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

SỬ DỤNG THƯ VIỆN SCIKIT-LEARN PHÂN TÍCH GIÁ LAPTOP

Nhóm 2			
Sinh viên thực hiện:			
STT	Họ tên	MSSV	Ngành
38	Nguyễn Ngọc Thanh Sang	21522544	CNCL
41	Phạm Minh Triết	21522712	CNCL
43	Lê Quang Trường	21522732	CNCL

TP. HỒ CHÍ MINH – 12/2024

1. GIỚI THIỆU

Đề tài này tập trung vào việc phân tích và dự đoán giá bán của laptop dựa trên bộ dữ liệu được thu thập từ Kaggle[1], với mục tiêu chính là xây dựng một mô hình dự đoán giá có độ chính xác cao. Để đạt được mục tiêu, nhóm đã sử dụng các công cụ hỗ trợ như Google Colab và Jupyter Notebook để lập trình, đồng thời lưu trữ mã nguồn và dữ liệu trên Github[8]. Quá trình xử lý và phân tích dữ liệu được thực hiện bằng Python với sự hỗ trợ của thư viện Pandas, giúp dễ dàng thao tác dữ liệu. Ngoài ra, nhóm còn áp dụng thư viện scikit-learn để triển khai các thuật toán học máy phổ biến, đồng thời thực hiện tiền xử lý dữ liệu như loại bỏ các giá trị khuyết, chuẩn hóa dữ liệu và thay thế các giá trị không hợp lệ.

Bộ dữ liệu được nhóm sử dụng là một tập mẫu do tác giả Pradeep Jangir cung cấp trên Kaggle[2]. Việc tiền xử lý được tiến hành kỹ lưỡng nhằm chuẩn bị cho quá trình xây dựng mô hình dự đoán hiệu quả. Nhóm đã lựa chọn mô hình XGBRegressor – một thuật toán mạnh mẽ trong việc dự đoán giá trị liên tục. Mô hình này đã được đánh giá bằng các thước đo như R2 score, Mean Squared Error (MSE), và Mean Absolute Error (MAE). Kết quả cho thấy mô hình đạt R2 score là 0.8179, MAE là 3.702117e+06 và MSE là 4.114120e+13, minh chứng cho khả năng dự đoán khá tốt của mô hình so với dữ liệu thực tế.

Nhóm cam kết thực hiện đề tài một cách minh bạch, chỉ sử dụng dữ liệu tham khảo từ Kaggle và không phụ thuộc vào bất kỳ nguồn nào khác. Điều này đảm bảo tính khách quan và chính xác trong quá trình nghiên cứu và xây dựng mô hình dự đoán giá laptop.

2. MÔ TẢ BỘ DỮ LỆU

Bộ dữ liệu được nhóm tham khảo tại Kaggle[2], được tác giả thu thập từ trang web Smartprix[4], một nền tảng trực tuyến liệt kê và so sánh các sản phẩm điện tử khác nhau, bao gồm cả máy tính xách tay. Dữ liệu được thu thập vào ngày 3/8/2024.

Bộ dữ liệu có tổng cộng 3976 dòng và 18 cột, trong đó bao gồm biến mục tiêu Price. Trong đó gồm 4 biến số, 13 biến phân loại và 1 cột không có nghĩa (Unnamed).

Bên dưới là bảng mô tả các thuộc tính trong bộ dữ liệu.

Tên thuộc tính	Mô tả thuộc tính	Loại biến	Kiểu giá trị
Unnamed: 0	Số chỉ mục (sẽ được loại bỏ).		Int64
Brand	Thương hiệu của laptop. Vd: HP, Lenovo, Dell…	Phân loại	Object
Name	Tên model cụ thể của laptop. Vd: Lenovo Ideapad Gaming 3 15IHU6 (82K101EEIN) Laptop…	Phân loại	Object
Price	Giá bán của laptop (theo đơn vị Indian Rupee).	Số	Int64
Processor_Name	Tên bộ xử lý được sử dụng. Vd: Intel Core i5, AMD Ryzen 5…	Phân loại	Object

