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

XÂY DỰNG MÔ HÌNH VÀ PHÂN TÍCH CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN GIÁ LAPTOP

Nhóm 4			
Sinh viên thực hiện:			
STT	Họ tên	MSSV	Ngành
1	Mã Kim Phát	22521071	KHDL
2	Đặng Chí Nguyên	22520963	KHDL
3	Võ Tấn Trung	22521573	KHDL
4	Nguyễn Tiến Nam	22520920	KHDL

TP. HỒ CHÍ MINH – 12/2024

1. GIỚI THIỆU

Đề tài này nhằm phân tích các yếu tố ảnh hưởng đến giá laptop trên thị trường và từ đó xây dựng mô hình dự đoán giá laptop dựa trên các yếu tố ảnh hưởng đã được phân tích. Trong quá trình thực hiện đề tài, chúng tôi đã sử dụng các công cụ: BeautifulSoup, Selenium, Pandas, Numpy, Sklearn và PowerBI. Chúng tôi đã áp dụng mô hình hồi quy tuyến tính và Hồi quy đa thức cho việc dự đoán giá Laptop. Sau khi hoàn thành, chúng tôi đã thành công phân tích được các yếu tố ảnh hưởng đến giá Laptop trên thị trường và đã xây dựng thành công mô hình Hồi quy tuyến tính dự đoán giá Laptop dựa trên các yếu tố ảnh hưởng đã được phân tích và chọn lựa.

Bộ dữ liệu phân tích được chúng tôi tự thu thập tại **[1]**. Đây là bộ dữ liệu và đề tài do nhóm tự phân tích thiết kế, không dựa trên đề tài nào khác. Đề tài và bộ dữ liệu này chưa từng được nhóm sử dụng làm đồ án ở bất kì môn nào khác.

2. MÔ TẢ BỘ DỮ LIỆU

Bộ dữ liệu này bao gồm giá của đa dạng các loại laptop từ nhiều nhãn hàng khác nhau đang có trên thị trường. Dữ liệu bao gồm các thuộc tính như Tên sản phẩm, Số nhân, Tốc độ CPU, RAM, Ổ cứng, Dài, Rộng, Giá,…

Bộ dữ liệu được chúng tôi thu thập bằng công cụ BeautifulSoup, Selenium; là thư viện được hỗ trợ bằng ngôn ngữ Python.

Sử dụng Python truy cập đường dẫn **[1]** để đến phần Laptop của trang web Thế Giới Di Động.

Sử dụng Selenium để tải trang web đầy đủ:

- Lưu trang web hiện tại về định dạng HTML thuận tiện cho việc thu thập lại dữ liệu tại thời điểm truy cập.
- Thu thập tất cả URL (*đường dẫn*) của laptop có trong trang web hiện tại và lưu về file ProductLink_list.csv.

Sử dụng BeautifulSoup để truy cập từng URL và lấy toàn bộ thông tin của Laptop.

Bộ dữ liệu phân tích giá Laptop được thu thập tại trang web **[1].**

Bộ dữ liệu bao gồm: 284 Laptop khác nhau và 35 biến.

Trong đó, bao gồm:

- *2 biến liên tục*: 'Giá mới nhất' và 'Số nhân'.
- *33 biến phân loại*: 'Tên sản phẩm', 'RAM', 'Công nghệ CPU', …

STT	Feature	None_null	Type	STT	Feature	None_null	Type
1	Tên sản phẩm	284	object	**19**	Card màn hình	284	object
2	Giá mới nhất	284	int64	**20**	Công nghệ âm thanh	283	object
3	Công nghệ CPU	284	object	**21**	Cổng giao tiếp	284	object
4	Số nhân	284	int64	**22**	Kết nối không dây	284	object
5	Số lượng	284	object	**23**	Tính năng khác	198	object
6	Tốc độ CPU	284	object	**24**	Đèn bàn phím	284	object
7	Tốc độ tối đa	284	object	**25**	Kích thước	284	object
8	Bộ nhớ đệm	284	object	**26**	Khối lượng tịnh	284	object
9	RAM	284	object	**27**	Chất liệu	284	object
10	Loại RAM	284	object	**28**	Thông tin PIN	284	object
11	Tốc độ BUS RAM	284	object	**29**	Công suất bộ sạc	267	object
12	Hỗ trợ RAM tối đa	284	object	**30**	Hệ điều hành	284	object
13	Ổ cứng	284	object	**31**	Thời điểm ra mắt	275	object
14	Màn hình	284	object	**32**	Khe đọc thẻ nhớ	104	object
15	Độ phân giải	284	object	**33**	Tản nhiệt	46	object
16	Tần số quét	284	object	**34**	Màn hình cảm ứng	21	object
17	Độ phủ màu	205	object	**35**	Webcam	284	object
18	Công nghệ màn hình	284	object				

