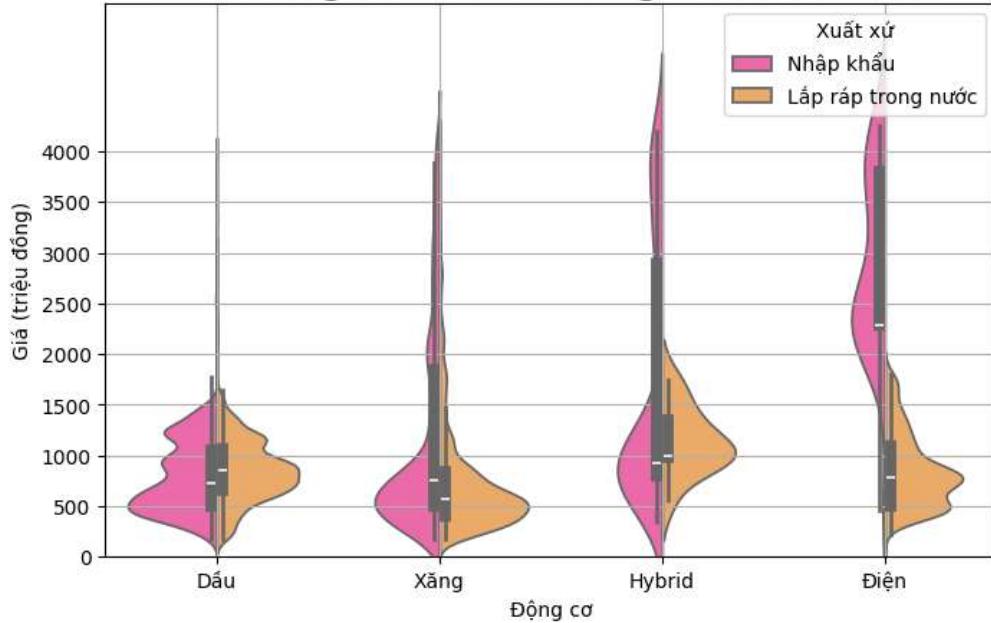
- Nhìn chung giá của xe giảm dần khi “Số km đã đi” càng tăng.



H.17 Biểu đồ giá xe trung bình theo hang xe16

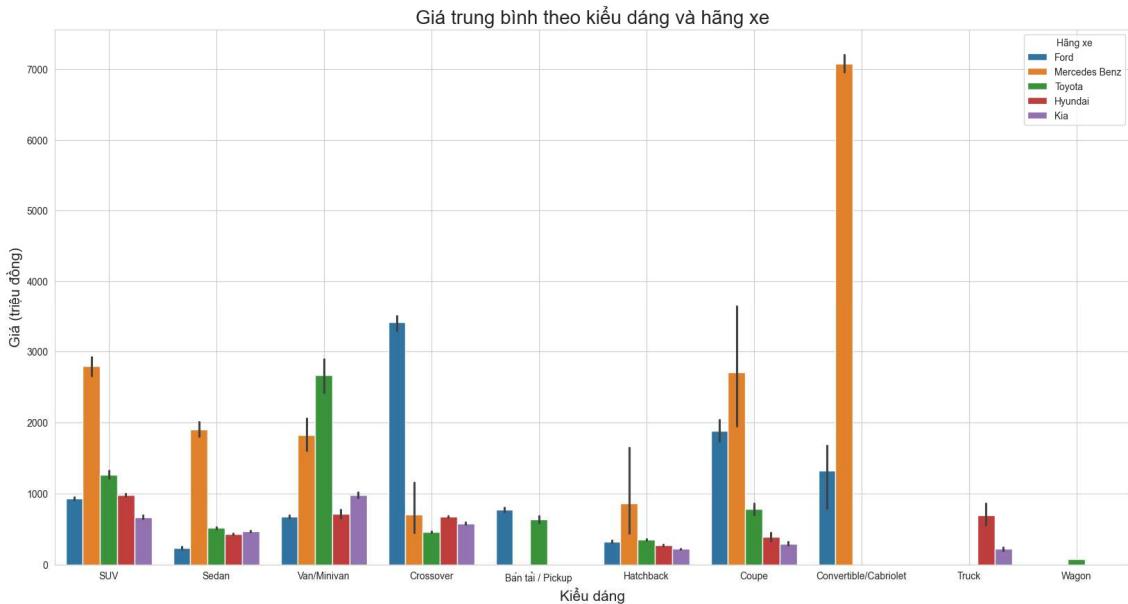
- Giá xe có sự ảnh hưởng nhiều bởi hang xe, các hang xe có giá trị cao như Roll Royce, Lamborghini, Ferrari,...

Giá trung bình theo động cơ và xuất xứ



H.18 Biểu đồ giá trung bình theo động cơ và xuất xứ17

- Tổng quan xe nhập khẩu có giá cao hơn xe lắp ráp trong nước.
- Xe điện nhập khẩu có giá cao hơn đáng kể so với xe điện lắp ráp trong nước.
- Trong tất cả các loại động cơ, xe hybrid có giá trung bình cao nhất, tiếp theo là xe điện, xe xăng và xe dầu.



H.19 Biểu đồ giá trung bình theo kiểu dáng và hãng xe18

- Nhìn chung xe mercedes có giá cao nhất ở 4 kiểu dáng suv, sedan, hatchback, couple và convertible/cabriolet
- Xe bán tải chỉ có 2 hãng Toyota và Ford với giá khoảng 700 triệu
- Toyota có giá đắt nhất trong kiểu xe van/minivan

4. Mô hình hóa dữ liệu

4.1. Dự đoán giá xe

4.1.1 Các thông số đánh giá mô hình

Mean squared Error (MSE):

Mean Squared Error (MSE) có lẽ là số liệu phổ biến nhất được sử dụng cho các bài toán hồi quy. Về cơ bản, nó tìm thấy sai số bình phương trung bình giữa các giá trị được dự đoán và thực tế. MSE là thước đo chất lượng của một công cụ ước tính - nó luôn không âm và các giá trị càng gần 0 càng tốt.

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

trong đó n là số điểm dữ liệu, y_i là giá trị quan sát và \hat{y}_i là giá trị dự đoán.

$$RMSE (\text{Root MSE}) = \sqrt{MSE}$$

R-squared (R2)

R-squared là một thước đo để xác định mức độ phù hợp của mô hình hồi quy với dữ liệu. Nó cho biết phần trăm phuơng sai của biến phụ thuộc được mô hình giải thích. R-squared có giá trị trong khoảng từ 0 đến 1, với 1 cho biết mô hình giải thích

toàn bộ sự biến động của dữ liệu và 0 cho biết mô hình không giải thích được bất kỳ biến động nào.

Công thức tính R-squared dựa trên tổng biệt phương của mô hình (SST) và tổng biệt phương còn lại sau khi sử dụng mô hình (SSE):

$$R^2 = 1 - (\text{SSE} / \text{SST})$$

Trong đó:

- SSE (Sum of Squared Errors) là tổng bình phương của các sai số (chênh lệch giữa giá trị thực tế và giá trị dự đoán).
- SST (Total Sum of Squares) là tổng bình phương của chênh lệch giữa các giá trị thực tế và giá trị trung bình.