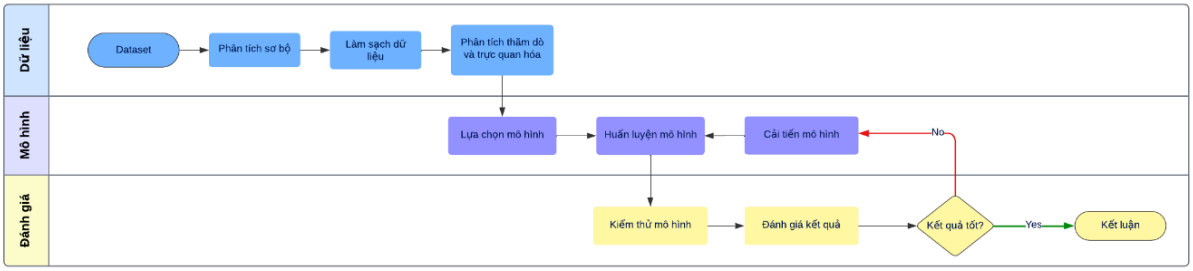
Processor_Brand	Thương hiệu của bộ xử lý. Vd: Intel, AMD, MediaTek…	Phân loại	Object
RAM_Expandable	Cho biết RAM có mở rộng được hay không và ở mức độ nào. Vd: 12 GB Expandable…	Phân loại	Object
RAM	Dung lượng RAM được lắp sẵn. Vd: 8 GB, 16 GB…	Phân loại	Object
RAM_TYPE	Loại RAM được sử dụng. Vd: DDR4, LPDDR4X…	Phân loại	Object
Ghz	Tốc độ xung nhịp của bộ xử lý (theo đơn vị Ghz). Vd: 2.3 Ghz Processor…	Số	Object
Display_type	Loại màn hình được sử dụng. Vd: LCD, LED…	Phân loại	Object
Display	Kích thước màn hình của laptop (theo đơn vị inches).	Phân loại	Object
GPU	Kiểu GPU (card đồ họa). Vd: UHD, Iris Xe, Geforce RTX 3050 GPU 4GB…hoặc Integrated nếu là card tích hợp.	Phân loại	Object
GPU_Brand	Thương hiện của GPU. Vd: NVIDIA, AMD, Intel…	Phân loại	Object
SSD	Dung lượng lưu trữ SSD (Solid State Drive). Vd: 512 GB SSD Storage, 1024 GB SSD Storage…	Phân loại	Object
HDD	Dung lượng lưu trữ HDD (Hard Disk Drive). Vd: No HDD, 1024 GB HDD Storage…	Phân loại	Object
Adapter	Công suất nguồn (theo đơn vị Watts).	Số	Object
Battery_life	Thời lượng pin dự kiến của laptop. Vd: Upto 7 Hrs Battery Life, Upto 10 Hrs Battery Life…	Số	Object

Bảng 1. Bảng mô tả dữ liệu

Trong số các thuộc tính này, có một số thuộc tính có kiểu dữ liệu trong bộ dữ liệu gốc khác với kiểu nó nên có. Ví dụ như biến Battery_life, vì giá trị của nó được ghi dưới dạng chuỗi văn bản như là "Upto 7 Hrs Battery Life" nên nó được nhận diện ban đầu là kiểu dữ liệu object, nhưng thực tế nó nên là biến số và có kiểu dữ liệu là int hoặc float, chẳng hạn như là 7. Nhóm sẽ chuẩn hóa lại trong quá trình tiền xử lý ở sau.

Nhóm đã thống kê số lượng giá trị bị khuyết như sau: Cột Battery_Life 418 giá trị, cột GPU khuyết 8 giá trị và cột GPU_Brand khuyết 4 giá trị.

3. PHƯƠNG PHÁP PHÂN TÍCH



Hình 1. Quy trình phân tích dữ liệu

3.1. Phân tích sơ bộ

Bộ dữ liệu mà chúng tôi sử dụng từ Kaggle[2] bao gồm 3976 dòng và 18 cột. Nhóm đã tiến hành phân tích thống kê nhằm khám phá và hiểu rõ các đặc điểm quan trọng của dữ liệu. Trong quá trình phân tích, chúng tôi đã phát hiện một số vấn đề trong bộ dữ liệu:

- Trước hết, dữ liệu bị khuyết xuất hiện trong một số cột. Điều này làm giảm tính đầy đủ của dữ liệu và yêu cầu thực hiện các biện pháp xử lý giá trị thiếu.Thứ hai, có các cột dữ liệu không có được kiểu dữ liệu và giá trị chính xác của mình. Điều này có thể gây ra sai sót và ảnh hưởng đến kết quả phân tích.
- Tiếp theo, một số cột chưa có kiểu dữ liệu hoặc giá trị được định dạng chính xác, gây ra nguy cơ sai lệch trong quá trình phân tích.
- Cuối cùng, các biến phân loại dạng chuỗi xuất hiện nhiều giá trị khác nhau dù có ý nghĩa tương đồng, điều này có thể ảnh hưởng đến quá trình phân loại và làm sai lệch kết quả dự đoán của mô hình.

Việc xác định và xử lý những vấn đề này là bước thiết yếu để đảm bảo bộ dữ liệu đạt được tính đồng nhất và chất lượng cần thiết cho các bước phân tích tiếp theo.

3.2. Làm sạch dữ liệu

Sau khi tiến hành một quá trình thăm dò sơ bộ và thu được kết quả thống kê chi tiết về quy mô của bộ dữ liệu thô, chúng tôi tiến hành làm sạch và tiền xử lý dữ liệu để đảm bảo chất lượng và tính nhất quán của thông tin thu thập được.

