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

□□□□□

DỰ ĐOÁN GIÁ ĐỒNG HỒ CASIO

Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Trần Nguyên Mẫn	21522325
2	Phan Lê Chí Trung	21522725
3	Nguyễn Đạt Tuấn	21522754

TP. HỒ CHÍ MINH – 12/2023

1. GIỚI THIỆU

Đề tài tập trung vào việc dự đoán giá của đồng hồ Casio. Mục tiêu chính là phân tích các yếu tố ảnh hưởng đến giá cả của các mẫu đồng hồ Casio và xây dựng một mô hình dự đoán giá dựa trên các thông tin đặc trưng. Bộ dữ liệu và đề tài do nhóm tự phân tích thiết kế và không dựa trên đề tài nào khác.

Dựa trên việc phân tích và xử lý dữ liệu, nhóm đã xây dựng được một mô hình dự đoán giá đồng hồ Casio với độ chính xác đáng kể. Kết quả có thể giúp người tiêu dùng, nhà sản xuất, và nhà đầu tư có cái nhìn tổng quan về giá cả và các yếu tố quyết định giá của sản phẩm này.

Bộ dữ liệu được nhóm tự thu thập từ các nguồn đáng tin cậy như trang thương mại điện tử Sendo, chuỗi cửa hàng Thế Giới Di Động và cửa hàng đồng hồ WatchStore (đại lý ủy quyền của nhiều thương hiệu nổi tiếng thế giới).

2. MÔ TẢ BỘ DỮ LIỆU

Với yêu cầu thao tác trực tiếp với trang web như click chuyển trang và điền tên đồng hồ vào khung tìm kiếm,... nhóm lựa chọn Selenium làm công cụ để thu thập dữ liệu từ các website https://www.sendo.vn/, https://www.thegioididong.com/ và https://www.watchstore.vn/.

Bộ dữ liệu thu thập được sau khi gộp và chọn lọc thuộc tính có 3533 mẫu, bao gồm 12 thuộc tính cơ bản liên quan đến thông số kỹ thuật, thiết kế, thương hiệu trong đó có 8 biến phân loại và 4 biến số. Sau đây là bảng thống kê một số đặc trưng cơ bản của các thuộc tính của bộ dữ liệu:

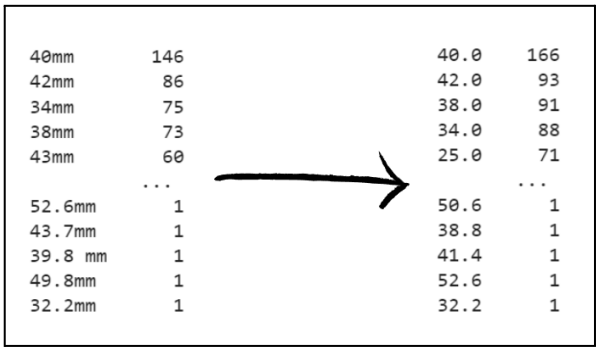
Thuộc tính	Kiểu biến (raw)	Kiểu biến thực	Số giá trị khuyết	Mô tả
watch_name	object	object	0	Tên của đồng hồ
price	object	float	0	Giá của đồng hồ
watch_type	object	object	388	Dòng máy
user	object	object	350	Đối tượng sử dụng
watch_dia	object	float	276	Đường kính mặt đồng hồ
glass_material	object	object	77	Chất liệu mặt kính
bracelet_material	object	object	64	Chất liệu dây
waterproof	object	int	128	Độ kháng nước
origin	object	object	532	Xuất xứ
watch_thickness	object	float	1481	Độ dày
website	object	object	0	Tên website bán
watch_shape	object	object	346	Hình dạng mặt kính

3. TIỀN XỬ LÝ DỮ LIỆU

3.1. Định dạng dữ liệu các thuộc tính

3.1.1. Nhóm 1: Chuẩn hóa/cập nhật kiểu dữ liệu các biến

- Đối với các thuộc tính đo lường có kiểu số thực float, khi crawl dữ liệu từ các web, các số được đi kèm với đơn vị đo nên ở dạng object.
- Nhóm này bao gồm các thuộc tính: 'watch_dia', 'waterproof', 'watch_thickness', 'price'.
- Ví dụ: Với thuộc tính 'watch_dia' (đường kính mặt kính): dữ liệu raw gồm các giá trị như: '40mm', '42mm' và '42 mm' (có thêm dấu khoảng trắng ' '). Với thuộc tính 'waterproof' (Độ kháng nước) thì gồm các giá trị như: '20atm', '5 ATM', '10 ATM - Tắm, bơi' hay 'Không chống nước'. Nhóm đã bỏ các đơn vị và chuyển thuộc tính này thành biến số.

