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

DỰ ĐOÁN VÀ PHÂN TÍCH CÁC YẾU TỐ ẢNH HƯỞNG TỚI GIÁ LAPTOP TRONG THỊ TRƯỜNG ĐỒ ĐIỆN TỬ 2023

Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Nguyễn Văn Quân	21521333
2	Nguyễn Kiếm Bảo Thắng	21521432
3	Nguyễn Viết Tiến	20520805

TP. HỒ CHÍ MINH – 12/2023

1. GIỚI THIỆU

Đồ án này tập trung vào việc dự đoán và phân tích các yếu tố ảnh hưởng đến giá của laptop trong thị trường điện tử tại Việt Nam năm 2023. Trong một thời đại mà công nghệ ngày càng tiến bộ và thị trường laptop trở nên cạnh tranh hơn, việc hiểu rõ những yếu tố nào có thể ảnh hưởng đến giá cả là quan trọng để các doanh nghiệp, nhà nghiên cứu, và người tiêu dùng có thể đưa ra các quyết định mua sắm và kinh doanh thông minh.

Để thực hiện đồ án, chúng tôi sử dụng selenium để thu thập dữ liệu laptop từ trang thương mại điện tử của tập đoàn FPT [1]. Đồ án chúng tôi sử dụng phương pháp dự đoán và phân tích thống kê để hiểu rõ hơn về xu hướng và biến động trong giá cả laptop. Áp dụng mô hình dụng Linear Regression đa biến, chúng tôi đạt được dự đoán tương đối chính xác với mean R2 cao nhất là 0.8005.

Bộ dữ liệu và đề tài là do nhóm tự thu thập và phân tích thiết kế, chỉ phục vụ cho môn học này, không dựa trên đề tài nào khác.

2. MÔ TẢ BỘ DỮ LỆU

Giới thiệu bộ dữ liệu: Bộ dữ liệu này bao gồm các thông tin về laptop. Thông tin tóm gọn thành 3 loại. Loại 1 là thông tin về thông số kỹ thuật của laptop như là cấu hình RAM, CPU, bộ nhớ, hãng sản xuất… Loại 2 là thông tin về chính sách bán hàng như là ưu đãi khi mua (cột discount), số tiền được trả góp (cột installment), thời gian bảo hành (cột warranty)… Loại 3 là thông tin về đánh giá sản phẩm như số sao của sản phẩm, số lượng phản hồi… Bộ dữ liệu do nhóm tự thu thập, không tham khảo bất kỳ nguồn nào.

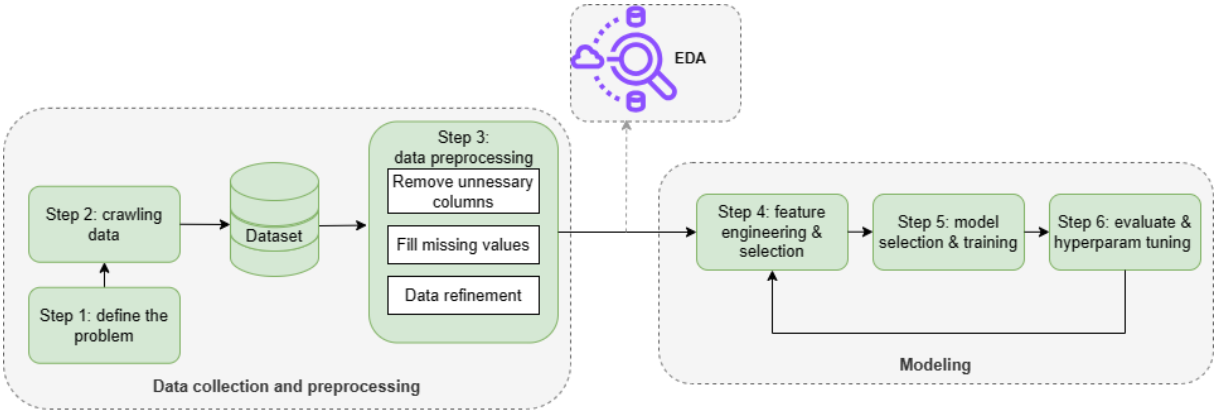
- Bộ dữ liệu tự thu thập tại website fptshop [1]
- Công cụ thu thập dữ liệu: Selenium [2]
- Kích thước: 257x39, bao gồm 23 non-numeric features và 16 numeric features.
- Mô tả các biến (features):

STT	Tên thuộc tính	Kiểu dữ liệu	Miền giá trị	Số giá trị khuyết	Mô tả
1	brand	object	{asus, msi,…}	0	Hãng máy tính
2	screen	float	[1.4, 17.3]	0	Kích thước màn hình (inch)
3	CPU	object	{Core i5, Ryzen 5,…}	0	Công nghệ chip
4	RAM	int	{4, 48}	0	Dung lượng RAM (GB)