Số lượng giá trị khuyết trên từng Feature trong **Bộ dữ liệu ban đầu**: Độ phủ màu (79), Công nghệ âm thanh (1), Tính năng khác (86), Công suất bộ sạc (17), Thời điểm ra mắt (9), Khe đọc thẻ nhớ (180), Tản nhiệt (238), Màn hình cảm ứng (263).

Ý nghĩa của từng Feature trong **Bộ dữ liệu ban đầu:**

STT	Features	Ý nghĩa	STT	Features	Ý nghĩa
1	Tên sản phẩm	Tên đầy đủ, cụ thể của laptop trên trang web TGDD	**19**	Card màn hình	GPU (*Integrated/Discrete*) xử lý đồ họa (Loại card, Thương hiệu, Thông số,…)
2	Giá mới nhất	Giá bán ngay tại thời điểm Crawl (có áp dụng giảm giá)	**20**	Công nghệ âm thanh	Các công nghệ cải thiện âm thanh

3	Công nghệ CPU	Loại vi xử lý (CPU) và công nghệ đi kèm	**21**	Cổng giao tiếp	Các cổng kết nối vật lý
4	Số nhân	Số lõi vật lý của CPU	**22**	Kết nối không dây	Chuẩn kết nối không dây của thiết bị. (Wifi, Bluetooth, HDMI,…)
5	Số luồng	Số luồng xử lý của CPU	**23**	Webcam	Thông số camera tích hợp trên thiết bị
6	Tốc độ CPU	Tốc độ cơ bản của CPU (*GHz*)	**24**	Tính năng khác	Các tính năng bổ sung (Bảo mật vân tay, Mở khóa khuôn mặt,…)
7	Tốc độ tối đa	Tần số tối đa của CPU khi kích hoạt chế độ tăng tốc (Turbo Boost).	**25**	Đèn bàn phím	Đèn nền của bàn phím (Hỗ trợ trong môi trường thiếu sáng)
8	Bộ nhớ đệm	Dung lượng bộ nhớ đệm (*cache*) của CPU	**26**	Kích thước	Kích thước vật lý (Dài x Rộng x Dày)
9	RAM	Dung lượng bộ nhớ tạm thời (RAM) của máy	**27**	Khối lượng tịnh	Trọng lượng của sản phẩm
10	Loại RAM	Công nghệ RAM	**28**	Chất liệu	Vật liệu chế tạo (nhựa, kim loại)
11	Tốc độ Bus RAM	Tần số hoạt động của RAM (*MHz*)	**29**	Thông tin Pin	Loại pin và dung lượng pin (*Wh*)
12	Hỗ trợ RAM tối đa	Dung lượng RAM lớn nhất mà máy hỗ trợ nâng cấp.	**30**	Công suất bộ sạc	Công suất của sạc (*W*)
13	Ổ cứng	Loại và dung lượng ổ cứng	**31**	Hệ điều hành	Hệ điều hành đi kèm (Windows, macOS, Linux,…)
14	Màn hình	Kích thước màn hình (*Inch)*	**32**	Thời điểm ra mắt	Thời gian sản phẩm được phát hành ra thị trường.
15	Độ phân giải	Độ sắc nét của màn hình	**33**	Khe đọc thẻ nhớ	Khả năng hỗ trợ đọc thẻ nhớ. (SD, microSD)
16	Tần số quét	Tốc độ làm mới màn hình (*Hz*)	**34**	Tản nhiệt	Công nghệ làm mát.
17	Độ phủ màu	Khả năng hiển thị màu sắc	**35**	Màn hình cảm ứng	Khả năng hỗ trợ cảm ứng trên màn hình.
18	Công nghệ màn hình	Loại công nghệ			

