ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

ĐỀ TÀI: DỰ ĐOÁN GIÁ XE MÁY CŨ

LỚP: IE224.O11

Sinh viên thực hiện:		
STT	**Họ tên**	**MSSV**
1	Kiều Minh Phước	20521774
2	Nguyễn Thành Long	20521571
3	Lưu Thượng Vỹ	20522179
4	Trương Minh Phong	20521751
5	Lâm Quốc Đạt	20520433

TP. HỒ CHÍ MINH – 12/2023

1 GIỚI THIỆU

Nhóm chọn đề tài phân tích dự đoán giá xe cũ trên thị trường. Bộ dữ liệu sử dụng công cụ thu thập: Selenium [2] và Beautiful Soup [3]. Và công cụ phân tích là Jupyter Notebooks. Nhóm đã chọn giải pháp phân tích là phân tích EDA và phương pháp thu thập là tự thu thập tại website Chợ Tốt [1] bằng cách thu thập tự động. Các độ đo đánh giá nhóm lựa chọn như: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), Correlation, ANOVA. Kết quả đạt được: Tìm ra các biến số, biến phân loại có ảnh hưởng đến biến mục tiêu(giá xe cũ) trong bộ dữ liệu, chọn ra được mô hình phù hợp nhất đó là Random Forest.

Bộ dữ liệu được nhóm tự thu thập tại website Chợ Tốt [1] bằng cách thu thập tự động thông qua các thư viện hổ trợ bởi python như Selenium [2] và Beautiful Soup [3]. Bộ dữ liệu và đề tài do nhóm tự phân tích thiết kế và không dựa trên đề tài nào khác.

2 MÔ TẢ BỘ DỮ LIỆU

Mô tả bộ dữ liệu: Dự đoán giá xe máy cũ.Bộ dữ liệu này chứa thông tin về các chiếc xe máy cũ đang được bán trên thị trường. Thông tin bao gồm các thuộc tính như: automaker, series, year, kilometers_traveled, condition, type, volume, origin, warranty_policy, weight, location, partner, price. Bộ dữ liệu chứa 20185 dữ liệu chiếc xe cũ.Bộ dữ liệu có kích thước 2,5 MB. Bộ dữ liệu được định dạng dưới dạng csv.

Bộ dữ liệu được nhóm tự thu thập tại website Chợ Tốt [1] bằng cách thu thập tự động thông qua các thư viện hổ trợ bởi python như Selenium [2] và Beautiful Soup [3].

Mô tả ý nghĩa các biến/cột dữ liệu:

STT	Thuộc tính	Kiểu dữ liệu	Ý nghĩa
1	automaker	object	Tên hãng sản xuất của chiếc xe
2	series	object	Tên dòng xe của chiếc xe
3	year	int	Năm sản xuất của chiếc xe

4	kilometers traveled	float	Số KM đã chạy của chiếc xe
5	condition	object	Tình trạng của chiếc xe
6	type	object	Loại xe(xe số, xe ga, xe tay côn)
7	volume	object	Dung tích của chiếc xe
8	origin	object	Tên nước sản xuất chiếc xe
9	warranty_policy	object	Chính sách bảo hàng của xe(còn hoặc không)
10	weight	object	Trọng lượng của chiếc xe(kg)
11	location	object	Vị trí bán xe của người đăng bán
12	partner	int	Đối tác liên kết với người đăng bán xe(có hoặc không)
13	price	float	Giá bán của chiếc xe

Bảng1. Mô tả ý nghĩa các biến dữ liệu.

Bộ dữ liệu xe máy cũ có **13 cột** gồm: automaker, series, year, kilometers_traveled, condition, type, volume, origin, warranty_policy, weight, location, partner, price. Bộ dữ liệu xe máy cũ có **20185 dòng**. Bộ dữ liệu xe máy cũ có **9 biến phân loại**. Bộ dữ liệu xe máy cũ có **4 biến số**.

Thống kê số lượng khuyết liên quan bộ dữ liệu:

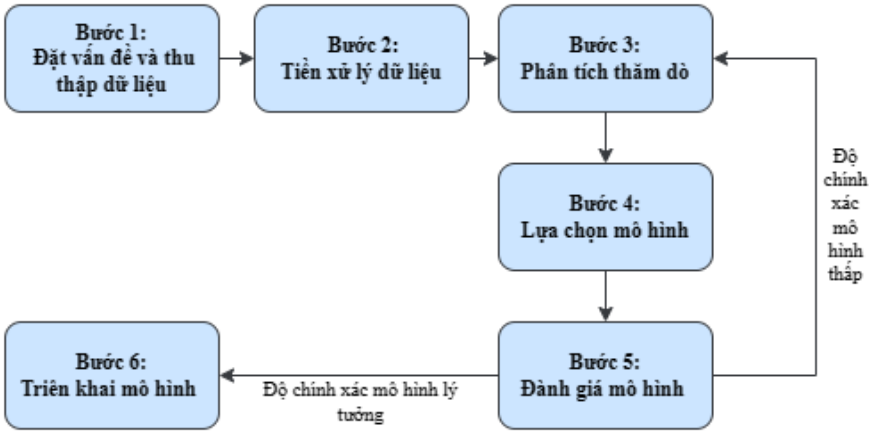
STT	**Cột**	**Kiểu dữ liệu**	**Số giá trị khuyết**	**STT**	**Cột**	**Kiểu dữ liệu**	**Số giá trị khuyết**
1	automaker	object	9	8	origin	object	0
2	series	object	10	9	warranty_policy	object	32
3	year	int	9	10	weight	object	45
4	kilometers_traveled	float	509	11	location	object	0