5	memory	int	{1, 512}	0	Dung lượng bộ nhớ (GB)
6	graphic card	object	{RTX 2050 4GB,…}	29	Hãng card đồ họa
7	weight	float	[0.879, 3.82]	0	Khối lượng (kg)
8	discount	int	[0, 30000000]	0	Số tiền được giảm (đồng)
9	warranty	int	{1, 36}	0	Thời gian bảo hành (tháng)
10	fgold	float	[0.0, 21.235]	0	Điểm tặng khi mua.
11	installment	int	[0, 975000000]	0	Số tiền trả góp (đồng/tháng)
12	evaluation	int	[0, 1259]	0	Số lượng khách hàng đã đánh giá sản phẩm
13	answer	int	[0, 2271]	0	Số lượng phản hồi của khách về sản phẩm
14	star	int	{0, 5}	0	Số sao của sản phẩm
15	size	object	{359 x 256 x 22.8,…}	1	Chiều dài, rộng, cao (mm)
16	color	object	{Đen, Xám,…}	0	Màu sắc
17	cpu_brand	object	{Intel, AMD,...}	0	Hãng chip
18	cpu_type	object	{11400H, 12450H,…}	3	Loại cpu
19	cpu_speed	float	[0.9, 33.0]	19	Tốc độ cpu (ghz)
20	cpu_core	int	[2.0, 16.0]	72	Số nhân
21	cpu_thread	int	[4.0, 24.0]	81	Số luồng

22	ram_type	object	{DDR4, DDR5,..}	21	Loại RAM
23	ram_speed	object	{3200MHz,…}	64	Tốc độ RAM (mhz)
24	ram_track	object	{2, 1,…}	0	Số khe cắm RAM
25	ram_onboard	object	{0, 1,...}	1	Số RAM onboard
26	screen_tech	object	{Anti-Glare,…}	1	Công nghệ màn hình
27	screen_resolution	object	{1920x1080 Pixels,...}	0	Độ phân giải
28	screen_plate	object	{VA, TN,…}	3	Tấm nền
29	screen_scan	object	{144, 120,…}	10	Tần số quét của màn hình
30	screen_coverage	object	{63% sRGB,…}	134	Độ phủ màu
31	screen_type	object	{OLED, …}	120	Loại màn hình
32	speaker	int	[1.0, 6.0]	147	Số lượng loa
33	keyboard_type	object	{Bàn phím cứng,…}	0	Loại bàn phím
34	keyboard_number	object	{Có, Không,…}	1	Bàn phím số
35	keyboard_lamp	object	{LED, Không,…}	1	Bàn phím có đèn không
36	battery_type	object	{Lithium-ion,…}	140	Loại pin
37	battery_supply	object	{48W, 65W,…}	122	Nguồn điện để sạc (W)
38	os	object	{windows,…}	0	Hệ điều hành
39	price	int	[3990000, 102490000]	0	Giá laptop (đồng)

3. THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU



Hình 1. Quy trình PTDL.

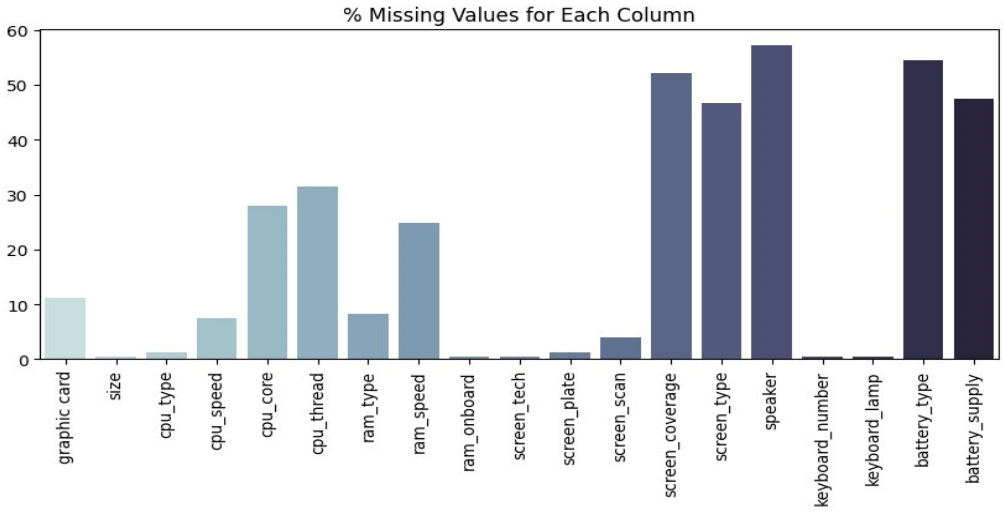
Trong phần này, chúng tôi sẽ giới thiệu về hai giai đoạn quan trọng trong quy trình tổng thể, bao gồm thu thập và tiền xử lý dữ liệu. Đây là hai bước cốt lõi đối với việc xử lý thông tin, và chúng đóng vai trò quan trọng trong việc đảm bảo chất lượng và độ tin cậy của dữ liệu.