Chúng tôi bằng đầu bằng việc kiểm tra các giá trị bị khuyết của các biến và phát hiện được 3 biến có giá trị khuyết là GPU, GPU_Brand và Battery_Life. Sau khi thực hiện quá trình phân tích, chúng tôi quyết loại bỏ biến Battery_Life vì có nhiều dữ liệu nhiễu và sử dụng các phương pháp điền khuyết cho thuộc tính GPU và GPU_Brand.

Chúng tôi tiếp tục định dạng lại giá trị không hợp lệ và chuẩn hóa lại giá trị trong các trường như “Name”, “Price”, “Processor_Name”, “RAM_Expandable”, ‘’RAM”, ‘’RAM_TYPE”, “Ghz”, “Display”, “GPU”, “GPU_Brand”, “SSD”, “HDD”, “Adapter”. Đồng thời chúng tôi cũng tạo thêm 2 cột mới là “OS” và “VRAM”, có nghĩa là hệ điều hành mà máy tính có đó và dung lượng VRAM mà máy hỗ trợ.

3.3. Phân tích thăm dò và trực quan hóa dữ liệu

Sau giai đoạn làm sạch dữ liệu, chúng tôi chuyển sang giai đoạn phân tích thăm dò và trực quan hóa chúng. Điều này nhằm mục đích tìm ra các cấu trúc và xu hướng quan trọng trong dữ liệu, đồng thời xác định độ tương quan giữa các biến dạng số và giá tiền (target variable). Bằng cách này, chúng tôi có thể chọn ra những thuộc tính quan trọng và có ảnh hưởng đáng kể đối với quá trình huấn luyện mô hình dự đoán giá máy tính cũ.

Quá trình thăm dò dữ liệu bao gồm việc tạo các biểu đồ để trực quan hóa sự phân phối của các biến và hiểu rõ hơn về mối quan hệ giữa chúng. Đồng thời, chúng tôi tập trung vào việc tìm hiểu về mối quan hệ giữa các biến phân loại và biến mục tiêu để xác định những thuộc tính quan trọng có thể đóng góp lớn cho việc dự đoán giá máy tính.

Những kết quả thu được từ quá trình này sẽ là cơ sở để chọn lọc các đặc trưng quan trọng để có thể xây dựng một mô hình dự đoán hiệu quả.

3.4. Huấn luyện mô hình

Trong quy trình của chúng tôi, bước lựa chọn mô hình đóng vai trò quan trọng trong việc xác định độ chính xác và khả năng dự đoán của hệ thống. Ở đây, nhóm đã cân nhắc và lựa chọn các mô hình hồi quy, bao gồm XGBRegressor, DecisionTreeRegressor, RandomForestRegressor, LinearRegression, Ridge Regression và KneighborsRegressor. Việc tham khảo và lựa chọn mô hình phù hợp giúp cho việc huấn luyện mô hình cho ra kết quả tốt hơn.

Ở bước này, bộ dữ liệu được chia theo tỉ lệ 8:2 thể phục vụ cho việc huấn luyện và kiểm thử. Trong quá trình thử nghiệm trên các mô hình khác nhau, chúng tôi đã giữ cài đặt mặc định của các thuật toán từ thư viện, không thực hiện bất kỳ tinh chỉnh nào. Điều này nhằm mục đích đánh giá hiệu suất ban đầu của các mô hình mà không có sự ảnh hưởng đến từ việc điều chỉnh tham số. Đối với mỗi mô hình, nhóm đã đánh giá hiệu suất dựa trên ba thang đo là R2 Score, Mean Absolute Error và Mean Squared Error.

4. PHÂN TÍCH SƠ BỘ

Đầu tiên, chúng tôi load dữ liệu tìm được bằng công cụ pandas và thực hiện tìm hiểu thông tin sơ bộ của bộ dữ liệu thông qua hàm info().

Kết quả trả về cho thấy hầu hết các cột đều có 3976 giá trị không null, riêng chỉ có duy nhất 3 cột là có số lượng giá trị nhỏ hơn 3976, lần lượt là GPU (3968 giá trị không null), GPU_Brand (3972 giá trị không null) và Battery_Life (3558 giá trị không null).

Tiếp đến, chúng tôi in ra 5 dòng đầu của bộ dữ liệu để kiểm tra giá trị của các cột có hợp lý hay không:

	Unnamed: 0	Brand	Name	Price	Processor_Name	Processor_Brand	RAM_Expandable	RAM	RAM_TYPE	Ghz	Display_type	Display	GPU	GPU_Brand	SSD	HDD	Adapter	Battery_Life
0	0	HP	HP Chromebook 11A-NA0002MU (2E4N0PA) Laptop (1...	22990	MediaTek Octa-core	MediaTek	Not Expandable	4 GB	DDR4 RAM	2.0 Ghz Processor	LED	11.6	Integrated Graphics	MediaTek	64 GB SSD Storage	No HDD	45	Upto 12 Hrs Battery Life
1	1	Lenovo	Lenovo Ideapad Slim 3 (82KU017KIN) Laptop (15....	36289	AMD Hexa-Core Ryzen 5	AMD	12 GB Expandable	8 GB	DDR4 RAM	4.0 Ghz Processor	LCD	15.6	Radeon	AMD	512 GB SSD Storage	No HDD	65	Upto 11 Hrs Battery Life
2	3	Dell	Dell G15-5520 (D560822WIN9B) Laptop (15.6 Inch...	78500	Intel Core i5 (12th Gen)	Intel	32 GB Expandable	16 GB	DDR5 RAM	3.3 Ghz Processor	LCD	15.6	GeForce RTX 3050 GPU, 4 GB	NVIDIA	512 GB SSD Storage	No HDD	56	Upto 10 Hrs Battery Life
3	4	HP	HP 15s-fy5007TU (91R03PA) Laptop (15.6 Inch | ...	55490	Intel Core i5 (12th Gen)	Intel	8 GB Expandable	8 GB	DDR4 RAM	4.2 Ghz Processor	LCD	15.6	Iris Xe	Intel	512 GB SSD Storage	No HDD	no	Upto 7.30 Hrs Battery Life
4	6	Infinix	Infinix Inbook Y2 Plus XL29 Laptop (15.6 Inch ...	21990	Intel Core i3 (11th Gen)	Intel	Not Expandable	8 GB LP	LPDDR4X RAM	1.7 Ghz Processor	LCD	15.6	UHD	Intel	512 GB SSD Storage	No HDD	45	Upto 8 Hrs Battery Life

Hình 2. Tổng quan về bộ dữ liệu

Qua bảng trên, chúng tôi nhận định một số biến như RAM_Expandable, RAM, GHz, SSD, HDD và Battery_Life sẽ hợp lý hơn nếu được chuyển đổi thành giá trị số. Ngoài ra, có thể thấy rằng biến Display và Adapter đều chứa các giá trị số, nhưng hàm info() lại trả về kiểu dữ liệu cho hai biến này là object, báo hiệu rằng có sự xuất hiện của dữ liệu nhiễu trong hai biến này.

Tiếp theo, chúng tôi thực hiện tìm giá trị unique trên từng biến và nhận thấy rằng có một số giá trị phân loại giống nhau nhưng khác cách viết , ví dụ như “NVIDIA” và “Nvidia”. Điều này sẽ gây ảnh hưởng đến phân tích về sau.

5. LÀM SẠCH DỮ LIỆU

Sau khi thực hiện thăm dò sơ bộ và có được góc nhìn tổng quát về bộ dữ liệu thô, nhóm chúng tôi tiến hành giai đoạn làm sạch và tiền xử lý dữ liệu để đảm bảo chất lượng và tính nhất quán của thông tin thu thập được. Sau đây là các bước chúng tôi thực hiện để làm sạch bộ dữ liệu này:

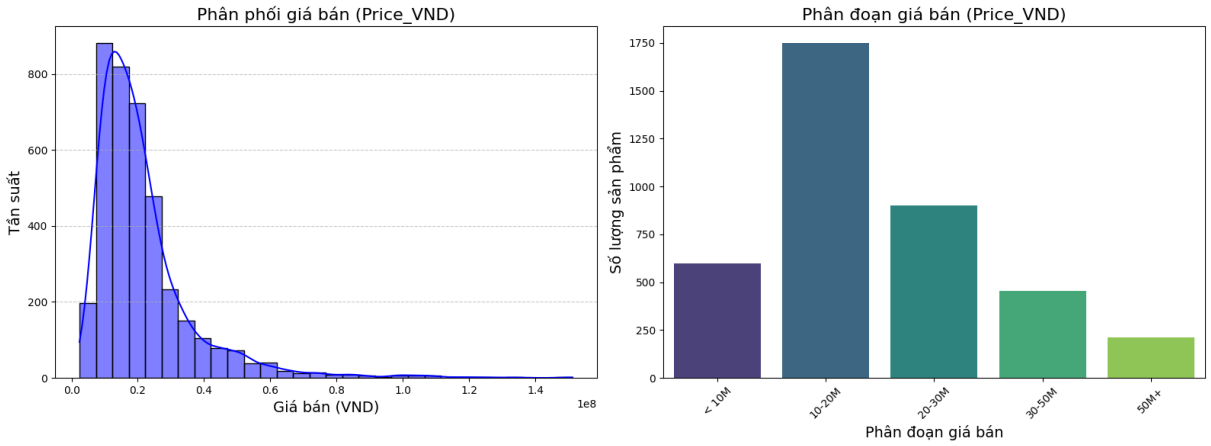
- Chúng tôi bắt đầu với việc xử lý các đặc trưng bị khuyết:
  + Thuộc tính Battery_Life đặc biệt hơn cả với khoảng 10% giá trị bị khuyết. Ngoài ra giá trị của cột này cũng bị mâu thuẫn với nhau vì tồn tại cùng lúc 2 loại giá trị chẳng hạn như "65W Adapter" và "Upto 6 Hrs Battery Life". Nhận thấy rằng giá trị về thời gian đúng với ý nghĩa thực tế của biến này hơn nên chúng tôi gán các giá trị chứa thông tin về công suất của Adapter tương đương với NaN. Sau quá trình xử lý, số lượng giá trị khuyết lên đến khoảng 62% nên chúng tôi đã quyết định không sử dụng cột này vào phân tích.
  + Chúng tôi nhận thấy có những giá trị của cột GPU bị khuyết và GPU_Brand tương ứng là Intel nên chúng tôi quyết định xem các biến này tương ứng với card tích hợp và điền khuyết với giá trị "Intergrated Graphics" cho chúng.
  + Về đặc trưng GPU_Brand, theo thông tin tham khảo từ trang pc-builds[5] và thegioididong[6], chúng tôi nhận ra 2 loại GPU "R5" và "Pro 555X" đều thuộc về hãng "AMD" nên quyết định điền khuyết cột GPU_Brand tương ứng của chúng với giá trị này.
- Tiếp theo, chúng tôi xử lý định dạng của các đặc trưng quan trọng:
  + Với biến mục tiêu Price, chúng tôi sẽ cột này bằng Price_VND vì đơn vị hiện tại của biến này là rupee - đơn vị tiền tệ của Ấn Độ.
  + Đối với biến RAM_Expandable đang có kiểu Object với giá trị như "8 GB Expandable", chúng tôi mong muốn biến này ở dạng số và giá trị "Not expandable" sẽ được gán là 0.
  + Cột Ghz – một trong những cột quan trọng đang có dữ liệu dạng "1.4 Ghz Processor", chúng tôi cũng sẽ trích xuất giá trị số của cột này. Tuy nhiên biến này có 1 số giá trị bằng 0, vì vậy chúng tôi sẽ xem chúng như bị khuyết và điền khuyết chúng bằng trung vị.