Giá trị R-squared càng gần 1, mô hình càng phù hợp với dữ liệu, là một chỉ số hữu ích để đánh giá hiệu suất của mô hình hồi quy, nhưng cần kết hợp với các chỉ số khác để có đánh giá toàn diện hơn.

4.1.2 Phân chia dữ liệu

Việc tách dữ liệu thành các phần khác nhau là một giai đoạn cần thiết trong việc phát triển các mô hình học máy. Chúng em đã tiến hành tách dữ liệu ra làm ba phần: tập dữ liệu huấn luyện, tập xác thực và tập kiểm tra, với tỷ lệ và kỹ thuật phân chia thích hợp. Dưới đây là thông tin chi tiết về cách thức phân chia:

- Tập dữ liệu huấn luyện (Training set): Đây là phần dữ liệu dùng để train mô hình và điều chỉnh các thông số. Chúng em đã dành 70% (trong 30 ngàn dữ liệu), nhằm đảm bảo mô hình có đủ thông tin để học hỏi từ các xu hướng và tính năng có trong dữ liệu.
- Tập dữ liệu xác thực (Validation set): Phần này chiếm 30% (trong 30 ngàn dữ liệu) để kiểm tra và tinh chỉnh hiệu suất của mô hình. Tập dữ liệu này không tham gia vào quá trình huấn luyện, giúp chúng tôi kiểm tra khả năng áp dụng mô hình trên dữ liệu chưa biết trước.
- Tập dữ liệu kiểm tra (Test set): Với khoảng 3000 dữ liệu mới hoàn toàn, phần này được dùng để đánh giá cuối cùng cho mô hình sau khi đã qua quá trình huấn luyện và điều chỉnh. Tập dữ liệu này hoàn toàn tách biệt từ quá trình huấn luyện và xác thực, đại diện cho những tình huống thực tế mà mô hình sẽ gặp phải.

Chúng em đã áp dụng kỹ thuật phân chia dữ liệu một cách ngẫu nhiên để đảm bảo tính đa dạng và không thiên lệch trong các mẫu dữ liệu, giúp mô hình có khả năng tổng quát hóa tốt khi gặp dữ liệu mới.

4.1.3 Linear Regression

1. Cơ sở lý thuyết

Linear regression là một thuật toán trong thống kê và học máy được sử dụng để xác định mối quan hệ tuyến tính giữa một biến phụ thuộc (được gọi là biến mục tiêu hoặc biến phản hồi) và một hoặc nhiều biến độc lập (được gọi là biến đầu vào hoặc biến giải thích).

Cơ sở lý thuyết của Linear regression dựa trên mô hình tuyến tính, trong đó giả định rằng mối quan hệ giữa biến phụ thuộc và biến đầu vào có thể được mô tả bằng một đường thẳng. Mô hình tuyến tính được biểu diễn bởi phương trình:

$$y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \dots + \beta_n \times X_n + \epsilon$$

Trong đó:

- y là biến mục tiêu (biến phụ thuộc).
- X_1, X_2, \dots, X_n là các biến đầu vào.
- $\beta_0, \beta_1, \beta_2, \dots, \beta_n$ là các hệ số hồi quy (được gọi là hệ số tương ứng với từng biến đầu vào)
- ϵ là sai số ngẫu nhiên (mô hình cho rằng một phần của sự biến động của biến phụ thuộc không thể được giải thích bởi các biến đầu vào)

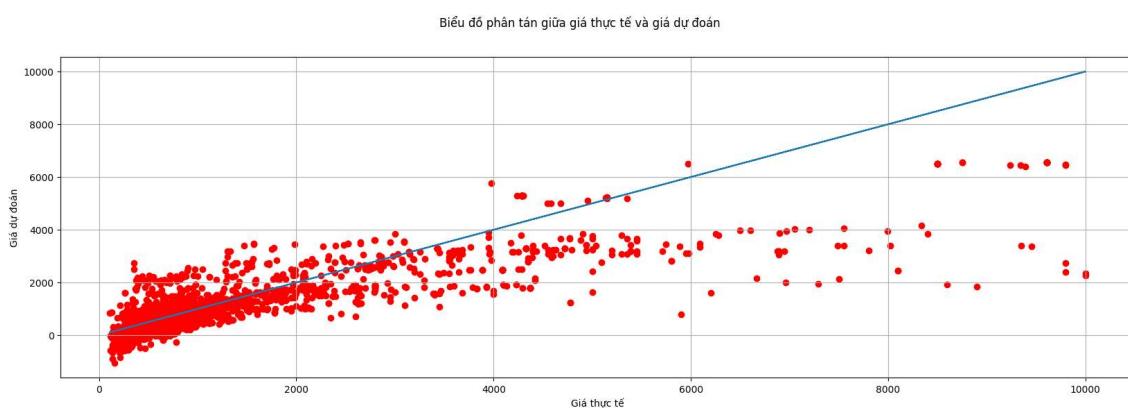
Mục tiêu của mô hình này là tìm ra các hệ số $\beta_0, \beta_1, \beta_2, \dots, \beta_n$ sao phù hợp với dữ liệu. Quá trình này được thực hiện bằng phương pháp Gradient Descent.

Linear regression có thể được áp dụng cho cả bài toán dự đoán (prediction) và bài toán phân tích (inference). Trong bài toán dự đoán, chúng ta sử dụng mô hình tuyến tính để dự đoán giá trị của biến phụ thuộc dựa trên các giá trị của biến đầu vào.

2. Kết quả thực thi mô hình

Chọn bộ tham số mặc định của thư viện

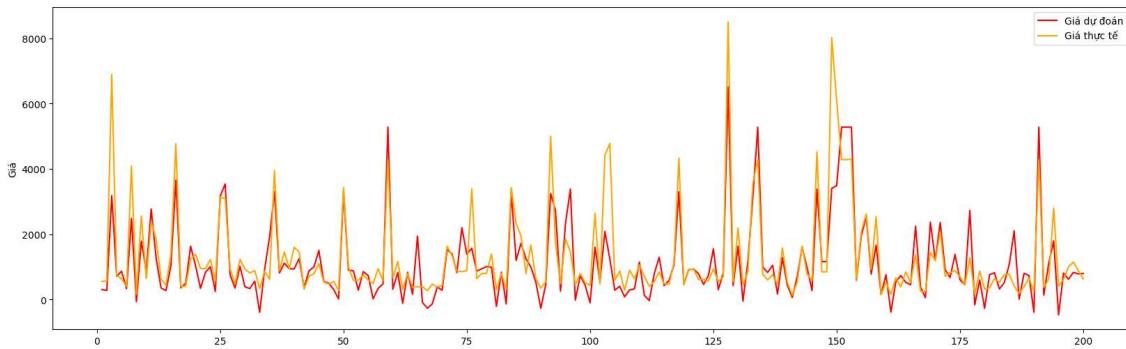
Mô hình	Metrics	Validate	Test
Linear Regression	R2 score RMSE(Triệu đồng)	0.727 751.468	0.667 852.050



H.20 Biểu đồ phân tán giữa giá thực tế và giá dự đoán

- Dễ dàng nhận thấy xe có giá dưới 2 tỷ dự đoán khá chính xác, trên phân khúc giá 2 tỷ thì mô hình có xu hướng dự đoán lệch về phía dưới, dự đoán giá thấp hơn so với thực tế.