3.1. Thu thập dữ liệu

Chúng tôi lựa chọn nguồn thu thập dữ liệu từ trang web https://fptshop.com.vn để đảm bảo tính khách quan và đáng tin cậy của thông tin, vì FPT Shop được biết đến là một trong những nhà cung cấp sản phẩm công nghệ uy tín tại Việt Nam. Chúng tôi cũng thực hiện quá trình thu thập dữ liệu một cách tự động thông qua công cụ Selenium, giúp tối ưu và mặt thời gian và giảm thiểu nguy cơ sai sót.

3.2. Tiền xử lý dữ liệu

3.2.1. Xóa bỏ những cột thuộc tính không cần thiết



Hình 2. Tỉ lệ phần trăm giá trị khuyết các cột.

Việc xóa những cột dữ liệu không cần thiết giúp loại bỏ thông tin không đầy đủ hoặc không quan trọng, từ đó cải thiện chất lượng và tính đồng nhất của tập dữ liệu. Trong tổng số 39 cột giá trị, chúng tôi quyết định loại bỏ 10 cột giá trị vì các lý do sau:

- Xóa do tỉ lệ giá trị khuyết lớn hơn 20%: screen_coverage, battery_type, cpu_core, screen_type, battery_supply, cpu_thread, ram_speed, ram_type.
- Xóa do không chứa thông tin quan trọng: speaker, ram_onboard.

3.2.2. Điền khuyết

Tác dụng của việc điền giá trị bị khuyết trong dữ liệu là giúp hoàn thiện thông tin, tăng độ chính xác của dữ liệu và cải thiện khả năng phân tích. Việc này cũng là bước quan trọng trong tiền xử lý dữ liệu, giúp đảm bảo tính toàn vẹn và hiệu suất của mô hình phân tích sau này. Do đó chúng tôi đã thực hiện điền khuyết trên các cột giá trị như sau:

- Biến numeric: Để xử lý giá trị thiếu trong cột cpu_speed, chúng tôi áp dụng phương pháp KNN (K-Nearest Neighbors) để tìm ra các giá trị thay thế tương đồng. Phương pháp này cho phép chúng tôi ước lượng giá trị còn thiếu bằng cách sử dụng thông tin từ những mẫu dữ liệu có đặc tính tương tự nhất.
- Biến non-numeric: graphic, card size, cpu_type, screen_tech, screen_plate, screen_scan, keyboard_number, keyboard_lamp có rất ít giá trị khuyết nên chúng tôi chọn xóa đi mà không cần quá lo lắng đến việc ảnh hưởng tới bộ dữ liệu.

3.2.3. Tinh chỉnh data

Việc tinh chỉnh dữ liệu giúp chuẩn hóa, loại bỏ nhiễu, và đồng nhất định dạng, tăng tính nhất quán và tiện lợi cho xử lý toán học. Dữ liệu sạch sẽ và thống nhất cũng tạo điều kiện thuận lợi cho việc trực quan hóa và chuẩn bị dữ liệu cho mô hình hóa. Các cột giá trị mà chúng tôi đã tinh chỉnh lại như sau:

- Cột os: nếu giá trị không phải là một trong hai giá trị ‘macos’ hoặc ‘windows’ thì sẽ đổi lại là ‘windows’ vì đa phần các laptop khi bán đều cài một trong hai hệ điều hành này.
- Cột screen_scan: đưa về dữ liệu dạng số như ‘60 Màn hình chính’ thành ‘60’. Nếu có các giá trị không đúng thì sẽ thay bằng ‘60’ vì đa số màn hình laptop hiện giờ đều có tần số từ 60 Hz trở lên.
- Cột keyboard_lamp: nếu giá trị là ‘Đang cập nhật’ thì thay lại bằng ‘Không’.
- Cột ram_track: nếu giá trị là ‘Không’ thì thay lại bằng ‘0’.
- Cột size: đổi thành ba cột mới là width, length, height tương đương với các kích thước của laptop. Nếu đơn vị là cm thì đổi về mm.
- Cột graphic card: tách thêm cột brand_card bằng cách lấy chữ đầu tiên của tên card như ‘NVIDIA GeForce MX330 2GB’ thành ‘NVIDIA’.

- Do tính chất của bộ dữ liệu, các cột như: weight, discount, fgold, installment, answer, height, width, length, cpu_speed, price chúng tôi chuyển chúng về kiểu dữ liệu float, còn lại chuyển thành kiểu dữ liệu category.
- Các cột discount, installment, price: chia cho 1000000, đơn vị là triệu đồng.

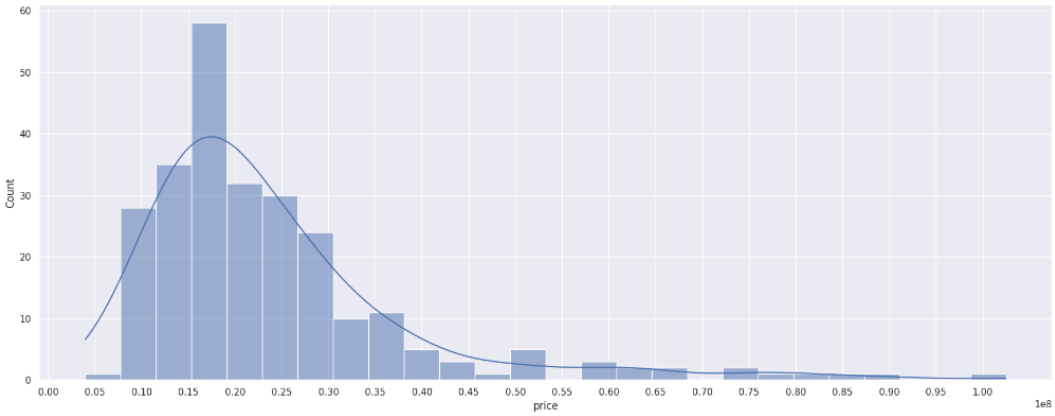
Cuối cùng chúng tôi thu được tidy data có kích thước 253x32, trong đó có 10 cột numeric và 22 cột category.