+ Ở cột DISPLAY, tất cả giá trị đều ở dạng số tuy nhiên xuất hiện 1 giá trị bất thường là "OLED Display With Touchscreen". Theo trang hpworldkbm[7] thì màn hình loại này có kích thước là 15.6 inch nên đã được điền khuyết bằng giá trị này.
+ Ở cột GPU, chúng tôi đã trích xuất thêm 1 biến quan trọng mới đó là VRAM. Tuy nhiên cột GPU chứa nhiều giá trị sai định dạng nên chúng tôi đã chỉnh sửa thủ công chúng.

Thông qua quá trình làm sạch dữ liệu, nhóm đã thu được bộ dữ liệu mới với các cột như sau:

- Các cột được xem như là biến số: RAM_Expandable, RAM, Ghz, Display, SSD, HDD, Adapter, VRAM.
- Các cột biến phân loại: Brand, Name, Processor_Name, Processor_Brand, RAM_TYPE, Display_type, GPU, GPU_Brand, OS.
- Cột biến mục tiêu: Price_VND.

6. PHÂN TÍCH THĂM DÒ VÀ TRỰC QUAN HÓA DỮ LIỆU

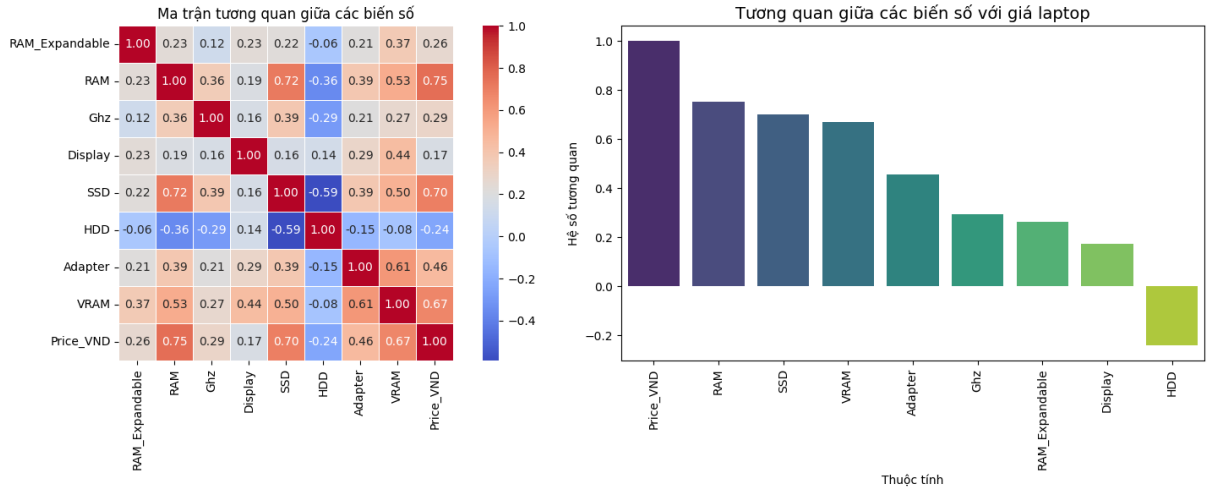


Hình 3. Phân bố của biến mục tiêu (Price)

Khi quan sát các biểu độ trực quan cho phân phối giá bán, nhóm đã rút ra một số nhận xét chung như sau:

- Phân khúc thị trường:
  + Giá bán trải dài từ phân khúc giá rẻ (~2.4 triệu) đến cao cấp (>150 triệu).
  + Khoảng giá từ Q1 đến Q3 (11.9 - 25.2 triệu) chiếm phần lớn, là phân khúc phổ biến.
- Mức giá trung bình:
  + Trung bình giá (21.5 triệu) cao hơn trung vị (17.6 triệu), cho thấy dữ liệu bị lệch phải do có những sản phẩm giá cao.