Dựa vào hiểu biết về thị trường Laptop, chúng tôi đã chọn lọc 18/34 biến độc lập có khả năng ảnh hưởng đến giá của 1 chiếc Laptop trong **Bộ dữ liệu ban đầu** và trình bày lại **Bộ dữ liệu mới** gồm 22 biến độc lập mới nhằm cho công việc phân tích thuận tiện và trơn tru hơn.

Dưới đây là **Bộ dữ liệu mới** mà chúng tôi đã chọn lọc, trình bày lại và nêu phương thức sử dụng:

STT	Tên Feature mới	Cách xử lý	Trích dẫn từ Feature	STT	Tên Feature mới	Cách xử lý	Trích dẫn từ Feature

1	Brand	Lưu lại tên thương hiệu	Tên sản phẩm	**13**	MaxSpeed_CPU	Lưu lại giá trị và loại bỏ đơn vị Ghz và các thông tin khác.	Tốc độ tối đa
2	Type	Lưu lại tên loại máy	Tên sản phẩm	**14**	Cache	Lưu lại giá trị và loại bỏ đơn vị Mb	Bộ nhớ đệm
3	Brand_CPU	Lưu lại tên thương hiệu	Công nghệ CPU	**15**	RAM	Lưu lại giá trị và loại bỏ đơn vị GB	RAM
4	Type_CPU	Loại vi xử lý và công nghệ đi kèm VD: Core_i5, Ryzen_7…	Công nghệ CPU	**16**	BusSpeed_RAM	Lưu lại giá trị và loại bỏ đơn vị Mhz	Tốc độ Bus RAM
5	Hard_Drive	Lưu lại giá trị và loại bỏ đơn vị GB cùng các thông tin khác.	Ổ cứng	**17**	MaxSup_RAM	Nếu có hỗ trợ sẽ lưu lại giá trị và loại bỏ đơn vị. Nếu không hỗ trợ sẽ lưu giá trị 0.	Hỗ Trợ Ram tối đa
6	Length	Lưu lại giá trị độ dài	Kích thước	**18**	Screen_Size	Lưu lại giá trị kích cỡ và loại bỏ kí hiệu đơn vị Inch	Màn hình
7	Width	Lưu lại giá trị độ rộng	Kích thước	**19**	Refresh_Rate	Lưu lại giá trị và loại bỏ đơn vị Hz	Tần số quyét
8	Thick	Lưu lại giá trị độ dày	Kích thước	**20**	Weight	Lưu lại giá trị khối lượng và loại bỏ đơn vị Kg	Khối lượng tịnh
9	Key_Light	Chỉ lưu giá trị "Không có đèn" hoặc nếu có thì lưu giá trị về loại đèn. Loại bỏ thông số về màu	Đèn bàn phím	**21**	Charging_Power	Lưu lại giá trị công suất và loại bỏ đơn vị W	Công suất bộ sạc
10	Cores_CPU	GIữ nguyên	Số nhân	**22**	Release	Lưu lại hai giá trị cuối của năm. VD: 2023 → 23	Thời điểm ra mắt
11	Threads_CPU	Giữ nguyên	Số luồng	**23**	Latest_Price	Giữ nguyên	Giá mới nhất
12	Speed_CPU	Lưu lại giá trị và loại bỏ đơn vị Hz	Tốc độ CPU				

Chúng tôi nhận thấy rằng trong 2 bộ dữ liệu đã trình bày đều tồn tại 2 vấn đề:

1. ***Dữ liệu của dòng Macbook thuộc Apple***: không cùng phân khúc với hầu hết các loại máy khác, sẽ gây nhiễu cho Bộ dữ liệu. Vì vậy chúng tôi lựa chọn loại bỏ tất cả dòng Macbook.
2. ***Dữ liệu của một vài laptop không được nhãn hàng công bố***: sẽ được phân về giá trị NaN và được xử lý chung với các giá trị bị khuyết khác.