4. PHÂN TÍCH THĂM DÒ

Khi một người bình thường không có chuyên môn về công nghệ quyết định mua laptop, họ thường tập trung vào những yếu tố cơ bản như giá cả, thương hiệu, kích thước màn hình, loại CPU, dung lượng RAM, dung lượng ổ cứng, loại card đồ họa để có thể lựa chọn phù hợp với nhu cầu cá nhân. Tuy nhiên, nếu là một chuyên gia trong lĩnh vực máy tính hoặc một doanh nghiệp đang lên kế hoạch sản xuất một dòng laptop mới tại thị trường Việt Nam, quá trình đánh giá trên nhiều yếu tố khác lại chi tiết và cẩn thận hơn. Dưới góc nhìn của một người bình thường có nhu cầu mua latop, trong phần này chúng tôi sẽ tập trung phân tích các yếu tố cơ bản ấy.

4.1. Gía cả

Thị trường laptop ở Việt Nam đa dạng về mức giá, tuy nhiên, sự chú ý chủ yếu đổ vào phân khúc từ 10 triệu đến 30 triệu đồng. Trong khi đó, phân khúc cao cấp với mức giá trên 80 triệu có sự hiện diện nhưng chiếm tỷ lệ nhỏ, phản ánh sự tập trung của đa số người tiêu dùng vào phân khúc tầm trung.

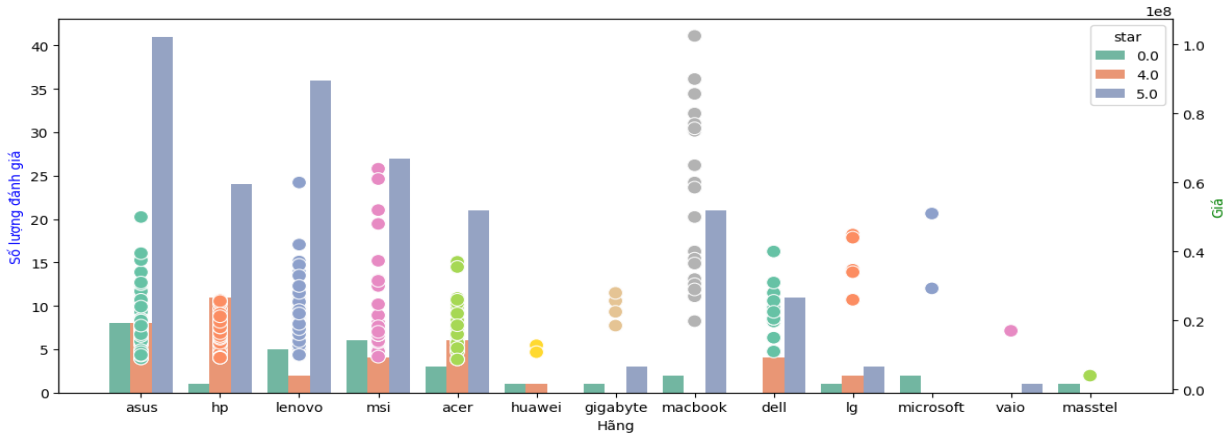


Hình 3. Phân phối về giá laptop được bán trên web.

4.2. Thương hiệu

Những thương hiệu nổi tiếng thường có danh tiếng tích cực, được xây dựng qua thời gian từ chất lượng sản phẩm và dịch vụ. Danh tiếng tốt có thể tạo ra sự tin tưởng từ phía người tiêu dùng, và vì vậy, thương hiệu có thể đặt giá cao hơn mà vẫn có sự chấp nhận từ thị trường điển hình như Apple. Trong *Hình 4*, Apple có mức giá trải dài từ 10 triệu cho đến 100 triệu, mức giá cao nhất cho một chiếc laptop mà FPT shop bán. Ngoài ra nếu quan tâm tới mức giá tầm trung nhưng chất lượng vẫn cực kì tốt thì người tiêu