5	condition	object	2	12	partner	int	0
6	type	object	9	13	price	float	42
7	volume	object	5033				

Bảng2: Thống kê số lượng khuyết liên quan bộ dữ liệu.

3 PHƯƠNG PHÁP PHÂN TÍCH

Để dự đoán giá xe cũ nhóm đã thực hiện theo quy trình gồm 6 bước như hình 1.



Hình 1. Quy trình dự đoán giá xe cũ

Hình 1. Mô tả quá trình thực hiện phân tích và đánh giá xe cũ. Đầu tiên cần đặt ra các vấn đề và thu thập bộ dữ liệu thô từ các nguồn ở bước 1 và tiến hành tiền xử lý trước khi phân tích ở bước 2. Bước tiếp theo là phân tích EDA bộ dữ liệu để đưa ra các mô hình có khả năng phù hợp với bài toán ở bước 3. Việc thực nghiệm và đánh giá các mô hình sẽ thực hiện ở bước 4 và 5. Kết quả đạt được là mô hình lý tưởng cho bài toán ở bước 6. Chi tiết về quá trình thực hiện mỗi bước sẽ được trình bài ở phần tiếp theo.

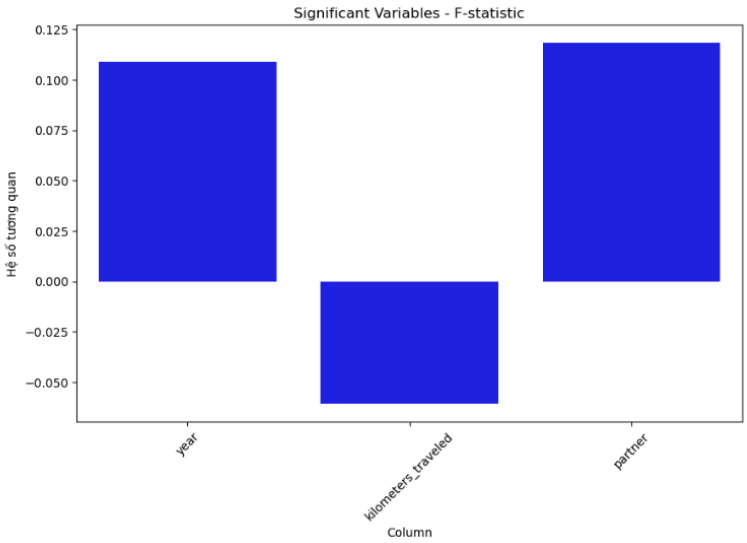
3.1 Đặt vấn đề và thu thập dữ liệu

Đề ra các yếu tố có thể ảnh hưởng đến giá xe cũ để lựa chọn các thuộc tính cần thu thập cho bộ dữ liệu. Bộ dữ liệu được thu thập hoàn toàn từ website Chợ Tốt [1] bằng cách thu thập tự động thông qua các công nghệ như Selenium [2] và Beautiful Soup [3].

3.2 Tiền xử lý dữ liệu

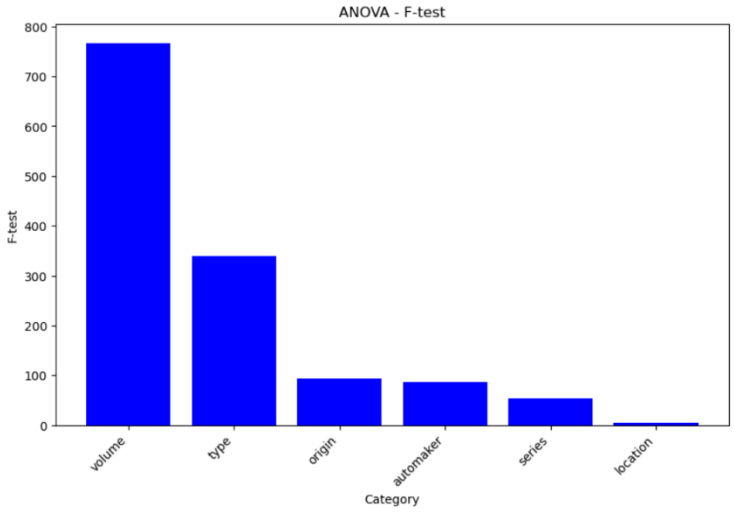
Để tăng độ chính xác cho mô hình, nhóm đã loại bỏ bớt các dòng dữ liệu bị khuyết. Đánh giá lại các kiểu dữ liệu phù hợp cho mỗi đặc trưng, lọc lại các dòng dữ liệu có giá không đúng với loại xe thực tế trên thị trường.