- Xu hướng:
  + Dữ liệu phản ánh sự đa dạng trong phân khúc sản phẩm, từ laptop bình dân đến cao cấp, phù hợp với nhiều nhu cầu người dùng.



Hình 4. Độ tương quan giữa biến mục tiêu và các biến

Tiếp đến, chúng tôi tiến hành đánh giá mức độ ảnh hưởng của các biến số tới biến mục tiêu là giá bán bằng các phương pháp như xem phân phối giá trị, mức độ tương quan đơn biến, mức độ tương quan khi kết hợp đa biến.

Qua ma trận tương quan và biểu đồ, có thể thấy biến Price_VND có độ tương quan cao với biến RAM, SSD và VRAM. Điều đó cho thấy các biến này có ảnh hưởng lớn đến giá laptop. Theo thứ tự, RAM sẽ có độ tương quan cao nhất (0.75), tiếp đến là SSD (0.70) và VRAM có độ tương quan thấp nhất trong ba biến (0.67)

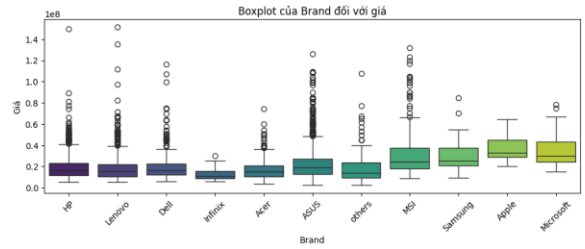
Phân tích dữ liệu biến số cho thấy RAM, SSD và VRAM là những yếu tố quan trọng nhất ảnh hưởng trực tiếp đến giá bán, với mối tương quan mạnh và mối quan hệ tuyến tính rõ ràng. RAM trung bình của các mẫu laptop là 11.13 GB, chủ yếu thuộc phân khúc tầm trung, phù hợp với nhu cầu sử dụng phổ thông lẫn hiệu năng cơ bản. Dung lượng SSD trung bình đạt 472.65 GB, cho thấy xu hướng phổ biến hóa ổ cứng SSD, mang lại tốc độ xử lý cao hơn so với HDD truyền thống. VRAM tập trung chủ yếu vào các dòng laptop phổ thông với GPU tích hợp, nhưng ở phân khúc cao cấp, VRAM lớn hơn lại là yếu tố quan trọng trong các mẫu laptop gaming hoặc workstation. Về kích thước màn hình, 15.6 inch là lựa chọn phổ biến nhất nhờ sự cân bằng giữa trải nghiệm thị giác và tính di động. Tần số xử lý CPU (Ghz) trung bình đạt 2.75, phản ánh hiệu suất phổ thông phù hợp với các nhu cầu cơ bản. Adapter lớn hơn thường xuất hiện trên các mẫu laptop hiệu năng cao, trong khi HDD dần mất vai trò do sự thay thế của SSD hiện đại. Giá bán trung bình của các mẫu laptop là 21.52 triệu VNĐ, với sự phân bổ mạnh ở phân khúc tầm trung, phù hợp với nhu cầu của đa số người tiêu dùng. Nhìn chung, RAM, SSD và VRAM đóng vai trò quyết định trong việc định giá sản phẩm, trong khi các yếu tố như HDD, Display và Adapter chỉ có tác động thấp hơn.

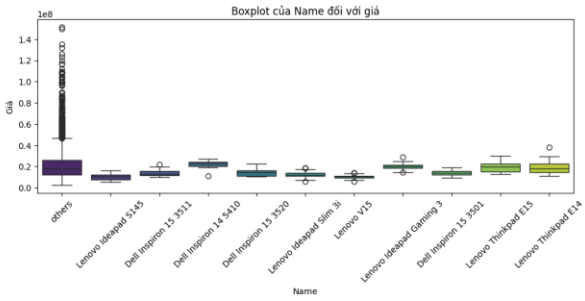
Đối với các biến phân loại, chúng tôi đã sử dụng kiểm định F-test để đánh giá khả năng ảnh hưởng của các biến này đến giá laptop.