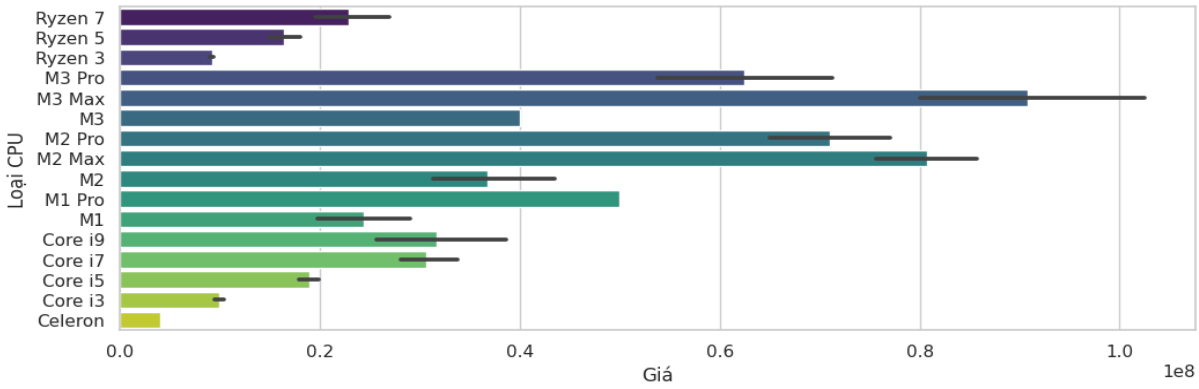
dùng có thể xem qua thương hiệu Asus. Thương hiệu này được đánh giá 5 sao nhiều nhất cũng như mức giá chỉ từ 5 triệu cho đến dưới 25 triệu.



Hình 4. Biểu đồ count plot cùng scatter plot thể hiện giá laptop theo thương hiệu và số lượt đánh giá qua các mức sao.

4.3. Loại CPU

CPU là yếu tố chủ chốt quyết định giá laptop. Thương hiệu nổi tiếng và dòng CPU mới đều ảnh hưởng đến giá bán. Trong ba dòng chính trên thị trường, CPU Ryzen của AMD thường có giá thấp hơn so với các CPU core i của Intel với hiệu suất tương đương. Sự cạnh tranh giá của AMD là do họ đang cố gắng chiếm thị phần bằng các sản phẩm có giá cạnh tranh nhưng với hiệu suất không kém. Còn đối với CPU dòng M của Apple, chúng được đánh giá cao với tốc độ xử lý nhanh và mượt mà, giải thích cho giá cao hơn so với các đối thủ khác.

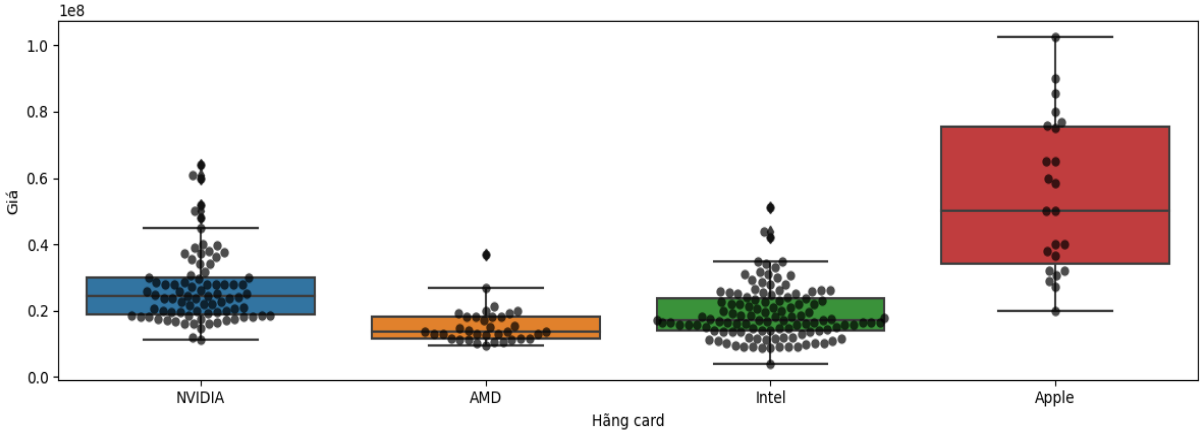


Hình 5. Biểu đồ bar plot thể hiện giá laptop theo loại CPU.

4.4. Card đồ họa

Card đồ họa cũng là một trong những yếu tố chủ chốt quyết định giá của laptop. Máy có card đồ họa mạnh mẽ thường cung cấp hiệu suất đồ họa cao, đặc biệt là khi xử lý các ứng dụng đồ họa và chơi game nặng. Trong phân khúc laptop tầm trung, card đồ họa NVIDIA hoặc AMD với dung lượng VRAM 4GB thường là lựa chọn phổ biến cho một chiếc máy gaming. Nếu là dân văn phòng một chiếc laptop với card đồ họa tích hợp của intel là một quyết định không tồi. Nhìn vào *Hình 9*, giá của những chiếc laptop sử