Thông tin Bộ dữ liệu mới sau khi đã xử lý:

- Số lượng biến liên tục: 5
- Số lượng biến kiểu phân loại: 18

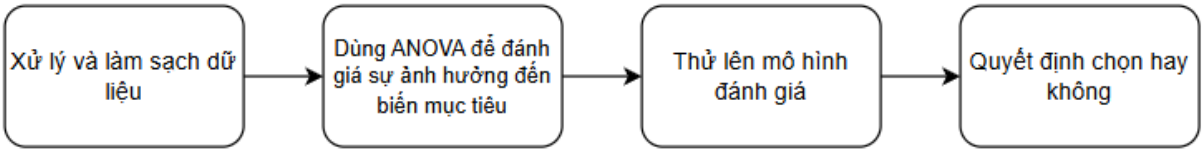
STT	Feature	None_null	Type	STT	Feature	None_null	Type
1	Brand	273	object	**13**	MaxSpeed_CPU	273	float64
2	Type	273	object	**14**	Cache	273	float64
3	Brand_CPU	273	object	**15**	RAM	273	float64
4	Type_CPU	273	object	**16**	BusSpeed_RAM	273	float64
5	Hard_Drive	273	float64	**17**	MaxSup_RAM	273	float64
6	Length	273	float64	**18**	Screen_Size	273	float64
7	Width	273	float64	**19**	Refresh_Rate	273	float64
8	Thick	273	float64	**20**	Weight	273	float64
9	Key_Light	273	object	**21**	Charging_Power	273	float64
10	Cores_CPU	273	int64	**22**	Release	273	float64
11	Threads_CPU	273	float64	**23**	Latest_Price	273	int64
12	Speed_CPU	273	float64	**24**		273	

3. PHƯƠNG PHÁP PHÂN TÍCH

Sau khi xử lý các biến, chúng tôi đã có được bộ dữ liệu đã được làm sạch và bắt đầu tới bước Phân tích và Đánh giá các biến.

1. Xử lý các biến phân loại:

Quy trình xử lý:



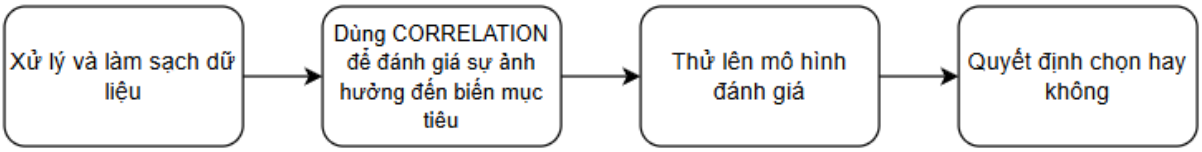
Bảng hệ số:

Biến	Chênh lệch phương sai	P_value	Biến	Chênh lệch phương sai	P_value
Brand	5.347591	1.055349e-04	Key_Light	64.843696	1.004263e-23
Type	6.320122	1.002697e-18	Screen_Size	22.368878	5.882569e-26
Type_CPU	22.773934	7.284644e-36	Brand_CPU	0.459609	6.320240e-01
Hard_Drive	99.142248	3.043909e-43	MaxSup_RAM	10.543747	7.797836e-13

Chúng tôi sử dụng phương pháp ANOVA để chọn những cột ảnh hưởng đến biến mục tiêu theo tiêu chí P_values ≤ 0.05 và ta chọn các biến là: **Brand, Type, Type_CPU, Hard_Drive, Key_Light, Screen_Size.**

2. Xử lý các biến liên tục:

Quy trình xử lý:



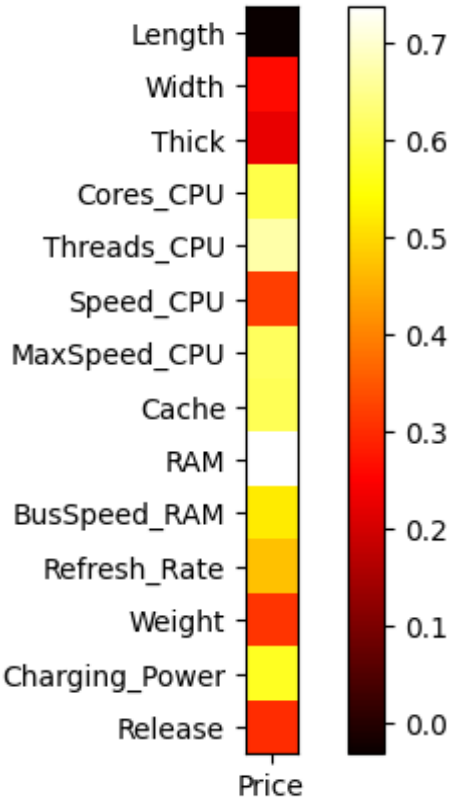
Sử dụng ĐỘ TƯƠNG QUAN, chúng tôi chọn ra được những biến có ảnh hưởng tốt đến biến mục tiêu theo tiêu chí:

ĐỘ TƯƠNG QUAN bé hơn - 0.3 hay lớn hơn 0.3

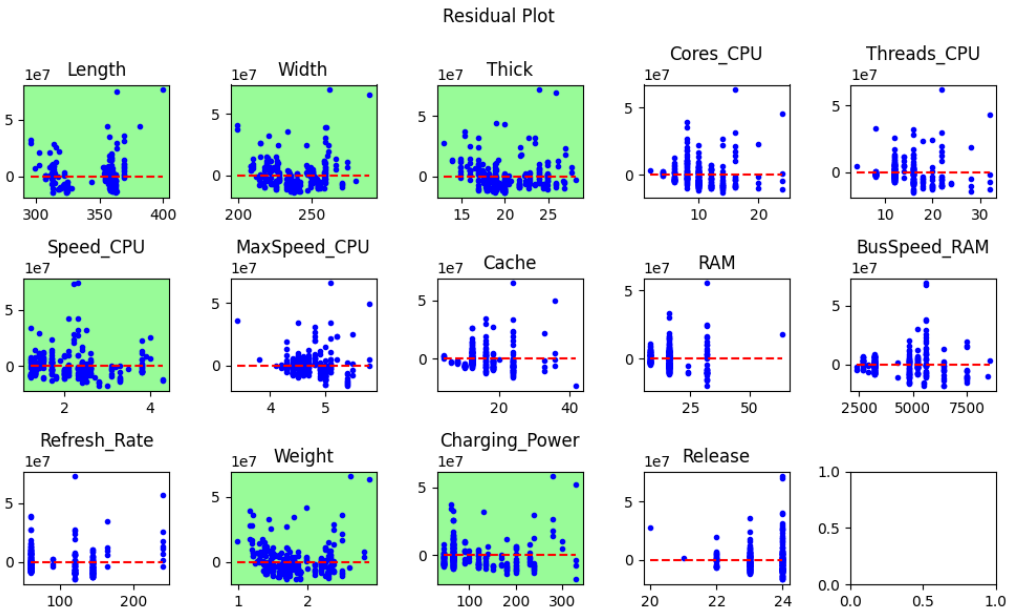
Chúng tôi chọn được các biến là :

Cores_CPU, Threads_CPU, Speed_CPU, MaxSpeed_CPU, Cache, RAM, BusSpeed_RAM, Refresh_Rate, Weight, Charging_Power.

Sau quá trình đánh giá và chọn lọc, chúng tôi sử dụng biểu đồ Residual để chọn các biến thực sự phù hợp với quá trình huấn luyện.