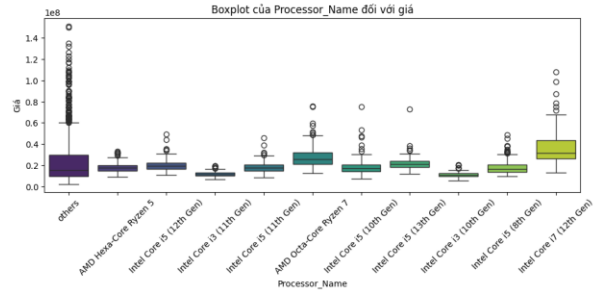
STT	Tên biến	F-statistic	p-value	Có ý nghĩa?
0	GPU_Brand	118.3844	6.703077e-178	True
1	RAM_TYPE	92.37151	1.975029e-257	True
2	Display_type	88.54996	8.209361e-21	True
3	Processor_Name	70.27055	0.000000e+00	True
4	GPU	54.74181	0.000000e+00	True
5	OS	27.90911	1.275768e-61	True
6	Brand	15.01939	6.449073e-72	True
7	Name	11.47409	7.300314e-320	True
8	Processor_Brand	5.949172	1.732426e-05	True

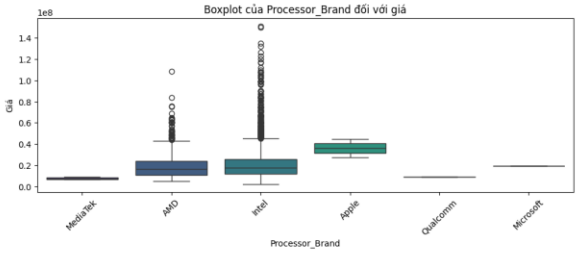
Bảng 2. Bảng thể hiện độ tương quan của các biến với biến mục tiêu

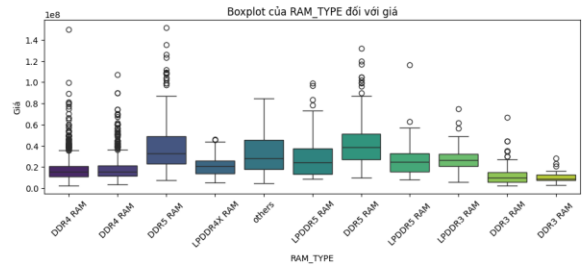
Từ bảng có thể thấy GPU_Brand có F-statistic cao nhất và p-value nhỏ (6.703077e-178). Điều đó thể hiện GPU_Brand có ảnh hưởng mạnh nhất đến giá laptop và có ý nghĩa thống kê. Các biến RAM_TYPE, Display_type và Processor_Name cũng có tầm quan trọng đáng kể. Biến Processor_Brand có ảnh hưởng yếu nhất.

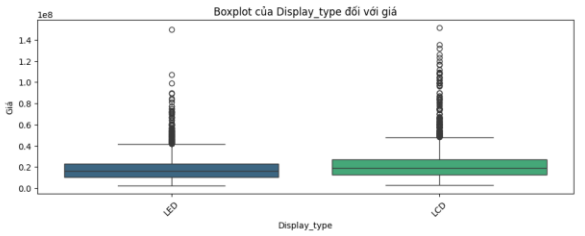


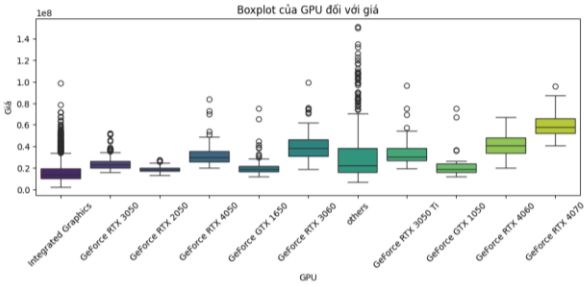


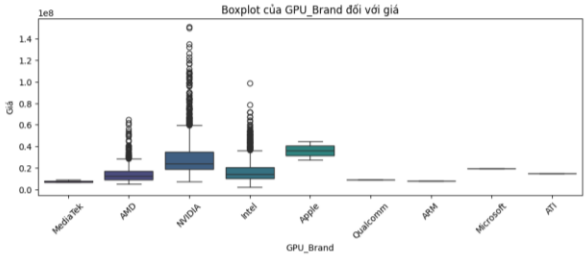


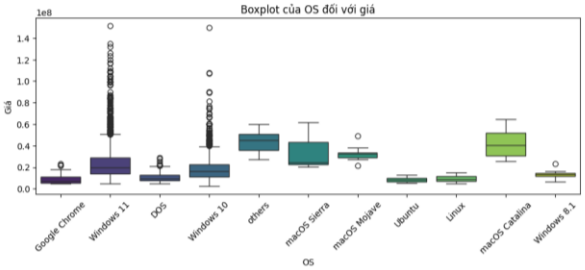












Tuy nhiên, khi nhìn vào phân phối giá tiền theo các biến Processor_Name, hoặc Brand, nhóm nhận thấy rằng có sự khác biệt về phân phối giá giữa các nhóm giá trị với nhau, nghĩa là các biến này có thể có ý nghĩa về mặt phân loại phân khúc giá trên thị trường, tức là cũng có thể có mức độ ảnh hưởng tới giá bán. Vì vậy nhóm đã tiến hành đánh giá thêm về mức độ quan trọng của các biến này thông qua trực quan giá trung

bình của laptop theo nhóm. Kết quả thu được cho thấy giá bán laptop cũng có sự phân hóa rõ rệt giữa các thương hiệu với nhau.