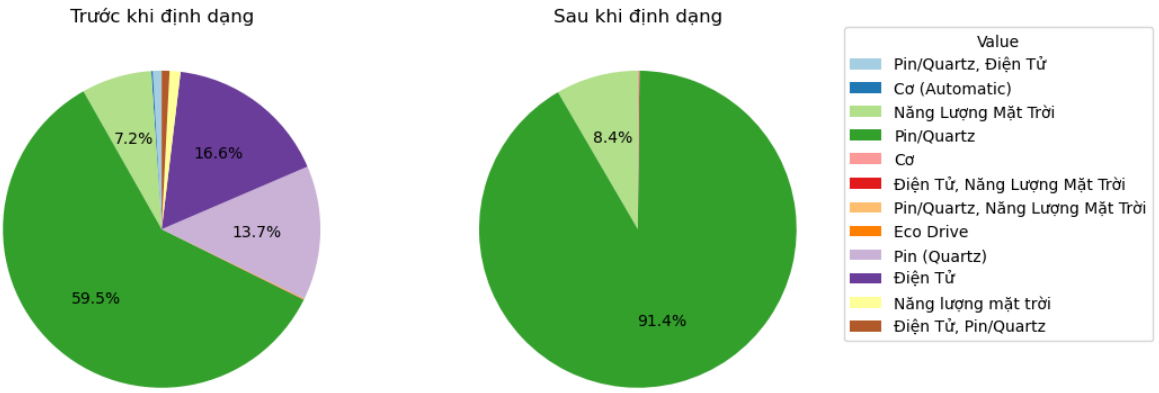


Hình 1. Danh sách giá trị của biến 'watch_dia' trước và sau khi chuẩn hóa

3.1.2. Nhóm 2: Ánh xạ thủ công các biến có giá trị không nhất quán

- Dữ liệu được thu thập trên 3 trang web khác nhau, nên xảy ra trường hợp không nhất quán. Vì vậy cần tiến hành chuyển đổi dữ liệu về một tiêu chuẩn chung.
- Nhóm này bao gồm các thuộc tính: 'watch_type', 'bracelet_material', 'glass_material' và 'origin'.
- Ví dụ: Với đồng hồ chạy bằng pin trong trang web *sendo* và *thegioididong* được lưu trữ là **Pin (Quartz)** nhưng trong *watchstore* lại được lưu là **Pin/Quartz** (thuộc tính 'watch_type').

  Thuộc tính 'watch_type': sau khi định dạng xong các loại đồng hồ, tổng cộng có 3 loại đồng hồ được phân loại, trong đó 'Pin/Quarts' chiếm số lượng lớn nhất (91.4%), 'Năng Lượng Mặt Trời' là 8.4% và thấp nhất là 'Cơ' (0.2%)



Hình 2. Biểu đồ tỉ lệ các nhãn của biến 'watch_type' trước và sau khi định dạng

3.1.3. Nhóm 3: Tạo biến mới

- Tạo thuộc tính Mã đồng hồ 'watch_sku' từ 'watch_name':

+ Thuộc tính 'watch_name' (tên đồng hồ): Đối với thuộc tính 'watch_name' thì nhóm chuẩn hóa về dạng 'Đồng hồ Casio SKU-UPC-MPN' (mã sản phẩm/số sản phẩm/nhà sản xuất) cho nhất quán, cũng như để thuận tiện trích xuất, tạo biến mới Mã đồng hồ 'watch_sku'. Đồng thời loại bỏ các đồng hồ không phải của thương hiệu Casio.

+ Mỗi mã đồng hồ thường có những đặc trưng chung, có thể dùng những đặc trưng này để điền giá trị khuyết cho những đồng hồ có chung mã.