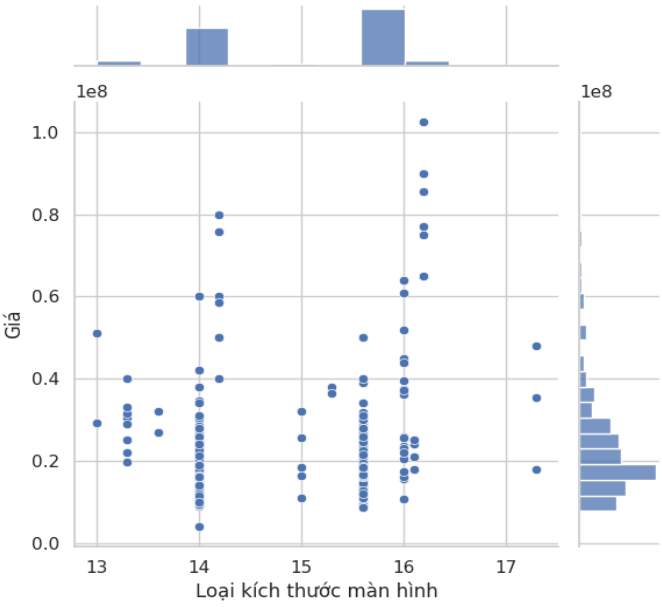
dụng card tích hợp của Apple thường cao hơn so với các hãng khác, thậm chí cao hơn các những chiếc laptop lắp card rời của NVIDIA. *Hình 9* cũng cho thấy FPT chủ yếu bán các dòng laptop gắn card tời của NVIDIA và card tích hợp của Intel. Tương tự CPU, giá của laptop lắp card AMD thường có giá mềm hơn của Intel hay NVIDIA.



Hình 9. Biểu đồ box plot thể hiện giá laptop theo thương hiệu card đồ họa.

4.5. Kích thước màn hình

Kích thước màn hình đóng một vai trò quan trọng trong việc định đoạt giá của laptop, với xu hướng là các máy tính xách tay có màn hình lớn thường có giá cao hơn một chút. Điều này rõ ràng khi nhìn vào sự chênh lệch giá giữa các laptop 16 inch so với các model 13 đến 15 inch, như thấy rõ trong *Hình 6*. FPT chủ yếu tập trung vào việc cung cấp các sản phẩm với kích thước màn hình 14 và 16 inch, vì đây được coi là các kích thước tiêu chuẩn cho một chiếc laptop. Điều này phản ánh sự hiểu biết về nhu cầu thị trường, nơi người tiêu dùng thường ưa chuộng những chiếc laptop vừa vặn và tiện ích cho cả công việc và giải trí.

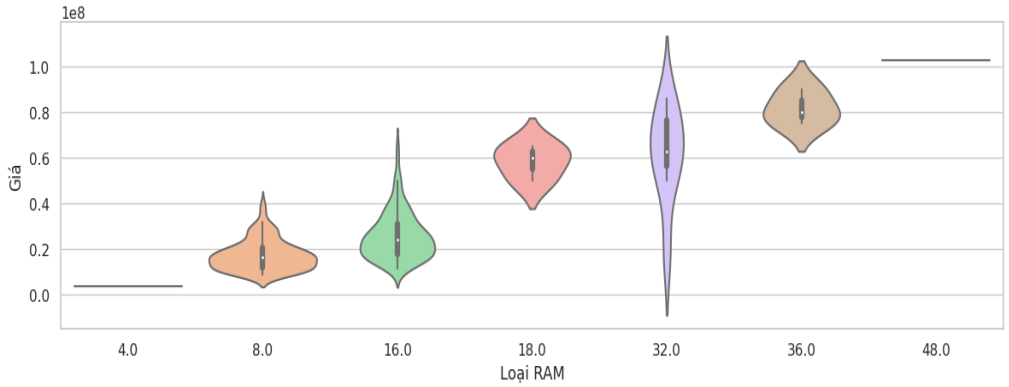


Hình 6. Biểu đồ scatter plot thể hiên giá laptop theo kích thước màn hình

4.6. Dung lượng RAM

RAM là một trong những yếu tố có ảnh hưởng tới giá laptop. Máy có RAM lớn thường mang lại hiệu suất cao, đặc biệt là khi xử lý các ứng dụng nặng. Người mua cần cân nhắc giữa nhu cầu sử dụng và ngân sách để chọn chiếc laptop phù hợp. Đối với laptop tầm trung, RAM 8GB thường là lựa chọn phổ biến. Tuy nhiên, những chiếc laptop

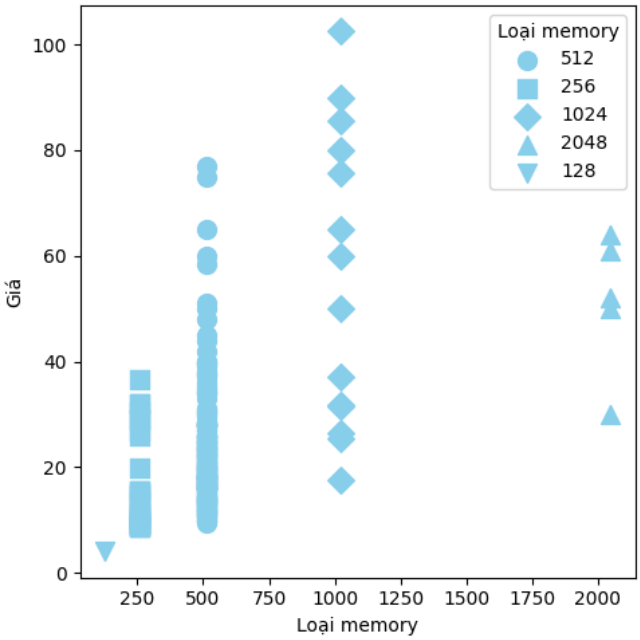
cao cấp thường sử dụng 32GB, thậm chí là 48GB RAM để đảm bảo trải nghiệm mượt mà và hiệu suất ổn định, điều này đồng thời làm tăng giá thành của sản phẩm.