3.3 Phân tích thăm dò



Hình 2. Phân tích Correlation các biến số

Áp dụng phương pháp Correlation, em sử dụng bar để vẻ biểu đồ. Đường dọc thể hiện hệ số tương quan của các biến so với biến mục tiêu và ngang qua là danh sách các biến. Dựa vào biểu đồ ta có thể thấy hệ số tương quan của ba biến này gần về 0 nên không có sự



tương quan giữa 3 biến này với biến ‘price’. Do đó ta có thể kết luận rằng 3 biến này không ảnh hưởng đến giá - ‘price’ của xe.

Hình 3. Phân tích ANOVA các biến phân loại

Sử dụng barplot để vẽ biểu đồ, các biến trong biểu đồ đã được lọc F-value < 0.05. Trục dọc thể hiện giá trị của F-test, trục ngang là các biến phân loại có khả năng ảnh hưởng đến biến mục tiêu. Dựa vào biểu đồ ta có thể thấy F-test của 5 biến ‘volume’, ‘type’, ‘origin’, ‘automaker’, ‘series’ có F-test cao và có P-value < 0.05 nên có khả năng ảnh hưởng mạnh giữa các biến này so với biến mục tiêu - ‘price’. Biến ‘location’ có F-test nhỏ nên khả năng ảnh hưởng yếu so với biến mục tiêu - ‘price’.

Dựa trên 2 kết quả phân tích có 5 biến ‘volume’, ‘type’, ‘origin’, ‘automaker’, ‘series’ có ảnh hưởng đáng kể đến giá - ‘price’.

3.4 Lựa chọn mô hình

Với mong muốn lựa chọn được mô hình có kết quả tốt nhất, nhóm đã thực hiện mô phỏng sự phân bố giữa các biến ảnh hưởng đối với biến mục tiêu bằng biểu đồ barplot như sau:

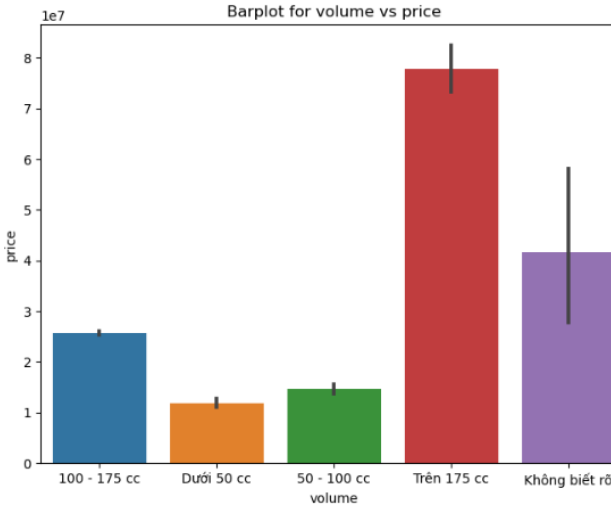
Biểu đồ cột (barplot) được sử dụng để thể hiện mối quan hệ giữa các biến ảnh hưởng và biến mục tiêu ‘Price’ . Dưới đây là mô tả ý nghĩa của biểu đồ:

Trục X (hoành độ): Trục này thể hiện giá trị của biến ảnh hưởng Mỗi cột trên trục X đại diện cho một giá trị cụ thể của biến đó.

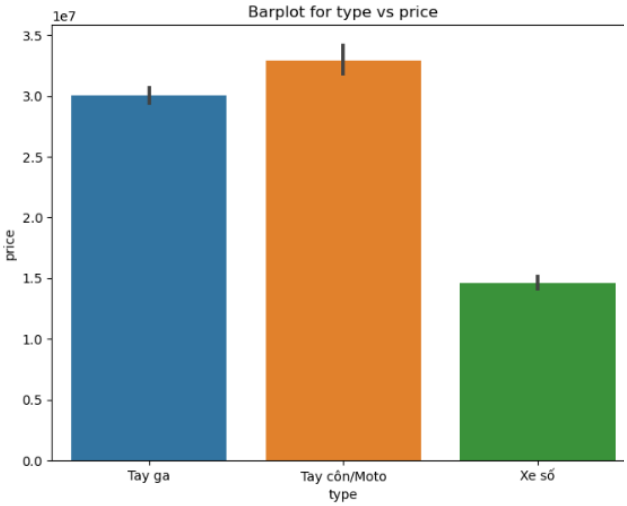
Trục Y (tung độ): Trục này thể hiện giá trị của biến mục tiêu ("Price"). Chiều cao của mỗi cột trên trục Y biểu thị giá trị tương ứng của "Price" dựa trên giá trị của biến ảnh hưởng tương ứng.