Giá dự đoán với giá thực tế



H.21 Biểu đồ giá dự đoán với giá thực tế

- Qua biểu đồ trên ta có thể thấy được mô hình dự đoán giá khác xác với thực tế ở khoảng giá dưới 2 tỷ và biến động mạnh trên khoảng 2 tỷ.

4.1.4 Gradient Boosting Regression

1. Cơ sở lý thuyết

Gradient Boosting Regression là một kỹ thuật học máy giám sát, thuộc họ Ensemble Learning. Nó dựa trên nguyên lý tăng cường dần dần (boosting) các mô hình học cơ bản (như cây quyết định) để tạo thành một mô hình mạnh mẽ hơn. Dưới đây là lý thuyết cơ bản về Gradient Boosting Regression:

Ý tưởng chính:

- Bắt đầu bằng một mô hình cơ bản (như cây quyết định) và dự đoán sai số (residual) so với kết quả thực tế.
- Xây dựng một mô hình mới để dự đoán sai số trên, sau đó cập nhật mô hình tổng thể bằng cách cộng thêm mô hình mới này.
- Lặp lại quá trình này nhiều lần để giảm dần sai số, tạo thành một mô hình tổng thể mạnh mẽ.

Thuật toán:

- Khởi tạo mô hình ban đầu $F_0(x) = 0$.
- Lặp lại.
 - Tính sai số $r_i = y_i - F(x_i)$.
 - Xây dựng một mô hình mới $h(x)$ để dự đoán sai số r .
 - Cập nhật mô hình $F(x) = F(x) + \eta * h(x)$, với η là tốc độ học.
- Lặp lại bước 2 cho đến khi đạt được kết quả mong muốn.

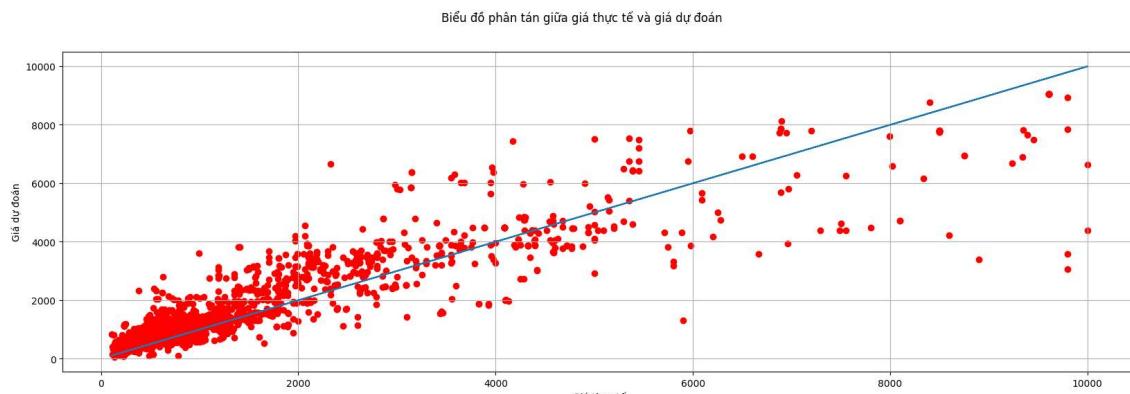
Gradient Boosting Regression là một kỹ thuật mạnh mẽ và phổ biến trong học máy, với khả năng xử lý các bài toán hồi quy phức tạp.

2. Kết quả thực thi mô hình

Với bộ tham số tối ưu:

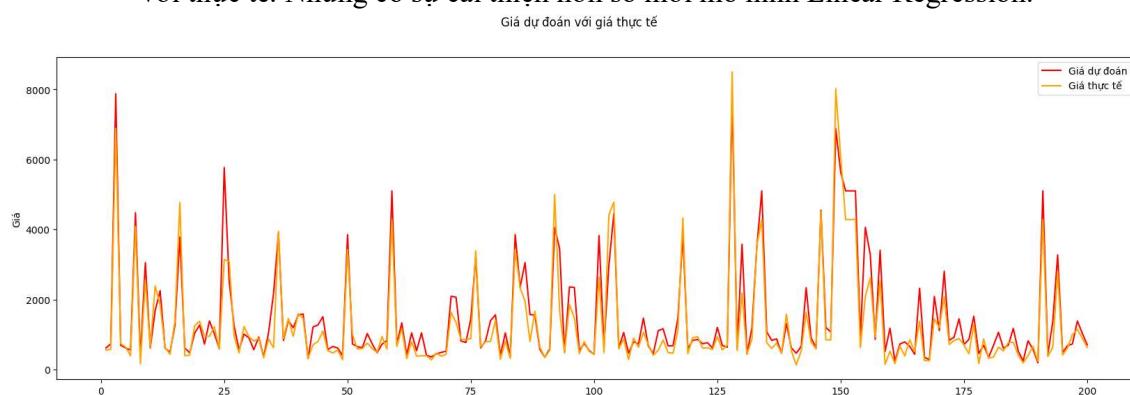
- Số lượng cây (n_estimators): 200
- Learning rate: 0.1,
- Chiều sâu tối đa (max_depth): 7
- min_samples_split: 2
- min_samples_leaf: 1

Mô hình	Metrics	Validate	Test
Gradient Boosting	R2 score	0.971	0.806
	RMSE(Triệu đồng)	244.207	650.429



H.22 Biểu đồ phân tán giữa giá thực tế và giá dự đoán

- Dễ dàng nhận thấy xe có giá dưới 4 tỷ dự đoán khá chính xác, trên phân khúc giá 4 tỷ thì mô hình có xu hướng dự đoán lệch về phía dưới, dự đoán giá thấp hơn so với thực tế. Nhưng có sự cải thiện hơn so với mô hình Linear Regression.



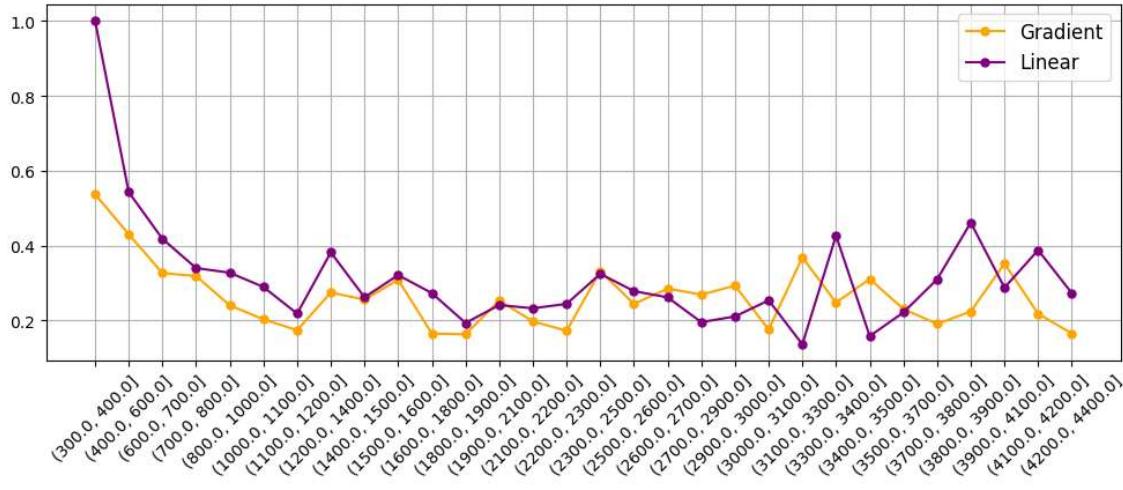
H.23 Biểu đồ phân tán giữa giá dự đoán và giá thực tế

Đường giá trị dự đoán và thực tế nhìn chung khá khớp nhau, chứng tỏ mô hình khá tốt, tốt hơn nhiều so Linear Regression.