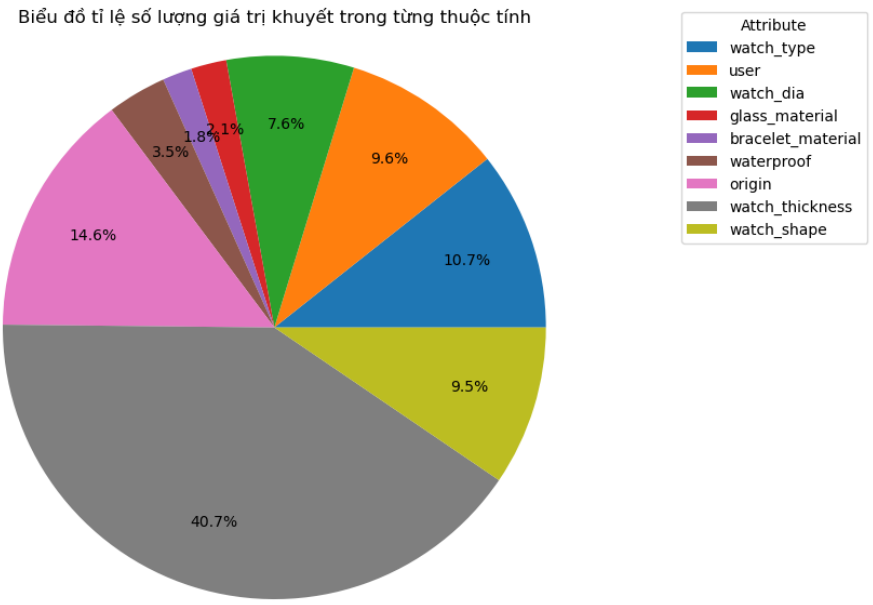
3.2. Xử lý các giá trị bị khuyết và trùng lặp

3.2.1. Tổng quan:

- Tổng cộng có 3642 dữ liệu bị khuyết, trong đó tỷ lệ giá trị khuyết trong mỗi thuộc tính được phân bổ khác nhau. Đặc biệt thuộc tính 'watch_thickness' có tỉ lệ cao nhất (chiếm 40.7% tổng số giá trị bị khuyết). (Hình 3)
- Và có tổng cộng 870 hàng trùng lặp.

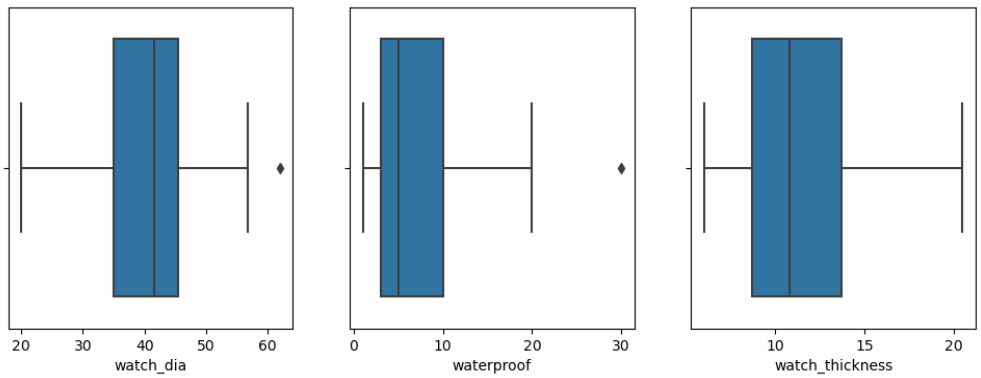
3.2.2. Xử lý các giá trị khuyết

- Bước đầu, nhóm loại bỏ các mẫu dữ liệu có từ 5 giá trị NaN trở lên.
- Các loại đồng hồ có chung mã đồng hồ (watch_sku) thường có chung một số đặc trưng cụ thể:



Hình 3. Biểu đồ tỉ lệ số lượng giá trị khuyết trong từng thuộc tính

- Đối với các biến phân loại: bộ dữ liệu có tổng cộng 9 biến phân loại, dùng giá trị phổ biến nhất (mode) từ các dòng dữ liệu không bị khuyết đã gộp từ mã đồng hồ (watch_sku) để thay thế.
- Đối với các biến số: có tổng cộng 4 biến số, bộ dữ liệu có khá ít giá trị ngoại lệ (Hình 4), nhóm đã quyết định điền giá trị trung bình (mean) cho những dữ liệu bị khuyết.



Hình 4. Biểu đồ phân phối giá trị trong các biến số

- Các giá trị khuyết còn lại: Sau thực hiện điền giá trị khuyết theo mã đồng hồ ‘watch_sku’:
- Các biến số còn lại được điền khuyết bằng giá trị trung bình.