Các cột: Mỗi cột trên biểu đồ đại diện cho một giá trị cụ thể của biến ảnh hưởng. Chiều cao của cột biểu thị giá trị trung bình hoặc tổng của biến mục tiêu ("price") tương ứng với giá trị của biến ảnh hưởng đó.

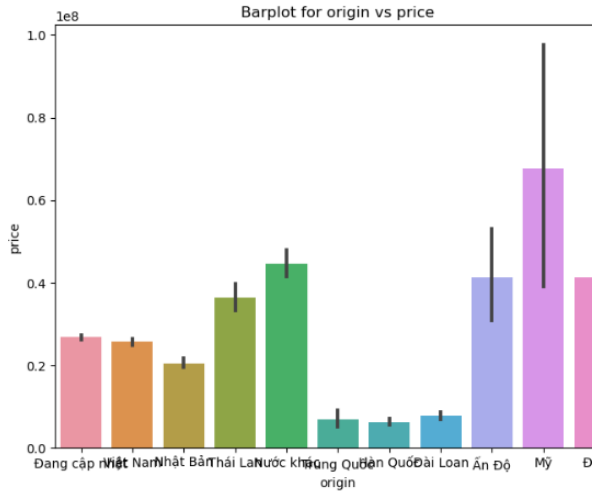
Đoạn thẳng ngắn trên đỉnh đầu của các cột thể hiện sử biến động của dữ liệu. Nếu đoạn thẳng ngắn thì dữ liệu ổn ddingj ít biến động, nếu đoạn thẳng dài thì dữ liệu có sự biến động lớn.



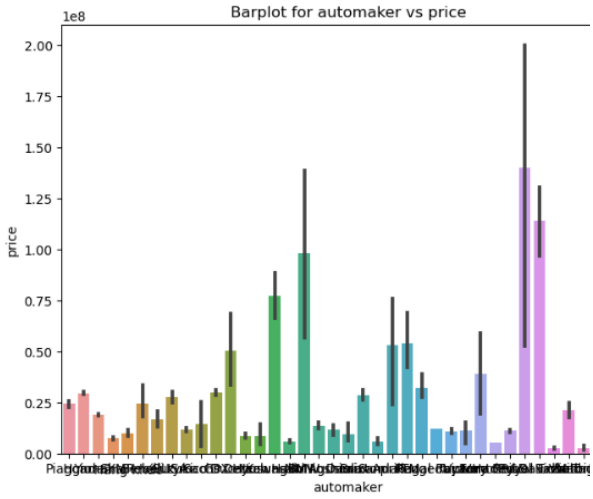
Hình 4. Biểu đồ thể hiện sự phân bố của Volume và Price



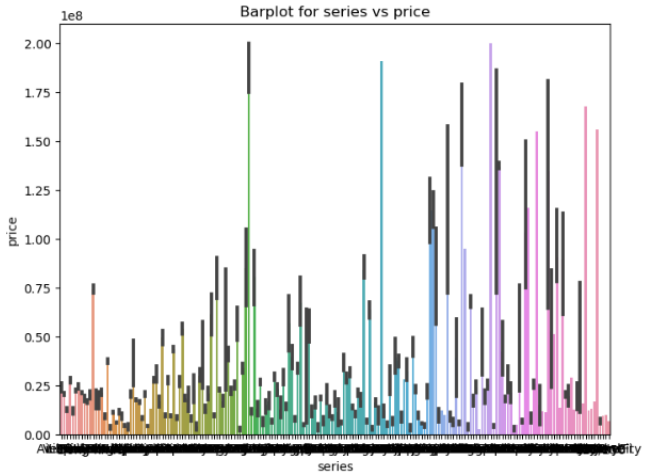
Hình 5. Biểu đồ thể hiện sự phân bố của Type và price



Hình 6. Biểu đồ thể hiện sự phân bố của Origin và Price



Hình 7. Biểu đồ thể hiện sự phân bố của Automaker và Price



Hình 8. Biểu đồ thể hiện sự phân bố của Seriesvà Price

Sau khi mô phỏng dữ liệu, nhóm nhận thấy rằng dữ liệu được phân bố rất phức tạp, các biến phân loại rất đa dạng nhưng lại được biểu diễn một cách rõ ràng. Xem xét các đặc điểm trên của bộ dữ liệu nhóm đã quyết định thực hiện các mô hình phù hợp nhất như Random Forest, Liner Regression, K-Nearest Neighbors Algorithm và Decicion Tree để đảm bảo tìm được mô hình cho kết quả tốt nhất.