Nhìn chung, các biến liên quan đến cấu hình phần cứng sẽ có ảnh hưởng lớn hơn các biến liên quan đến thương hiệu hoặc tên sản phẩm. Điều này là hợp lý vì giá laptop hoặc máy tính thường bị ảnh hưởng nhiều bởi cấu hình phần cứng.

7. KẾT QUẢ PHÂN TÍCH

Sau quá trình phân tích, chúng tôi đã rút ra được một số biến có ảnh hưởng tới giá của laptop, có thể áp dụng vào mô hình dự đoán như sau:

- Biến số: RAM, SSD, VRAM
- Biến phân loại: Brand, Processor_Name, Processor_Brand, RAM_TYPE, Display_type, GPU, GPU_Brand, OS

Chúng tôi chọn các mô hình Regressor do các biến đầu vào có tuyến tính với giá cả, kết quả có được như sau:

	MAE	MSE	R2
XGBRegressor	2.207033e+06	9.270887e+12	0.770631
DecisionTreeRegressor	2.673671e+06	1.465028e+13	0.637541
RandomForestRegressor	2.377512e+06	1.102642e+13	0.727198
LinearRegression	2.317085e+06	9.790896e+12	0.757766
Ridge	2.308289e+06	9.788094e+12	0.757835
KNeighborsRegressor	2.674143e+06	1.436497e+13	0.644600

Có thể thấy XGBRegressor là mô hình tốt nhất trong bộ so sánh, với mực chênh lệch trung bình là 2,2 triệu đồng và KneighborsRegressor là mô hình kém nhất, cho thấy nó không phù hợp với dữ liệu.

Nhóm đã thực hiện dự đoán thử với một vài sample máy tính thực tế như là Lenovo LOQ 2024 15ARP9 2024[9], giá dự đoán gần đúng với thực tế, chênh lệch từ 600.000 đến 2.500.000 đồng.

8. KẾT LUẬN

Sử dụng bộ dữ liệu về giá laptop có được từ Kaggle, nhóm đã thực hiện điền những giá trị null một cách hợp lý, xác định những giá trị nhiễu trong bộ dữ liệu và loại bỏ chúng. Trong quá trình làm sạch dữ liệu, nhóm đã loại bỏ các cột không có giá trị sử dụng trong dự đoán. Nhóm cũng đã tách ra được dữ liệu mới có ảnh hưởng đến giá cả từ các cột khác của bộ dữ liệu. Sau khi hoàn tất quá trình làm sạch, nhóm thu được bộ dữ liệu không null và các cột ở đúng định dạng nên có.

Nhóm đã thực hiện phân tích thăm dò và đã tìm ra được các biến có ảnh hưởng lớn đến giá cả: RAM, SSD, VRAM, Brand, Processor_Name, Processor_Brand, RAM_TYPE, Display_type, GPU, GPU_Brand, OS.

Tiến hành huấn luyện mô hình, nhóm đã tìm ra một mô hình có kết quả khả quan, với R2 là 0.77, MSE là 9270887279339.035 và MAE là 2207033.0366556835, tức là độ sai lệch của mô hình khoảng 2,2 triệu đồng.

TÀI LIỆU THAM KHẢO

[1] Kaggle. Link: Kaggle: Your Home for Data Science (Truy cập 20/10/2024).

[2] Kaggle. Link: Laptop Dataset (Truy cập 20/10/2024).

[3] Kaggle. Link: laptop_price_prediction.ipynb (Truy cập 20/10/2024)

[4] Smartprix. Link: Smartprix - Best Online Comparison Shopping (Truy cập 25/10/2024).

[5] PC Builds. Link: pc-builds (Truy cập 10/12/2024).

[6] Thế giới di động. Link: thegioididong (Truy cập: 10/12/2024).

[7] HP World KPM. Link: hpworldkbm (Truy cập 10/12/2024).

[8] Github. Link: Github (Truy cập 14/12/2024).

[9] LaptopAZ. Link: Lenovo LOQ giá tốt nhất thị trường - LaptopAZ.vn (Truy cập 13/12/2024).

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
38	Nguyễn Ngọc Thanh Sang	- Viết báo cáo phần 1, 2, - Làm sạch dữ liệu - Xây dựng mô hình dự đoán
41	Phạm Minh Triết	- Viết báo cáo phần 5, 6, 7, 8. - Phân tích thăm dò biến số. - Làm slide báo cáo. - Hỗ trợ thiết kế dashboard.
43	Lê Quang Trường	- Viết báo cáo phần 3 và 4. - Phân tích thăm dò biến phân loại. - Thiết kế dashboard