- Biến phân loại ‘origin’ được điền mặc định bằng ‘Nhật’. Vì Casio là thương hiệu đồng hồ của Nhật, hầu hết các đồng hồ Casio sẽ được sản xuất và lắp ráp tại Nhật.
- Biến phân loại ‘bracelet_material’, sẽ loại bỏ các mẫu chứa giá trị khuyết (gồm 6 mẫu), bởi vì số lượng khá ít, và các đồng hồ này có thông số tương tự nhau, chỉ khác giá hoặc giống.

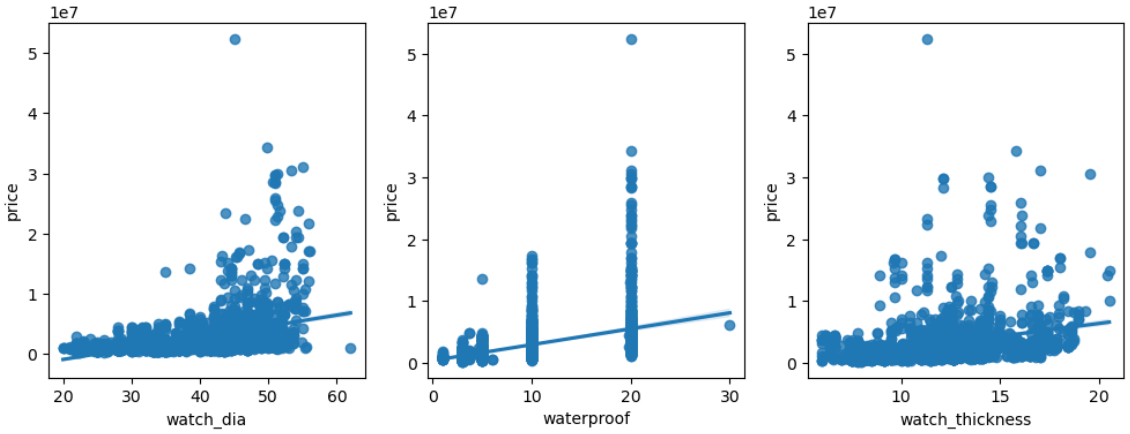
3.2.3. Xử lý các giá trị trùng lặp:

- Đối với các đồng hồ trùng thông số, website và giá tiền: Loại bỏ mẫu trùng lặp, giữ lại mẫu đầu tiên.
- Đối với các đồng hồ trùng thông số, website nhưng khác giá tiền: Cập nhật giá tiền lại bằng giá trị trung bình của giá và loại bỏ các bộ dữ liệu cũ.
- Đối với các đồng hồ trùng thông số, khác website: Giữ nguyên mẫu dữ liệu.

4. PHÂN TÍCH, THĂM DÒ DỮ LIỆU

4.1. Biến số

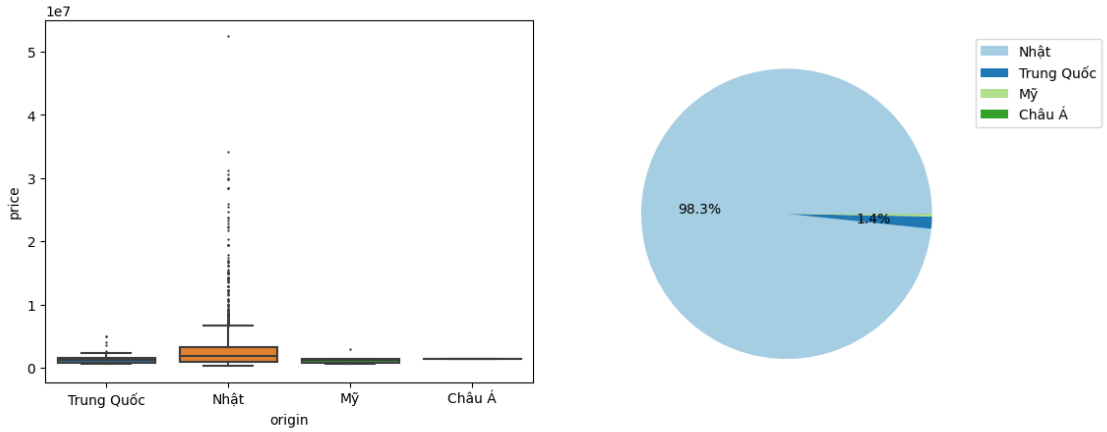
- Sử dụng độ đo pearsonr của spicy.stats với mức ý nghĩa 95% (p_value < 0.05), kiểm tra sự ảnh hưởng của các biến số với biến mục tiêu ‘price’. Cho thấy các biến số đều có thể có ảnh hưởng tương quan tuyến tính với biến mục tiêu tuy không cao. Và sử dụng biểu đồ Regplot để kiểm chứng lại kết quả (Hình 5)