Hình 7. Biểu đồ violin plot thể hiện giá laptop theo loại RAM.

4.7. Dung lượng ổ cứng

Tương tự RAM, dung lượng ổ cứng cũng là một trong những yếu tố quyết định giá của laptop. Máy có ổ cứng lớn thường cung cấp không gian lưu trữ đủ cho dữ liệu và ứng dụng, người mua cần xem xét kỹ lưỡng giữa nhu cầu lưu trữ và ngân sách để chọn chiếc laptop phù hợp. Trong phân khúc laptop tầm trung, ổ cứng 256GB hoặc 512GB thường là lựa chọn phổ biến. Đối với các laptop cao cấp, dung lượng ổ cứng có thể lên đến 1TB hoặc thậm chí 2TB, nhằm đáp ứng nhu cầu lưu trữ lớn và đồng thời làm tăng giá thành của sản phẩm.



Hình 8. Biểu đồ reg scatter plot thể hiện giá laptop theo loại ổ cứng.

5. HUẤN LUYỆN VÀ ĐÁNH GIÁ MÔ HÌNH

5.1. Feature engineering & feature selection

Bộ dữ liệu được chia làm hai nhóm biến chính là biến numeric và category cùng biến mục tiêu (price) là biến liên tục. Chúng tôi chọn phương pháp tương quan Pearson [4] nhằm xác định mức độ tương quan của các biến numeric đối với biến mục tiêu. Các đặc trưng có độ tương quan cao với biến mục tiêu thường được ưu tiên. Với các biến category, chúng tôi sử dụng phương pháp One-Way ANOVA [4] để tìm ra mức độ ảnh

hưởng giữa các biến này và biến mục tiêu. *Hình 10* và *11* là 23 cột thuộc tính mà chúng tôi chọn được với p value < 0.05 bằng hai phương pháp trên. Sau khi chọn được các cột thuộc tính, chúng tôi sử dụng LabelEncoder để chuyển đổi giá trị của các cột category về dạng số. Cuối cùng là áp dụng StandardScaler để chuẩn hóa các giá trị nhằm thuận lợi cho việc huấn luyện mô hình.

	Column	Correlation	p-value
0	fgold	0.499884	1.183923e-17
1	weight	0.246504	6.488505e-05
2	installment	-0.157839	1.127941e-02

Hình 10. Các thuộc tính numeric tương quan với giá laptop.

	Column	F-score	p-value
0	cpu_branch	90.574970	2.026461e-30
1	os	86.728928	1.951787e-29
2	graphic_brand	77.699006	1.178474e-35
3	RAM	71.928081	1.158861e-51
4	keyboard_lamp	54.013806	2.842409e-20
5	CPU	50.158279	2.782947e-65
6	screen	40.340821	7.604032e-49
7	memory	33.768842	1.469429e-22
8	screen_scan	33.192518	2.502333e-29
9	screen_resolution	31.680741	7.271777e-59
10	cpu_type	29.089559	3.763955e-71
11	graphic card	28.124741	4.753457e-58
12	brand	17.812505	2.734885e-27
13	keyboard_type	12.930871	3.587290e-14
14	screen_tech	8.556641	2.017862e-25
15	ram_track	5.946924	6.208116e-04
16	keyboard_number	5.408126	5.010872e-03
17	screen_plate	4.990984	9.459606e-06
18	color	2.646847	6.062084e-03
19	warranty	2.483828	4.424742e-02

Hình 11. Các thuộc tính category ảnh hưởng tới giá laptop.

5.2. Mô hình máy học

Để phù hợp với bài toán dự đoán giá, chúng tôi sử dụng độ đo R-Squared (R2) để đánh giá độ tin cậy của mô hình. Dữ liệu đã được chia thành hai tập Train và Test (8-2). Để dự đoán giá laptop, chúng tôi sử dụng các mô hình máy học sau:

- Linear Regression [5] đơn và đa biến.
- Polynomial Regression [5] đơn và đa biến.

Chúng tôi cũng xây dựng hàm để phù hợp với kích thước đầu vào bất kỳ. Hàm này sẽ tự chọn mô hình tốt nhất cho dữ liệu đầu vào từ 4 mô hình trên. Ở đây, chúng tôi sẽ test trên các đầu vào có số lượng thuộc tính khác nhau: 1, 3, 6 và toàn bộ. Trong đó 6 cột: CPU, brand, screen, RAM, memory, graphic_brand theo góc nhìn của người tiêu dùng và toàn bộ thuộc tính mà feature selection chọn được theo góc nhìn của doanh nghiệp.