3.5 Đánh giá mô hình

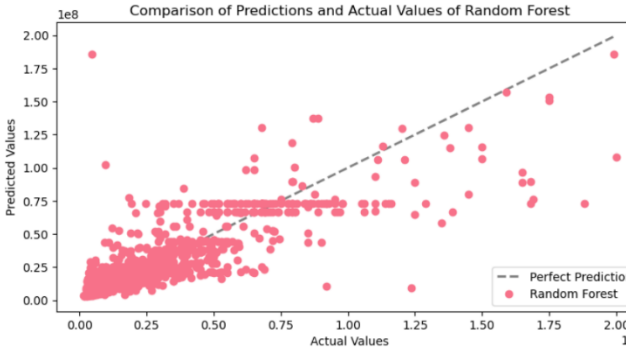
Nhóm lựa chọn đánh giá bằng các thang đo: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE) và R^2. MAE cung cấp một cái nhìn trực tiếp và dễ hiểu về sai số trung bình tuyệt đối trong dự báo, giúp nhận biết mức độ chính xác tổng thể của mô hình. RAE được sử dụng để đánh giá hiệu suất của mô hình so với một mô hình dự báo cơ bản. Điều này cho phép chúng ta hiểu rõ hơn về hiệu suất tương đối của mô hình trong bối cảnh cụ thể, so sánh hiệu quả của mô hình so với những phương pháp dự báo đơn giản hoặc truyền thống. RMSE là thang đo quan trọng bởi vì nó đặc biệt nhạy cảm với những sai số lớn, giúp nhận diện và đánh giá ảnh hưởng của chúng đối với mô hình. R^2 là một thống kê được sử dụng để đánh giá mức độ phù hợp (fit) của một mô hình hồi quy với dữ liệu quan sát. Nó là tỷ lệ phần trăm của tổng biến thiên của biến phụ thuộc được giải thích bởi mô hình. Adjusted R^2, là một biến thể của R^2 thường được sử dụng trong hồi quy tuyến tính để đánh giá mức độ phù hợp của mô hình. Adjusted R^2 giúp

khắc phục vấn đề overfitting mà R^2 thông thường có thể gặp phải trong quá trình đánh giá mô hình hồi quy.

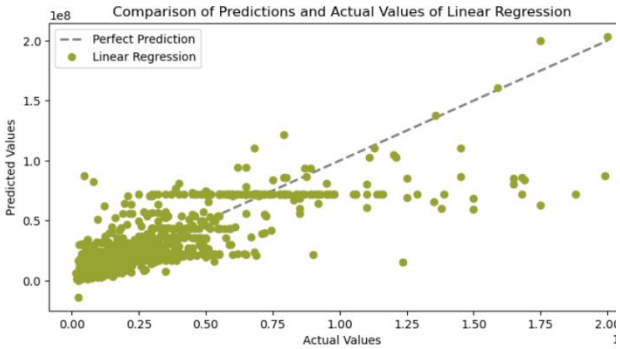
3.6 Deloy model

Tiến hành triển khai mô hình vào thực tế khi kết quả mô hình đạt được độ chính xác phù hợp nhất đựa trên kết quả phân tích các thang đo đánh giá.

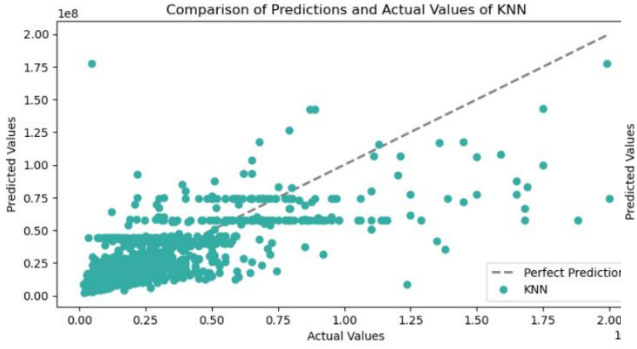
4 KẾT QUẢ PHÂN TÍCH



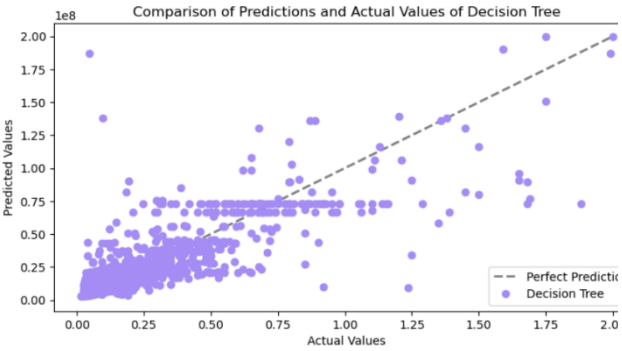
Hình 9. Scatter plots so sánh giá thực tế và dự đoán của Random Forest