4.1.5 So sánh 2 mô hình

Mô hình	Metrics	Validate	Test
Linear Regression	R2 score	0.727	0.667
	RMSE(Triệu đồng)	751.468	852.050

Gradient Boosting	R2 score	0.971	0.806
	RMSE(Triệu đồng)	244.207	650.429



H.24 Biểu đồ giá dự đoán của Linear và Gradient

- Xe có giá khoảng 200-500 triệu có tỉ lệ sai số cao (do đối với xe giá rẻ thì sai số / giá cao hơn, đối với xe giá lớn thì sai số / giá thấp hơn).
- Mô hình Gradient cho thấy tỉ lệ sai số thấp nhất so với mô hình còn lại.
- Mô hình Gradient Boosting ổn định hơn Linear Regression ở khoảng giá cao trên 2 tỷ, các thông số đánh giá cũng cao hơn. Khoảng giá dưới 2 tỷ thì cả 2 đều tương đối ổn định.

4.2. Phân cụm

4.2.1 Mô hình K-means.

Cơ sở lý thuyết.

Phân cụm (Clustering): K-Means là một thuật toán phân cụm, có nghĩa là nó nhóm các điểm dữ liệu vào k cụm khác nhau dựa trên sự tương đồng của chúng.

Trung tâm cụm (Cluster Centroid): Mỗi cụm có một điểm trung tâm, còn gọi là trung tâm cụm. Trung tâm cụm là trung bình của tất cả các điểm trong cụm đó.

Hàm mục tiêu (Objective Function): Thuật toán K-Means cố gắng tối thiểu hóa tổng bình phương khoảng cách giữa các điểm dữ liệu và trung tâm cụm gần nhất. Đây được gọi là hàm mục tiêu.

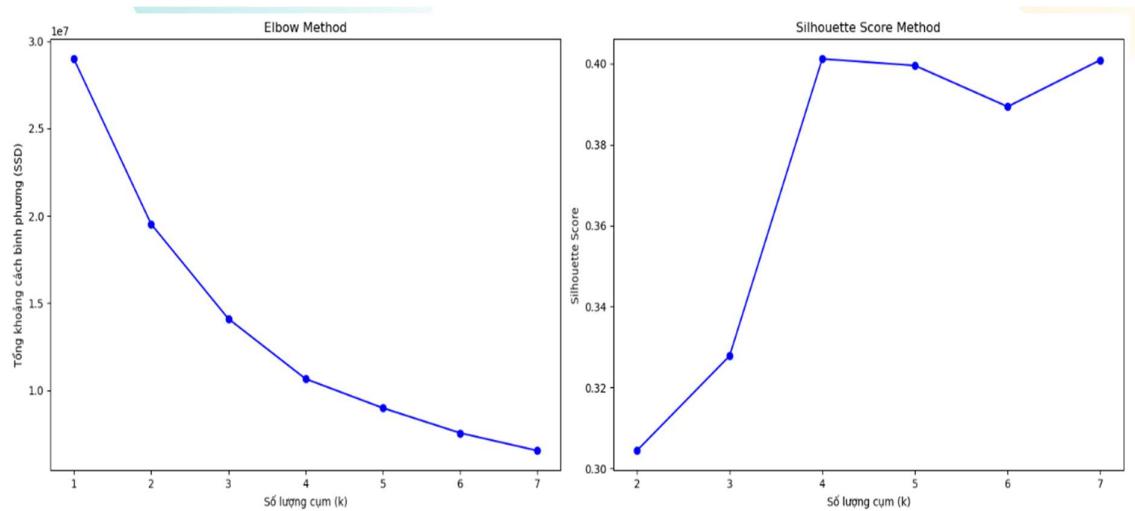
Thuật toán K-Means: Thuật toán có các bước sau:

- Chọn ngẫu nhiên k điểm dữ liệu làm các trung tâm cụm ban đầu.
- Gán mỗi điểm dữ liệu vào cụm có trung tâm gần nhất.
- Tính lại vị trí của các trung tâm cụm bằng cách lấy trung bình của các điểm trong cụm.
- Lặp lại các bước 2 và 3 cho đến khi hàm mục tiêu không thể giảm thêm.

Kết quả thực thi mô hình

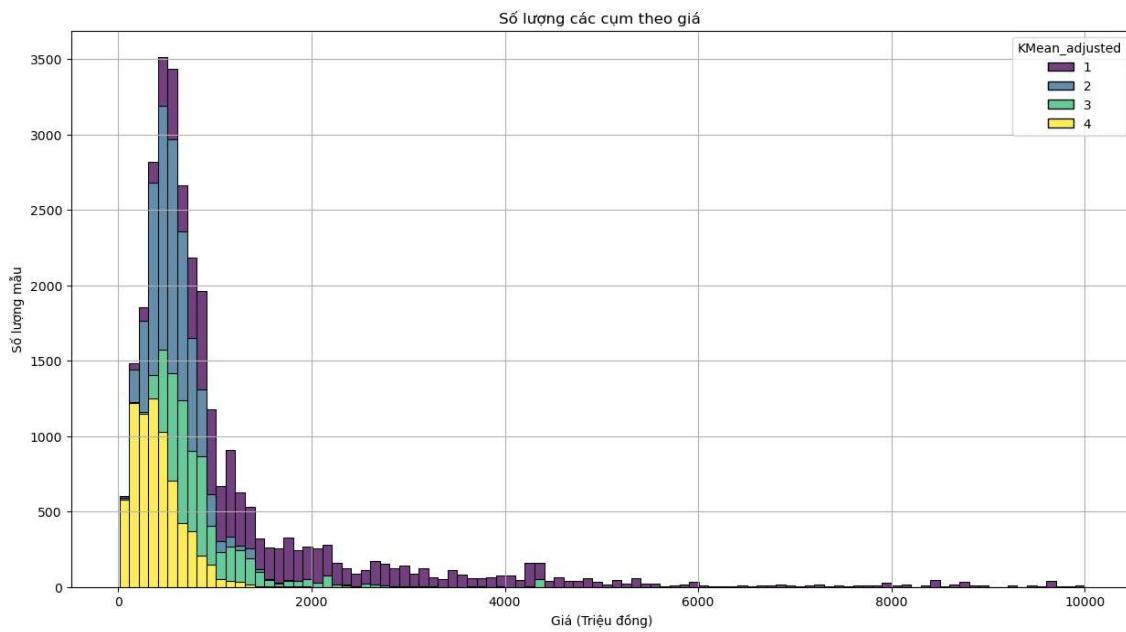
Xác định số cụm K ban đầu:

- Sử dụng Elbow Method và Silhouette Score.
- Elbow Method: Xác định K ở điểm mà giá tốc đồ thị đột ngột thay đổi.
- Silhouette Score: Xác định K ở điểm cực đại của đồ thị.