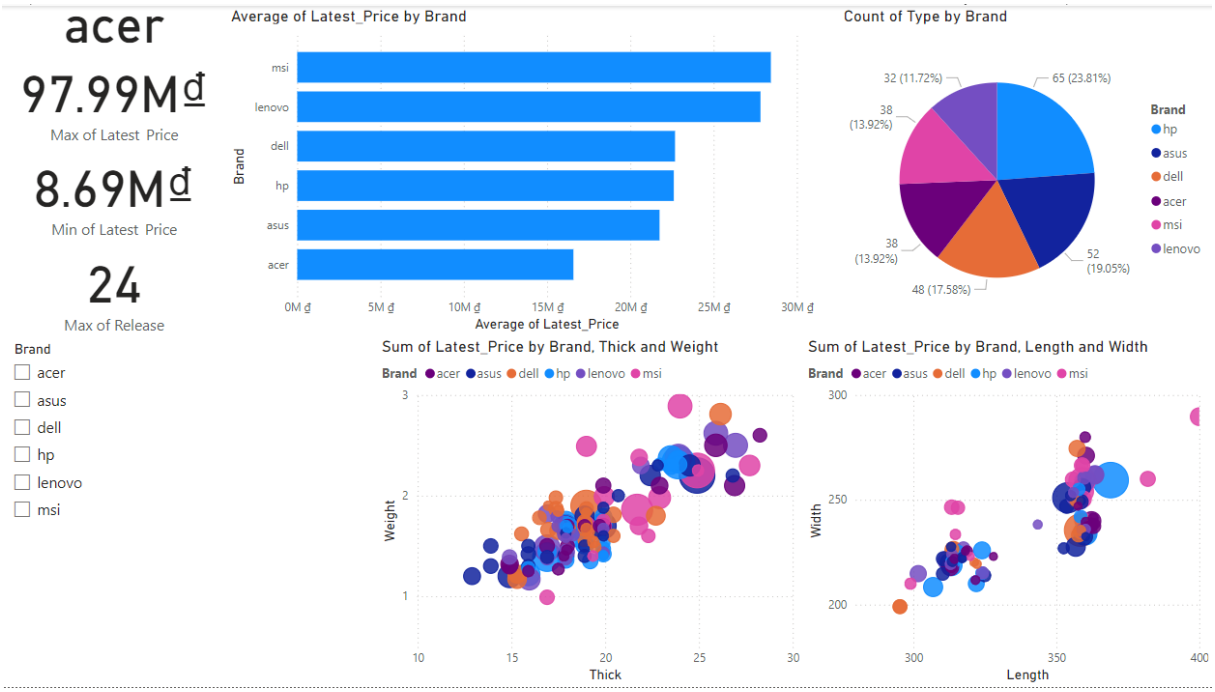


Biểu đồ Residual:



Dựa vào các biểu đồ trên, chúng tôi nhận thấy rằng các biến: **Length, Width, Thick, Speed_CPU, Weight, Charging_Power** sẽ phù hợp với mô hình đơn bội, còn các biến còn lại dường như không phù hợp để huấn luyện mô hình.

4. PHÂN TÍCH THĂM DÒ



Chúng tôi có thấy được loại laptop có giá tiền lớn nhất có giá khoảng 98 triệu và giá tiền nhỏ nhất có giá khoảng 8,7 triệu.

Bộ dữ liệu gồm có 6 hãng laptop: HP, Asus, Dell, Acer, MSI, Lenovo.

Trong đó, số loại laptop mà mỗi hãng sản xuất gồm có:

- HP có 65 loại (23,81%).
- Asus có 52 loại (19,05%).
- Dell có 48 loại (17,58%).
- Acer có 38 loại (13,92%).
- MSI có 38 loại (13,92%).
- Lenovo có 32 loại (11,72%).

Nhìn vào biểu đồ cột, chúng tôi thấy được hãng có giá trung bình cao nhất là **MSI** (khoảng 29 triệu), ngược lại hãng có giá trung bình thấp nhất là **Acer** (khoảng 17 triệu).

Dựa vào biểu đồ Scatter tương quan giá độ nặng (*Weight*) và độ dày (*Thick*), ta thấy được giá của một sản phẩm bị ảnh hưởng bởi độ nặng (*Weight*) và độ dày (*Thick*). Nếu độ nặng và độ dày của một sản phẩm càng cao thì giá của sản phẩm cũng sẽ càng cao.