Hình 10. Scatter plots so sánh giá thực tế và dự đoán của Linear Regression



Hình 11. Scatter plots so sánh giá thực tế và dự đoán của KNN

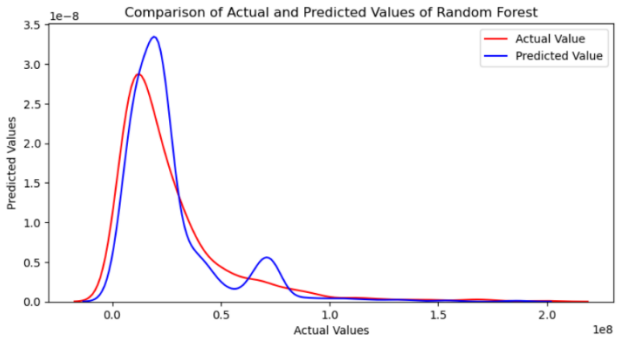


Hình 12. Scatter plots so sánh giá thực tế và dự đoán của Decision Tree

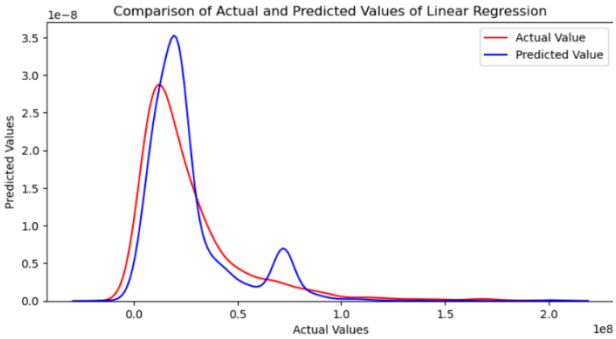
Các biểu đồ phân tán được cung cấp đánh giá hiệu suất của bốn mô hình học máy khác nhau (Random Forest, Linear Regression, KNN, và Decision Tree) trong việc dự đoán giá xe máy cũ. Trong mỗi biểu đồ, giá thực tế được biểu diễn trên trục hoành, trong khi giá dự đoán của mô hình được biểu diễn trên trục tung. Đường chấm đứt mô tả dự đoán hoàn hảo, nơi giá dự đoán trùng khớp với giá thực tế.

Từ phân tích các biểu đồ, có thể thấy rằng mô hình Random Forest cho kết quả khá chính xác, với nhiều điểm dữ liệu gần với đường dự đoán hoàn hảo, đặc biệt ở phần giá thấp và trung bình. Trong khi đó, mô hình hồi quy tuyến tính có hiệu suất tốt ở phần giá thấp nhưng không duy trì được độ chính xác này ở các giá trị cao hơn. Mô hình KNN dường như không phù hợp với dữ liệu này, với sự phân tán rộng của các dự đoán xung quanh đường hoàn hảo. Mô hình Decision Tree thể hiện một số độ chính xác nhưng không nhất quán trên toàn bộ phạm vi giá.

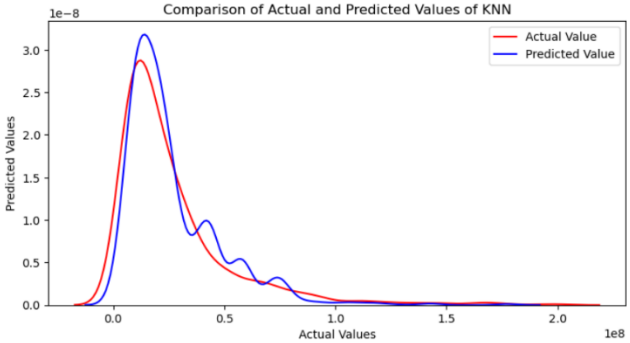
Dựa trên những quan sát này, có thể kết luận rằng mô hình Random Forest có vẻ như là mô hình tối ưu nhất trong số các mô hình đã được kiểm tra. Nó cung cấp sự cân bằng tốt giữa độ chính xác và khả năng tổng quát hóa trên các phạm vi giá của xe máy.



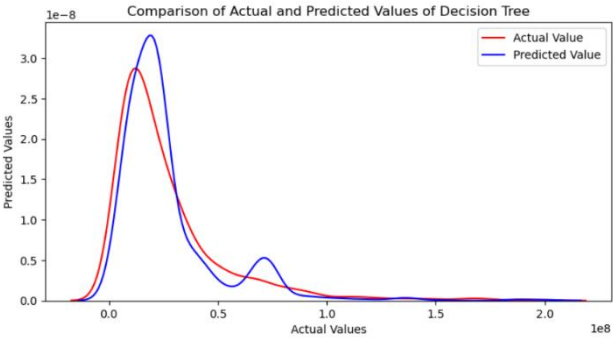
Hình 13. KDE plot biểu diễn phân phối giá thực tế và dự đoán của Random Forest



Hình 14. KDE plot biểu diễn phân phối giá thực tế và dự đoán của Linear Regression



Hình 15. KDE plot biểu diễn phân phối giá thực tế và dự đoán của KNN



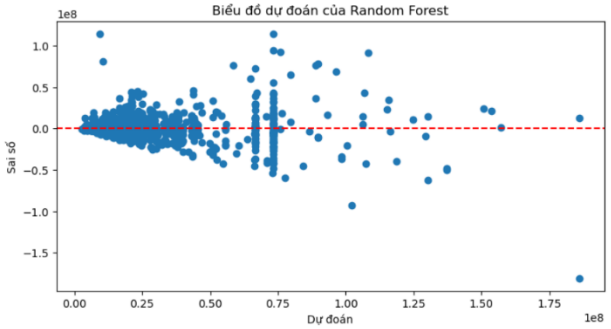
Hình 16. KDE plot biểu diễn phân phối giá thực tế và dự đoán của Decision Tree