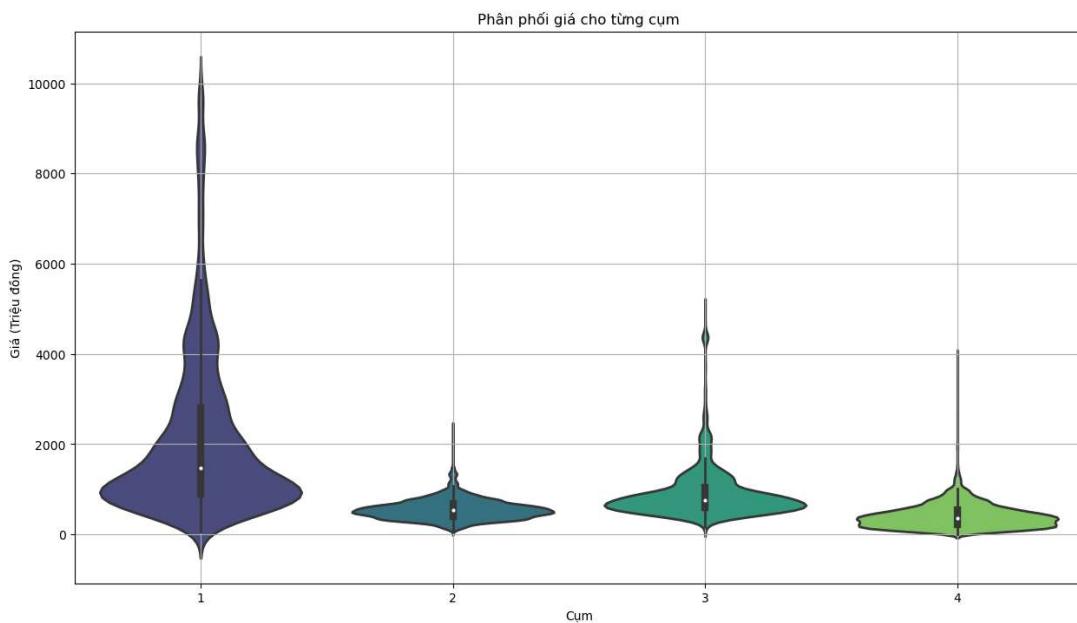


H.25 Biểu đồ đánh giá cách lựa chọn k

- Dựa vào đồ thị, ta thấy điểm K tối ưu là K = 4.

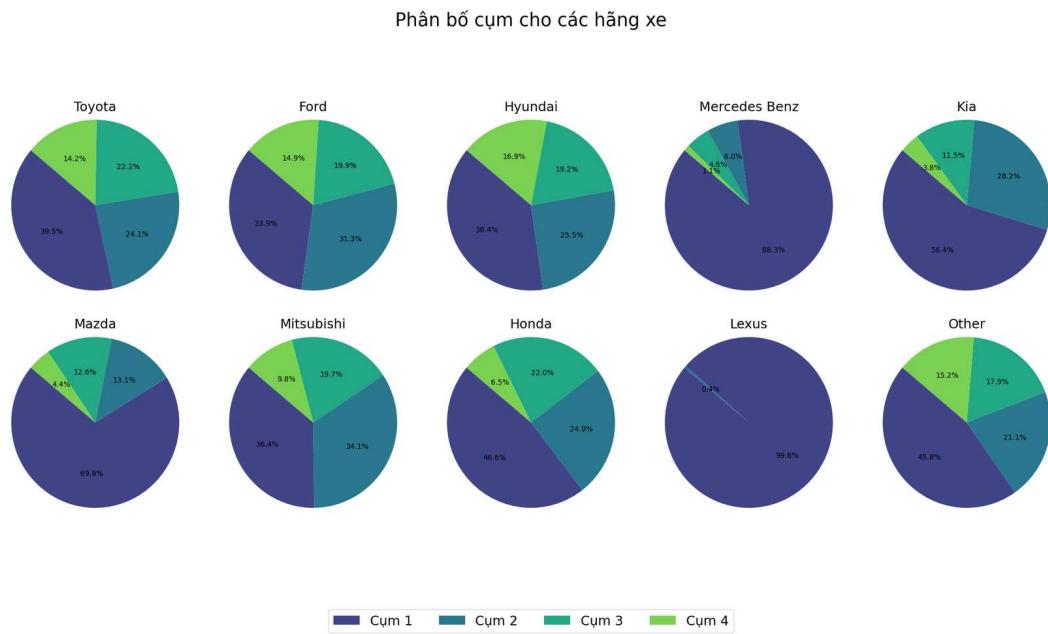


H.26 Biểu đồ phân bố các cụm trên đặc trưng giá



H.27 Biểu đồ phân bố của các cụm trên đặc trưng Giá

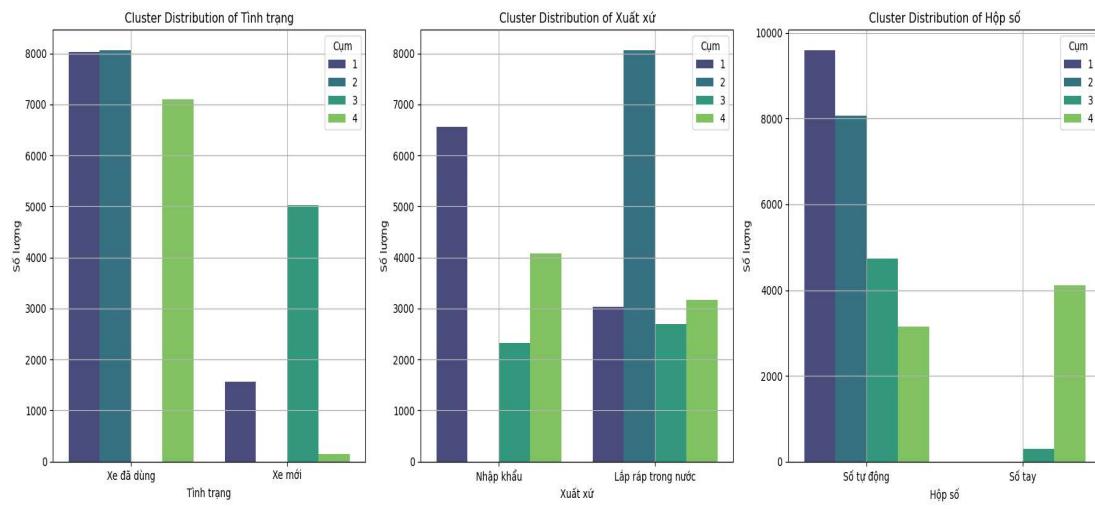
- Cụm 1 chiếm số lượng nhiều nhất và phân bố trải dài nhất trong 4 cụm.
- Ở khoảng giá 1.5 tỉ trở xuống, số lượng mẫu cụm 2 nhiều hơn cụm 3 nhưng cụm 3 trải dài hơn và trải đến 4 tỉ.
- Cụm 4 nhìn chung có số lượng ít nhất và phân bố chủ yếu trong khoảng 0 tới 1 tỉ.