Hình 5. Biểu đồ Regplot thể hiện sự tương quan các biến số với biến ‘price’

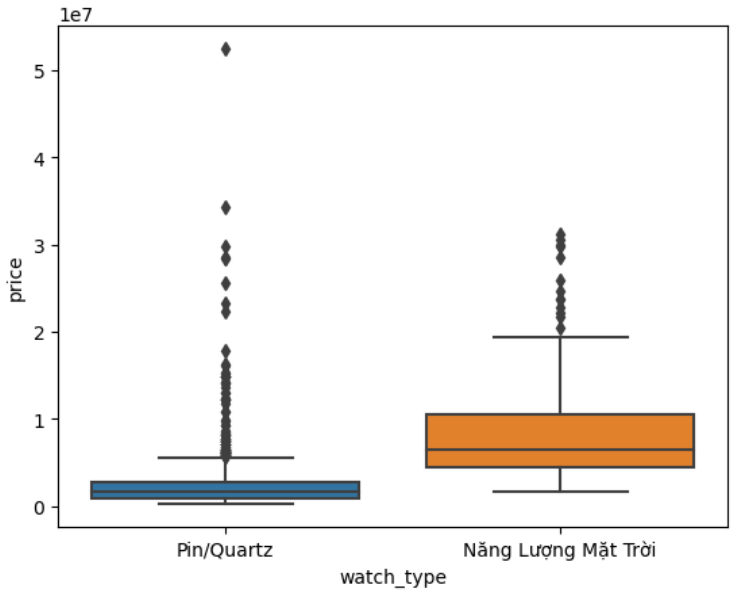
4.2. Biến phân loại

- Sử dụng f_oneway của scipy.stats để đánh giá sự tương quan của các biến với biến mục tiêu, ngoài thuộc tính xuất xứ ‘origin’ các thuộc tính còn lại đều có ảnh hưởng đến giá đồng hồ ‘price’.
- Thuộc tính ‘origin’ ảnh hưởng rất yếu đến giá đồng hồ, vì hầu hết đồng hồ Casio đều của ‘Nhật’, các khu vực khác chỉ chiếm 1.7%.



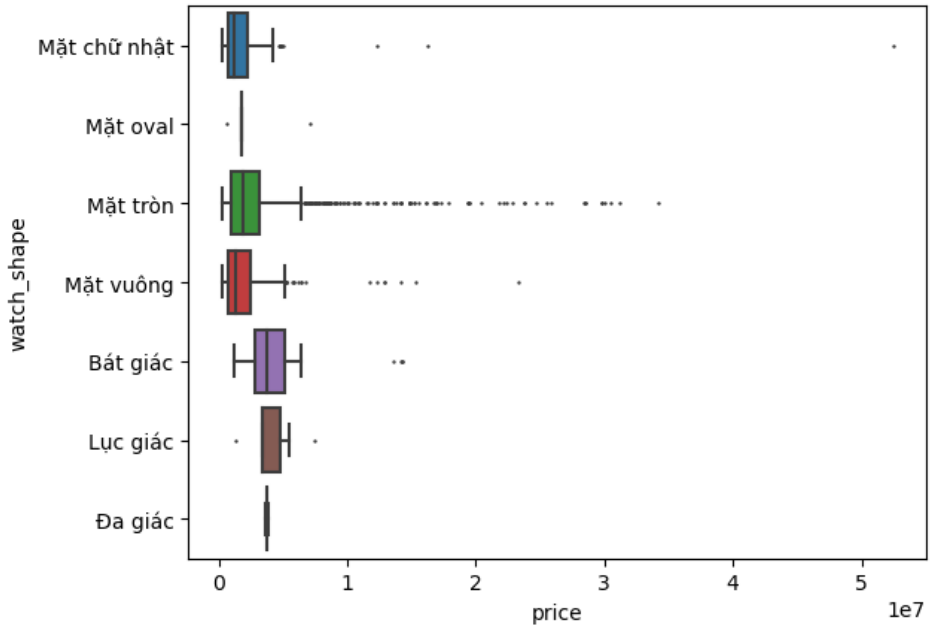
Hình 7. Biểu đồ Boxplot, Pie thể hiện phân bố và tỉ lệ giá trị của ‘origin’