Các biểu đồ Kernel Density Estimate (KDE) trên mô tả phân phối của giá thực tế so với giá dự đoán từ bốn mô hình Random Forest, Linear Regression, KNN và Decision Tree. Mỗi biểu đồ cung cấp một cái nhìn trực quan về mức độ mô hình dự đoán phù hợp

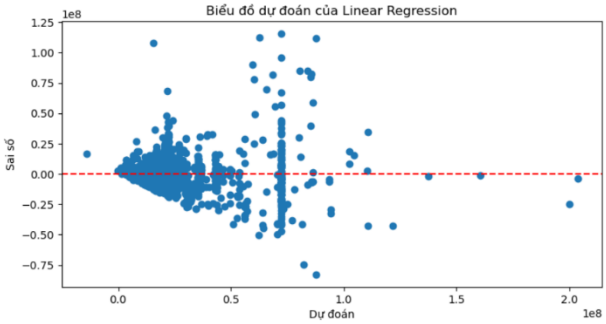
với dữ liệu thực tế thông qua việc so sánh hai đường KDE: một cho giá thực tế (màu đỏ) và một cho giá dự đoán (màu xanh).

Từ các biểu đồ cho thấy các mô hình đều cho ra kết quả khá giống nhau nhưng vẫn có vài điểm khác biệt. KNN thì phía bên sườn dốc có nhiều lượn sóng hơn các mô hình còn lại cho thấy nó kém hiệu quả hơn. Các mô hình khác đều có dạng giống như hai quả núi nhưng của Linear Regression thì cao hơn một chút ở đỉnh nhỏ cho thấy mô hình đôi khi cho giá cao hơn thực tế một chút. Random Forest và Decision Tree có hình dạng giống nhau và chênh lệch không nhiều so với giá gốc.

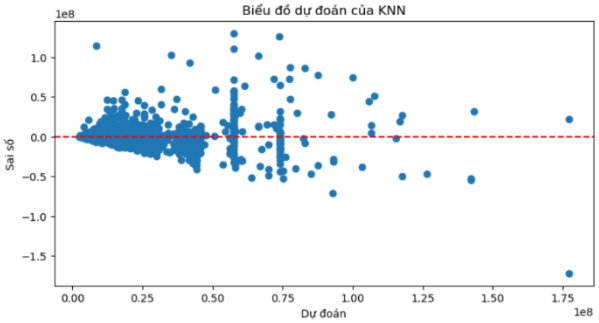
Từ các biểu đồ KDE, Random Forest và Decision Tree là hai mô hình cho kết quả gần với thực tế nhất trong các mô hình được chọn.



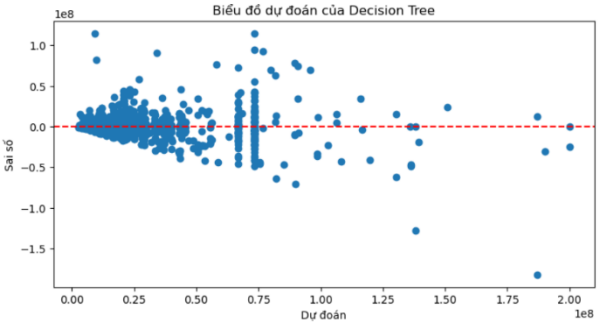
Hình 17. Residual scatter plot biểu diễn sai số giữa giá thực tế và dự đoán của Random Forest



Hình 18. Residual scatter plot biểu diễn sai số giữa giá thực tế và dự đoán của Linear Regression



Hình 19. Residual scatter plot biểu diễn sai số giữa giá thực tế và dự đoán của KNN



Hình 20. Residual scatter plot biểu diễn sai số giữa giá thực tế và dự đoán của Decision Tree

Biểu đồ phân tán thể hiện sai số - tức là sự chênh lệch giữa giá dự đoán và giá thực tế của xe máy cũ từ bốn mô hình Random Forest, Linear Regression, KNN và Decision Tree. Điểm nằm chính giữa dọc theo đường màu đỏ ngang (đường sai số bằng 0) biểu thị những

dự đoán chính xác hoàn hảo, trong khi các điểm xa đường này biểu thị dự đoán có sai số lớn hơn.

Random Forest có nhiều điểm tập trung gần đường ngang màu đỏ, cho thấy rằng mô hình này có nhiều dự đoán chính xác. Các điểm phân tán ra xung quanh chỉ ra một số dự đoán có sai số lớn. Linear Regression có các điểm phân tán rộng lớn hơn so với Random Forest, điều này cho thấy mô hình này có sai số cao hơn trong dự đoán. Các điểm của KNN có vẻ tập trung ở một vùng nhất định nhưng cũng phân tán khá đều qua lại, cho thấy mô hình này vẫn có sai số đáng kể. Decision Tree thì giống như Random Forest, có nhiều điểm gần đường ngang màu đỏ, nhưng cũng có những điểm sai lệch lớn, thể hiện sự không ổn định trong dự đoán.