H.28 Biểu đồ phân bố cụm theo các hãng xe

- Về hãng xe, nhìn chung Cụm 1 có tỉ lệ số lượng cao nhất trong 4 cụm ở mọi hãng xe.
- Tiếp đến là cụm 2 và cụm 3 có tỉ lệ số lượng khá tương đồng.

- Cụm 4 có tỉ lệ về số lượng thấp nhất.



H.29 Biểu đồ phân bố của các cụm trên các đặc trưng nhị phân

- Về tình trạng, có sự phân bổ số lượng không đều ở 4 cụm.
- Về xuất xứ, cụm 1 và 2 có phân bổ không đều về nhập khẩu và trong nước, cụm 3 và 4 có số lượng phân bổ khá đều.
- Về hộp số, cụm 4 phân bổ đều ở 2 hộp số, 3 cụm còn lại không đều nhau.

4.2.1 Mô hình GMM

Cơ sở lý thuyết.

Gaussian Mixture Model (GMM) là một mô hình thống kê dùng để biểu diễn dữ liệu bao gồm nhiều phân phối Gaussian khác nhau.

Mô hình này thường được sử dụng để thực hiện phân cụm (clustering) và mô hình hóa phân phối của dữ liệu phức tạp.

Thuật toán:

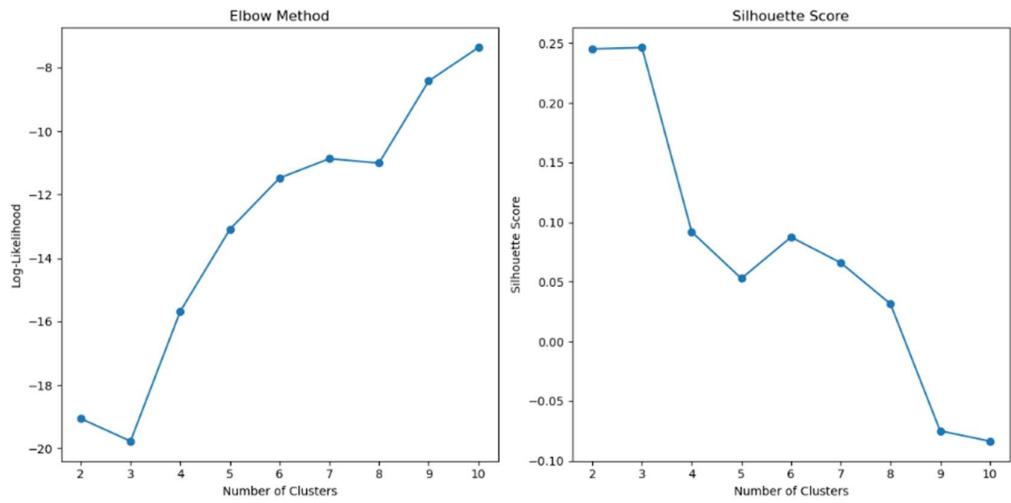
Thành phần của GMM: GMM bao gồm nhiều phân phối Gaussian (còn gọi là thành phần) với các tham số riêng biệt.

- Bước 1: Khởi tạo
- Khởi tạo các tham số của mô hình (trọng số, giá trị trung bình và phương sai của các phân phối Gaussian).
- Bước 2: E-step (Expectation Step)
- Tính toán xác suất một điểm dữ liệu thuộc về mỗi thành phần Gaussian.
- Bước 3: M-step (Maximization Step)
- Cập nhật các tham số của mô hình dựa trên các xác suất đã tính toán ở bước E-step.
- Bước 4: Lặp lại
- Lặp lại bước 2 và 3 cho đến khi hội tụ (các tham số không thay đổi hoặc thay đổi rất ít).

Kết quả thực thi

Xác định số cụm

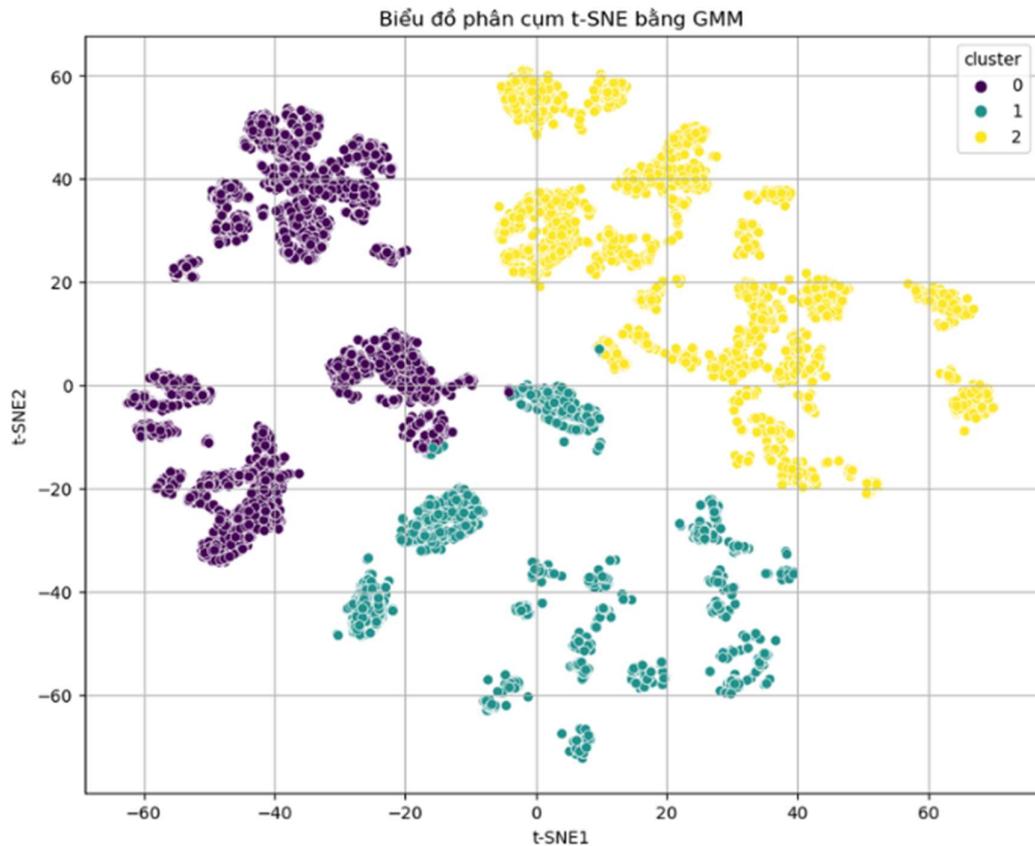
Sử dụng phương pháp elbow và Silhouette Score