- Thuộc tính dòng máy ‘watch_type’ thực sự có ảnh hưởng đến giá trị của đồng hồ, điều này phù hợp với thực tế vì đồng hồ năng lượng mặt trời tích hợp công nghệ và vật liệu phức tạp hơn so với đồng hồ pin.



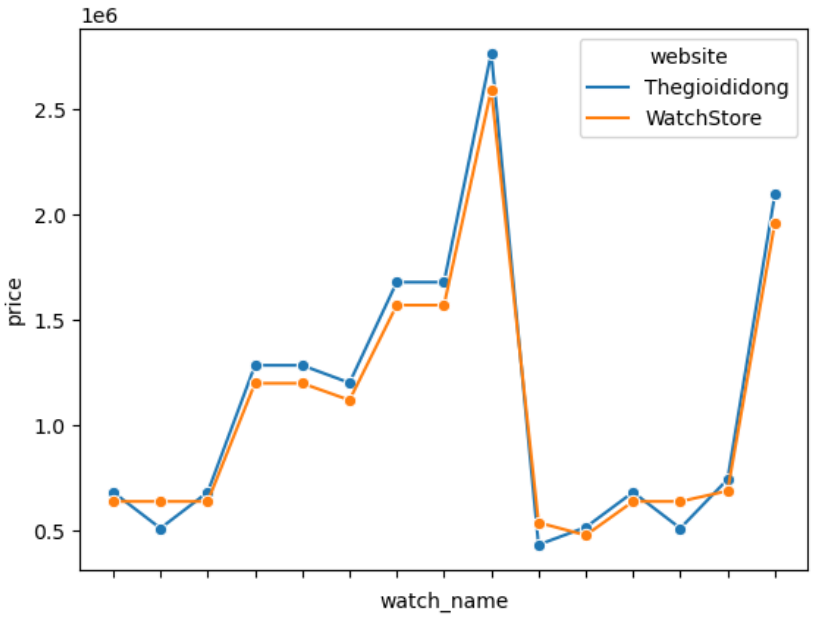
Hình 8. Biểu đồ Boxplot thể hiện phân bố giá trị của thuộc tính ‘watch_type’

- Tương tự, đồng hồ mặt chữ nhật, oval, tròn hay vuông đều là những đồng hồ cơ bản nên thường có giá thấp hơn so với đồng hồ mặt bát giác, lục giác hay đa giác. Nhưng có nhiều giá trị outlier của đồng hồ mặt tròn có giá rất cao, điều này cũng hợp lí, vì trên thực tế, các đồng hồ sang trọng thường có mặt tròn.



Hình 9. Biểu đồ Boxplot thể hiện phân bố giá trị của thuộc tính 'watch_shape'

- Đối với cùng đồng hồ, cùng thông số thì ở Thegioididong thường có giá nhỉnh hơn so với WatchStore. (Các đồng hồ ở Sendo không trùng với 2 web còn lại)



Hình 10. Biểu đồ lineplot thể hiện giá cả của các đồng hồ theo từng 'website'

5. CÀI ĐẶT MÔ HÌNH

Nhóm đã cài đặt Pipeline để triển khai mô hình thực thi tự động qua các bước:

1. **Mã hóa One-hot Encoder:** mã hóa dữ liệu các biến phân loại và chuyển chúng về dạng vector One-hot để huấn luyện mô hình.
2. **Tạo ra đặc trưng đa thức PolynomialFeatures:** chuyển đổi vector đặc trưng đầu vào thành một ma trận chứa tất cả các đặc trưng đa thức từ bậc 0 đến một bậc đã cho.