5.3. Kết quả mô hình

Số thuộc tính đầu vào	Mô hình tốt nhất	Bậc	R2
1 thuộc tính: CPU	Polynomial Regression đơn biến	4	0.5364

3 thuộc tính: CPU, RAM, memory	Polynomial Regression đa biến	2	0.5934
6 thuộc tính: CPU, brand, screen, RAM, memory, graphic_brand	Polynomial Regression đa biến	2	0.6411
Toàn bộ thuộc tính	Linear Regression đa biến	-	0.8882

Từ các kết quả trên, chúng tôi nhận xét như sau:

- Khi thêm nhiều thuộc tính hơn vào mô hình, hiệu suất của mô hình có tăng lên (R2 tăng từ 0. 5364 đến 0. 8882).
- Mô hình với tất cả các thuộc tính sử dụng Linear Regression đa biến có hiệu suất tốt nhất với R2 = 0. 8882. Điều này cho thấy mô hình dự đoán tốt và đáng tin cậy.

6. KẾT LUẬN

6.1. Kết quả đạt được

- Tự thu thập và xây dựng một bộ dữ liệu phục vụ cho đồ án.
- Thành thạo trong việc tiền xử lý, phân tích và trực quan dữ liệu để đảm bảo sự sạch sẽ và thống nhất của dữ liệu.
- Phân tích các đặc trưng quan trọng mà người mua laptop quan tâm và thành công trong việc xây dựng các mô hình học máy cơ bản để dự đoán giá của laptop dựa trên những đặc trưng này.

6.2. Khó khăn

- Thách thức xuất phát từ khả năng xử lý dữ liệu, đặc biệt là khi hiểu biết về các đặc trưng trong lĩnh vực công nghệ còn hạn chế.
- Số lượng mẫu trong bộ dữ liệu tương đối ít, làm cho việc phân tích xu hướng và mối quan hệ trở nên khó khăn do nhiễu và che khuất.
- Với số lượng mẫu nhỏ, đánh giá chính xác của các mô hình học máy có thể bị hạn chế. Điều này gây khó khăn trong quyết định dựa trên kết quả phân tích và tăng nguy cơ rủi ro trong quá trình đưa ra quyết định.

7. SAU BÁO CÁO

7.1. Mô hình máy học

Để phù hợp với bài toán dự đoán giá, chúng tôi sử dụng độ đo R-Squared (R2) để đánh giá độ tin cậy của mô hình. Dữ liệu đã được chia thành hai tập Train và Test (8-2). Để dự đoán giá laptop, chúng tôi sử dụng các mô hình máy học sau:

- Linear Regression [5] đơn và đa biến.

- Polynomial Regression [5] đơn và đa biến.

Chúng tôi cũng xây dựng hàm để phù hợp với kích thước đầu vào bất kỳ. Hàm này sẽ tự chọn mô hình tốt nhất cho dữ liệu đầu vào từ 4 mô hình trên và đều sử dụng kiểm chứng chéo (cross validation). Ở đây, chúng tôi sẽ test trên các đầu vào có số lượng thuộc tính khác nhau: 1, 3, 6 và toàn bộ. Trong đó 6 cột: CPU, brand, screen, RAM, memory, graphic_brand theo góc nhìn của người tiêu dùng và toàn bộ thuộc tính mà feature selection chọn được theo góc nhìn của doanh nghiệp.

7.2. Kết quả mô hình

Số thuộc tính đầu vào	Mô hình tốt nhất	Bậc	Mean R2
1 thuộc tính: CPU	Polynomial Regression đơn biến	5	0.4548
3 thuộc tính: CPU, RAM, memory	Polynomial Regression đa biến	2	0.6210
6 thuộc tính: CPU, brand, screen, RAM, memory, graphic_brand	Linear Regression đa biến	-	0.5577
Toàn bộ thuộc tính	Linear Regression đa biến	-	0.8005

Từ các kết quả trên, chúng tôi nhận xét như sau:

- Khi thêm nhiều thuộc tính hơn vào cùng mô hình, hiệu suất của mô hình có xu hướng tăng lên.
- Mô hình với tất cả các thuộc tính sử dụng Linear Regression và đa biến có cùng hiệu suất tốt nhất với mean R2 = 0.8005. Điều này cho thấy mô hình dự đoán tốt và đáng tin cậy.

TÀI LIỆU THAM KHẢO

[1] fptshop : https://fptshop.com.vn/may-tinh-xach-tay

[2] Selenium : https://www.selenium.dev

[3] Seaborn: https://seaborn.pydata.org/

[4] Tutorials: Feature Selection In Machine Learning [2023 Edition] - Simplilearn

[5] Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to linear regression analysis. John Wiley & Sons, 2021.

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Nguyễn Văn Quân	Xây dựng mô hình, viết báo cáo, tổng hợp
2	Nguyễn Kiến Bảo Thắng	Tiền xử lý, EDA, viết báo cáo
3	Nguyễn Viết Tiến	Thu thập dữ liệu, mô tả dữ liệu, làm slide