K = 3

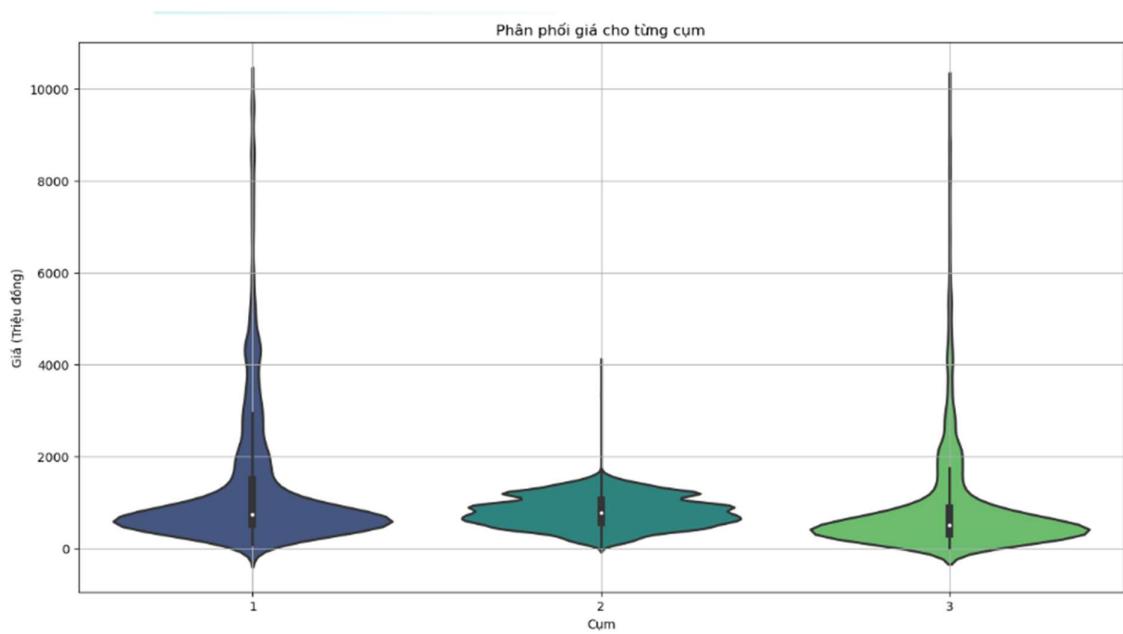
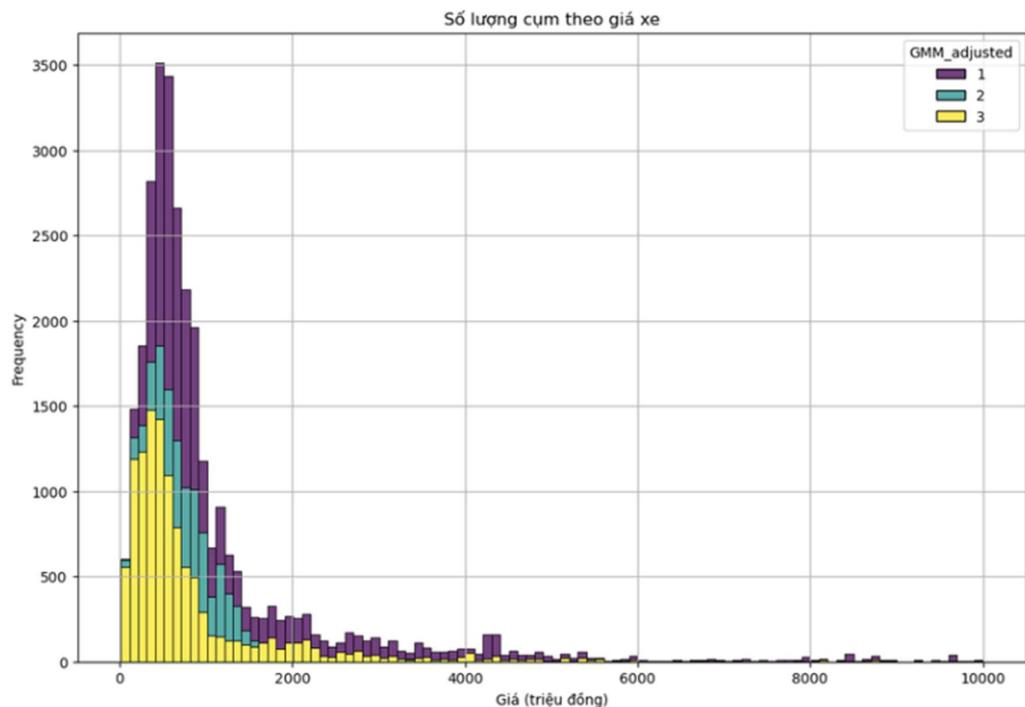
H.30 Biểu đồ đánh giá lựa chọn số cụm của GMM

- Dựa vào cả 2 đồ thị, ta xác định được K tối ưu là K = 3.



H.31 Biểu đồ phân cụm bằng phương pháp GMM

Một số kết quả trực quan hóa:

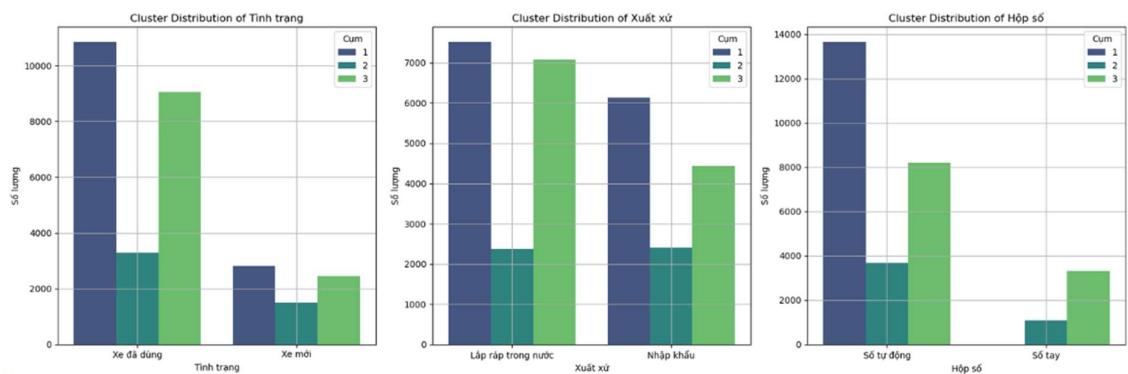


- Cụm 1 có số lượng nhiều nhất ở tất cả các giá.
- Cụm 2 có số lượng nhỉnh hơn cụm 3 1 chút ở khoảng dưới 2 tỉ, tuy nhiên cụm 3 trải dài hơn.



H.34 Biểu đồ phân bố các cụm trên đặc trưng Hạng xe

- Cụm 1 có tỉ lệ số lượng nhiều nhất.
- Tiếp đến là cụm 2 và cụm 3 thấp nhất.