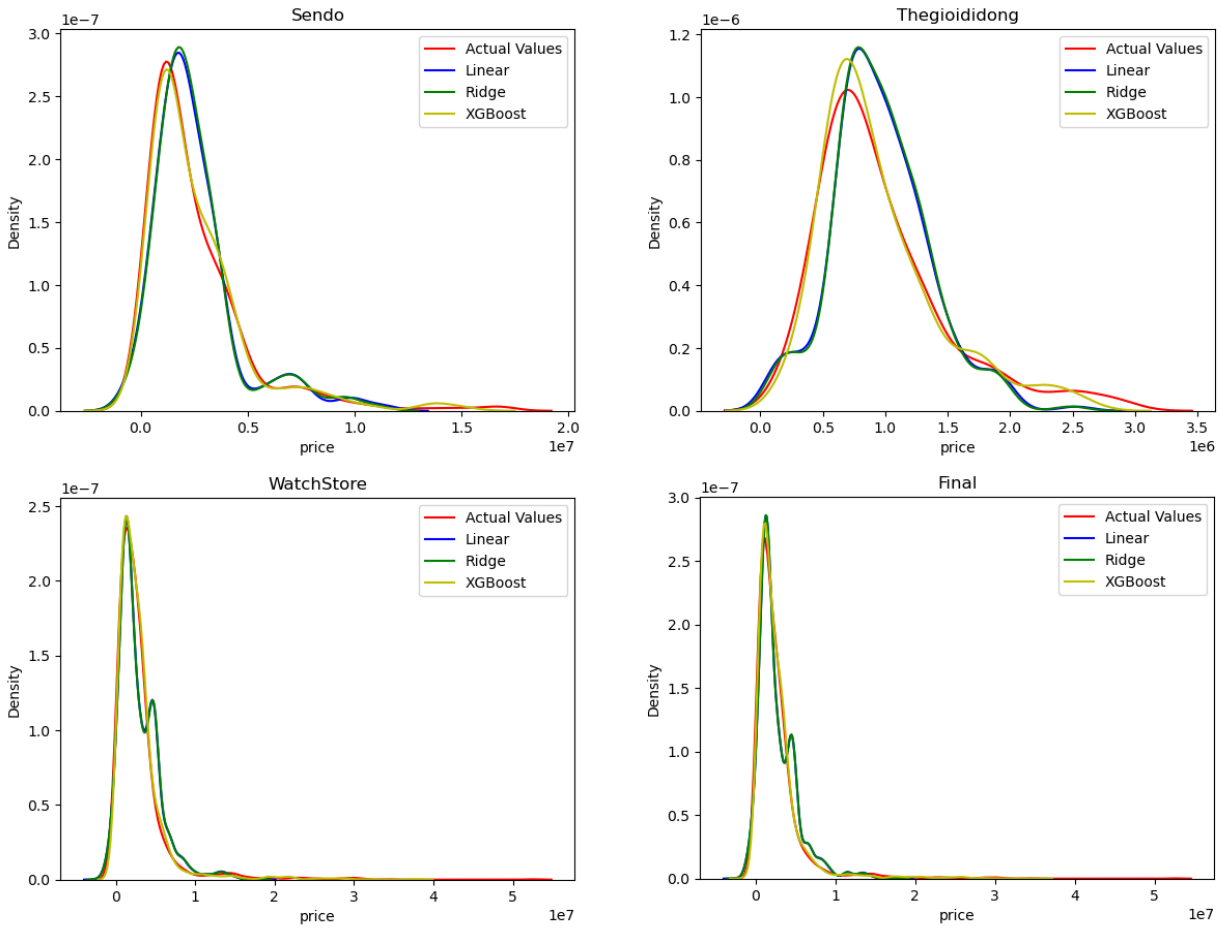
3. **Mô hình sử dụng:**
    a. Linear Regression: là phương pháp trong thống kê để mô hình hóa mối quan hệ tuyến tính giữa biến độc lập và biến phụ thuộc. Nó dự đoán giá trị dự báo bằng cách tối thiểu hóa sai số giữa dự đoán và dữ liệu thực tế, sử dụng đường thẳng tuyến tính.
    b. Ridge: là phương pháp trong máy học để giảm overfitting trong mô hình hồi quy tuyến tính bằng cách thêm một thành phần giảm số liệu trọng số. Nó kiểm soát độ lớn của các hệ số, giúp cải thiện hiệu suất và ổn định mô hình.
    c. XGBoost Regressor: là thuật toán máy học độc lập mạnh mẽ dùng trong dự đoán giá trị số. Nó kết hợp nhiều cây quyết định để tạo ra mô hình ổn định và hiệu quả, với khả năng xử lý overfitting và tối ưu hóa hàm mất mát.