Từ các biểu đồ, ta có thể thấy các điểm tập trung gần đường sai số bằng 0 của Random Forest và Decision Tree nhiều hơn hai mô hình còn lại. Chứng tỏ chúng có mức độ dự đoán chính xác hơn hai mô hình còn lại.

	Model	MAE	RAE	RMSE	R^2	Adjusted R^2
0	Random Forest	8669311.15777876	0.4834772595333917	15224198.12776448	0.67696931	0.60937718
1	Linear Regression	9256188.34958316	0.5162067084148807	15434337.47932254	0.66799019	0.59851923
2	KNN	10118720.91819444	0.5643091325799547	17558567.17412219	0.57031215	0.48040267
3	Decision Tree	8733885.04201585	0.4870784689051836	15574738.09409830	0.66192237	0.59118176

Bảng 3. Kết quả đánh giá của các thang đo

Kết quả thống kê cho thấy mô hình Random Forest xuất hiện với cả ba chỉ số đánh giá MAE, RAE, và RMSE tốt nhất so với các mô hình còn lại. Và khi đánh giá mức độ phù hợp thì Random Forest vẫn cho kết quả cao nhất. Random Forest không chỉ cung cấp dự báo gần với giá trị thực tế nhất (theo MAE) mà còn cho thấy khả năng dự báo tốt hơn mô hình cơ bản (theo RAE) và ít bị ảnh hưởng bởi các sai số lớn (theo RMSE). Trong khi đó, Linear Regression và Decision Tree có kết quả khá giống nhau về RAE, đồng thời Linear Regression lép vế hơn một chút với MAE và RMSE. KNN, ngược lại, thể hiện hiệu suất kém nhất với cả ba chỉ số, có thể là dấu hiệu rằng mô hình này không phù hợp hoặc cần được điều chỉnh thêm. Khi xét về mức độ phù hợp của mô hình thì Random Forest cho kết quả tốt nhất (theo R2 và Adjusted R2) theo sau là Linear Regression và Decision Tree. Vì

vậy, Random Forest không chỉ cho thấy sự ưu việt thông qua các con số mà còn qua việc cân bằng giữa độ chính xác và sự nhạy cảm với các sai số, làm cho nó trở thành lựa chọn tốt nhất trong số các mô hình được xem xét.

5 KẾT LUẬN

Sau khi phân tích cho thấy một số biến số, biến phân loại có khả năng ảnh hưởng đến biến mục tiêu trong bộ dữ liệu. Mô phỏng mô hình barplot nhận xét dữ liệu, xây dựng pipeline tiền xử lý, mã hóa nhị phân các biến phân loại. Sau quá trình phân tích chi tiết và xây dựng mô hình, nhóm đã xác định rằng Random Forest là lựa chọn ưu việt nhất cho bộ dữ liệu hiện tại. Kết quả đánh giá chi tiết của các mô hình đã chọn mô tả một hiệu suất xuất sắc của Random Forest, đặc biệt là trong việc dự đoán và xử lý biến đổi của dữ liệu. Tương lai, nhóm sẽ tiến hành thử nghiệm thêm các mô hình khác như Gradient Boosting, Support Vector Machines, hoặc Neural Networks có thể mang lại cái nhìn toàn diện về sự phù hợp của các mô hình khác nhau với dữ liệu.

TÀI LIỆU THAM KHẢO

[1] Chợ tốt. Địa chỉ: https://www.chotot.com/. [Truy cập lần cuối 10/12/2023]

[2] "The Selenium Browser Automation Project," 2023. [Trực tuyến]. Địa chỉ: https://www.selenium.dev/documentation/. [Truy cập lần cuối 20/11/2023].

[3] "Beautiful Soup Documentation," 2015. [Trực tuyến]. Địa chỉ: https://beautiful-soup-4.readthedocs.io/en/latest/. [Truy cập lần cuối 20/11/2023].

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Kiều Minh Phước	• Thu thập và tiền xử lý bộ dữ liệu. • Phân chia công việc và các giai đoạn làm đồ án. • Tổng kết file báo cáo cuối cùng (định dạng, viết trích dẫn, phụ lục,..) • Tổng hợp code.
2	Nguyễn Thành Long	• Viết giới thiệu. • Mô tả bộ dữ liệu. • Kết luận. • Thuyết trình.
3	Lưu Thượng Vỹ	• Phân tích thăm dò bộ dữ liệu. • Trực quan hóa và viết đánh giá kết quả thăm dò. • Làm slides.
4	Trương Minh Phong	• Lựa chọn các mô hình khả thi. • Hiện thực các mô hình. • Thuyết trình
5	Lâm Quốc Đạt	• Lựa chọn độ do phù hợp. • Thực hiện đánh giá các mô hình. • Viết đánh giá mô hình qua các thang đo. • Thuyết trình.