H.35 Biểu đồ phân bố các cụm trên đặc trưng nhị phân

- Ở phần tình trạng, cả 3 cụm đều phân bố không đều: 3 cụm phân bố ở xe đã dùng nhiều hơn xe mới.
- Ở phần xuất xứ, cả 3 cụm phân bố và tương đồng nhiều.
- Về phần hộp số thì 3 cụm 1 2 3 phân bố nhiều hơn đáng kể ở số tự động, còn số tay ít.

5. Kết luận

5.1. Kết luận:

5.1.1 Bài toán dự đoán:

2 mô hình dự đoán tốt và ổn định ở trong khoảng giá thấp (dưới 4 tỷ), còn càng cao thì mô hình Gradient Boosting dự đoán ổn hơn so với Linear Regression nhưng nhìn chung cũng không được tốt. Lý do là vì dữ liệu của những xe có giá trị cao thì có ít hơn nhiều so với dữ liệu của những xe có giá tầm trung

Vấn đề với đặc trưng: có thể có những đặc trong dữ liệu mà mô hình chưa thể hiện tốt. Việc lựa chọn đặc trưng là 1 việc rất quan trọng giúp giảm thời gian huấn luyện cũng như tăng hiệu quả của mô hình.

5.1.2 Bài toán phân cụm:

Mô hình GMM cho thấy ưu điểm, tốt hơn mô hình K-means. Có thể nhận diện và phân cụm rõ ràng ưu việt hơn.

Đối với mô hình GMM dữ liệu phân chia thành 3 cụm là hợp lý nhất, cụ thể đa số các đặc trưng GMM phân cụm đồng đều và hợp lý, tuy nhiên vẫn có những đặc trưng mô hình phân bổ không đều như tình trạng, hộp số, ...

Đối với mô hình K-means dữ liệu được phân chia thành 4 cụm là tối ưu nhất cụ thể như sau đa số các đặc trưng mô hình dự đoán không đồng đều đáng kể, nhất là các đặc trưng như giá, số Km đã đi, năm sản xuất, ...

Những đặc trưng như Giá, Tình trạng, Hộp số cả 2 mô hình đều dự đoán không đồng đều

5.2. Hướng phát triển:

a) Bài toán dự đoán:

Cải thiện chất lượng dữ liệu: Tập trung vào việc thu thập thêm dữ liệu cho phân khúc giá cao, để bù đắp sự thiếu hụt dữ liệu hiện tại.

Kiểm tra và loại bỏ các giá trị ngoại lai (outliers) có thể gây ảnh hưởng đến chất lượng mô hình.

Nghiên cứu các phương pháp xử lý dữ liệu bị thiếu (missing data) phù hợp.

Cải tiến quá trình chọn đặc trưng (feature engineering): thử nghiệm các kỹ thuật chọn đặc trưng tiên tiến như Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), hoặc sử dụng các mô hình machine learning để đánh giá tầm quan trọng của các đặc trưng.

Kết hợp các đặc trưng hiện có để tạo ra các đặc trưng mới có ý nghĩa hơn.

Xem xét các yếu tố bối cảnh như tình hình kinh tế, chính sách, xu hướng thị trường... để bổ sung thêm đặc trưng.

Thử nghiệm các mô hình học máy nâng cao: ngoài linear regression và gradient boosting, hãy thử nghiệm các mô hình như Random Forest, Support Vector Regression, Neural Networks,...

So sánh hiệu suất của các mô hình và tìm ra mô hình phù hợp nhất.

Kết hợp các mô hình (model ensemble) để tận dụng ưu điểm của từng mô hình.

Dánh giá và điều chỉnh mô hình: Sử dụng các phương pháp đánh giá mô hình khác như cross-validation, holdout set để có đánh giá chính xác hơn.

Phân tích lỗi dự báo để tìm ra các nguyên nhân và cải thiện mô hình.

Theo dõi hiệu suất mô hình trên dữ liệu mới và cập nhật/tinh chỉnh mô hình khi cần thiết.

b) Bài toán phân cụm:

➤ Cải thiện Hiệu suất và Độ chính xác

Tối ưu hóa các tham số: Tối ưu hóa các tham số của thuật toán như số cụm (K) trong K-means hoặc số thành phần trong GMM để cải thiện độ chính xác của phân cụm.

Sử dụng phương pháp khởi tạo tốt hơn: Đối với K-means, sử dụng K-means++ để khởi tạo các centroid giúp cải thiện sự hội tụ và kết quả phân cụm.

Tích hợp các tiêu chí đánh giá: Sử dụng các tiêu chí như Silhouette Score, Davies-Bouldin Index hoặc Akaike Information Criterion (AIC) và Bayesian Information Criterion (BIC) để đánh giá và chọn số cụm hợp lý.

➤ Xử lý dữ liệu phức tạp và không đồng nhất

Phân cụm dữ liệu không đồng nhất: Sử dụng các biến đổi dữ liệu hoặc kỹ thuật tiền xử lý dữ liệu như PCA, t-SNE hoặc UMAP để giảm chiều dữ liệu và làm cho dữ liệu dễ phân cụm hơn.

Phân cụm dữ liệu có phân phối không đồng nhất: Đối với dữ liệu có phân phối không đồng nhất, GMM có thể được cải thiện bằng cách sử dụng các phương pháp nâng cao như Variational Bayesian Gaussian Mixture Model.

➤ Khả năng mở rộng và tính toán phân tán

Mở rộng thuật toán: Phát triển các phiên bản phân tán của K-means và GMM để xử lý các tập dữ liệu lớn hơn bằng cách sử dụng các nền tảng tính toán phân tán như Apache Spark hoặc Dask.

Tối ưu hóa hiệu suất tính toán: Sử dụng các kỹ thuật tối ưu hóa như giảm kích thước dữ liệu hoặc sử dụng các cấu trúc dữ liệu hiệu quả để cải thiện hiệu suất tính toán.

6. Tài liệu tham khảo

- [1] Giới thiệu về Feature Engineering – Pham Dinh Khanh, https://phamdinhkhanh.github.io/deepai-book/ch_ml/index_FeatureEngineering.html
Giới thiệu về Feature Engineering – Pham Dinh Khanh, https://phamdinhkhanh.github.io/deepai-book/ch_ml/index_FeatureEngineering.html
- [2] Slide môn “Khoa học dữ liệu” – Thầy Ninh Khánh Duy - Trường ĐH Bách Khoa
Đà Nẵng, <https://drive.google.com/drive/folders/12KNleapSgtcwCJglSsbpLtkIDVN4pG47>
- [3] Slide môn “Trí tuệ nhân tạo” – Thầy Nguyễn Văn Hiệu – Trường ĐH Bách Khoa
Đà Nẵng, <https://drive.google.com/drive/folders/1GC5-GAIub1F5kWazw702ovxsgNyAPZbP>
- [4] Gradient Boosting - Tát tần tật về thuật toán mạnh mẽ nhất trong Machine Learning – Bui Tien Tung, <https://viblo.asia/p/gradient-boosting-tat-tan-tat-ve-thuat-toan-manh-me-nhat-trong-machine-learning-YWOZrN7vZQ0>
- [5] Ebook Machine Learning cơ bản – Vũ Hữu Tiệp,
<https://machinelearningcoban.com>