6. HUẤN LUYỆN VÀ ĐÁNH GIÁ MÔ HÌNH

- Sau khi huấn luyện và đánh giá mô hình bằng 2 độ đo Mean Squared Error và R-Squared trên từng bộ dữ liệu của từng website và toàn tập dữ liệu, thu được kết quả:

MSE	LinearRegression	RidgeRegression	XGBoostRegression
Sendo	2.457e+12	2.350e+12	0.992e+12
Thegioididong	0.153e+12	0.153e+12	0.104e+12
WatchStore	7.122e+12	7.103e+12	4.342e+12
Final	5.891e+12	5.871e+12	4.029e+12

R2-score	LinearRegression	RidgeRegression	XGBoostRegression
Sendo	0.6041	0.6212	0.8401
Thegioididong	0.4508	0.4515	0.6250
WatchStore	0.4773	0.4787	0.6813
Final	0.4981	0.4998	0.6567



Hình 11. Biểu đồ distplot so sánh giá thực tế và dự đoán của các mô hình trên từng tập dữ liệu

- Mô hình dự đoán ở mức tương đối, có thể do bộ dự liệu chưa có nhiều thuộc tính chủ chốt có độ ảnh hưởng lớn đến giá đồng hồ.
- Bộ dữ liệu WatchStore có nhiều mẫu dữ liệu nhất nhưng dự đoán lại chưa thực sự tốt so với Sendo. Điều này có phải do bộ dữ liệu Sendo khái quát hơn so với WatchStore ? Vấn đề này nhóm sẽ tìm hiểu và giải quyết trong tương lai.

7. KẾT LUẬN

Bài báo cáo này nhóm đã thực hiện các công việc sau:

1. Thu thập dữ liệu: đã thu thập từ 3 website: sendo, thegioididong và watchstore
2. Tiền xử lý dữ liệu: đã định dạng lại dữ liệu, xử lý khuyết, tổng cộng đã xử lý được 3642 giá trị bị khuyết, xử lí lặp,...
3. Phân tích, thăm dò dữ liệu: nhóm đã chọn ra được những thuộc tính có tương quan ổn định với giá đồng hồ ‘price’
4. Huấn luyện và đánh giá mô hình: đánh giá trên 2 độ đo, MSE và R-Squared

8. HƯỚNG PHÁT TRIỂN

Tuy hiện tại mô hình đã tìm ra được siêu tham số tối ưu, song kết quả dự đoán vẫn sẽ được cải thiện nếu khắc phục được một số hạn chế nhất định:

- Bộ dữ liệu bị bias: cần phải cân bằng dữ liệu, để nó đều hơn bằng cách bổ sung dữ liệu cho các website crawl bị ít data về Đồng hồ Casio, hiện tại dữ liệu từ trang watchstore là khá nhiều (chiếm khoảng hơn 75% trên bộ dữ liệu).
- Bộ dữ liệu chưa được đa dạng: hiện tại nhóm chỉ crawl trên ba website (sendo, thegioididong, watch store).
- Số lượng thông tin về đồng hồ còn thiếu nhiều, chẳng hạn như: Màu sắc, Chức năng, Chế độ bảo hành, Tính chất (đồng hồ thời trang, đồng hồ thể thao,...), Hàng tặng kèm,...
- Dữ liệu được thu thập từ các trang web chưa được tối ưu hóa vì bỏ qua các mô tả sản phẩm bằng đoạn văn, dẫn đến việc bỏ qua một số thông tin quan trọng. Trong tương lai có thể đề xuất ra một phương pháp hiệu quả hơn để thu thập thông tin từ các miền dữ liệu này.

TÀI LIỆU THAM KHẢO

[1] Hoerl, Arthur E., and Robert W. Kennard. “Ridge Regression: Biased Estimation for Nonorthogonal Problems.”

[2] Lederer, J. (2022). Linear Regression. In: Fundamentals of High-Dimensional Statistics.

[3] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Nguyễn Đạt Tuấn	Crawl data watchstore, tiền xử lý dữ liệu
2	Phan Lê Chí Trung	Crawl data sendo, phân tích thăm dò
3	Trần Nguyên Mẫn	Crawl data thegioididong, cài đặt & cải thiện mô hình
Làm báo cáo được thảo luận bởi các thành viên trong nhóm		