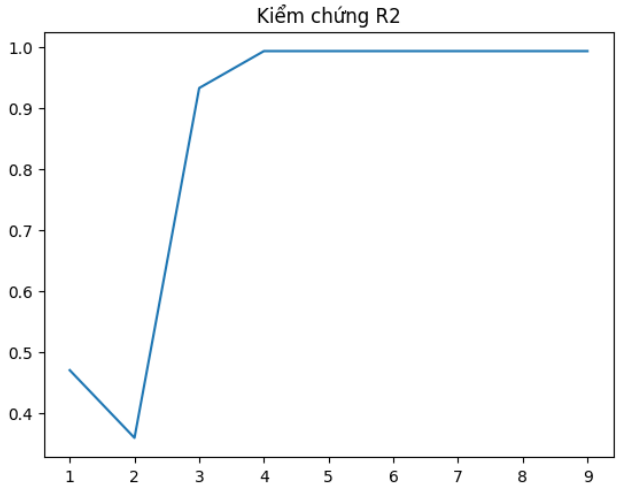
Ví dụ:

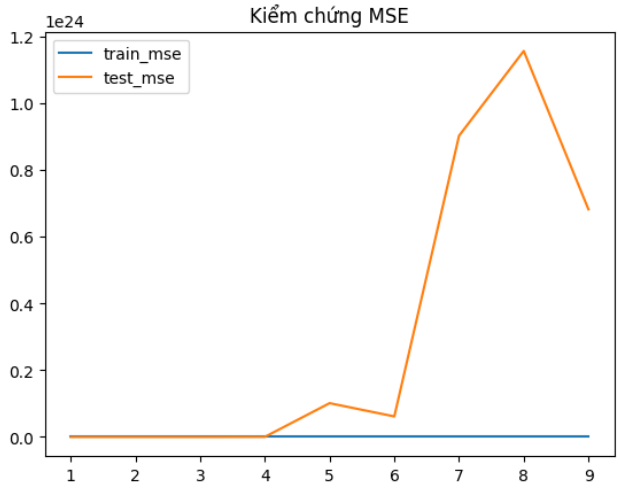
Điều này cũng được áp dụng tương tự trên biểu đồ Scatter tương quan giữa độ dài (Length) và độ rộng (Width).

Ví dụ:

Sau khi đã cân nhắc các biến thông qua phương pháp Correlation và ANOVA, chúng tôi quyết định chọn 6 biến bao gồm: **Length, Width, Thick, Speed_CPU, Weight, Charging_Power**.

Dưới đây lần lượt là kiểm chứng **R-Squared** và kiểm chứng **MSE**. Hai kiểm chứng này sẽ là tiền đề để chúng tôi chọn ra được bậc phù hợp với mô hình và sẽ được trình bày cụ thể trong phần **Kết quả phân tích.**





Nhóm cũng sử dụng thêm phần kiểm chứng chéo để đánh giá mô hình:

Bảng chỉ số MSE, R2 của kiểm chứng chéo

Số CV	1	2	3	4	5	Trung bình
MSE	-3.846e+13	-3.656e+13	-3.157 e+13	-7.99e+13	-3.232e+14	-101965671816249
R2	-1.80154297	-0.45869742	0.60596699	-0.543862	-0.0726734	-0.454161

5. KẾT QUẢ PHÂN TÍCH

Sau kiểm chứng R2 và MSE, chúng tôi chọn bậc 3 có chỉ số R2 cao và MSE thấp để chạy polynomial sau đó huấn luyện cho mô hình. Nhưng sau khi huấn luyện mô hình, chúng tôi đã thử tính lại giá trị R2 và MSE của mô hình và kết quả là:

R2	MSE
0.9936128614911366	3.7677164732654544e+20

Mặc dù có chỉ số R2 rất tốt những chỉ số MSE lại rất rất cao, do đó chúng tôi quyết định bỏ qua mô hình có sử dụng polynomial vì mô hình có vẻ đang bị quá khớp (overfitting).

Vì vậy, chúng tôi đã sử dụng lại một cách chọn cột khác để chọn lọc các cột dùng trong huấn luyện mô hình đó là sử dụng các có hệ số tốt (Độ tương quan – Biến liên tục,

ANOVA – Biến phân loại) và sử dụng toàn bộ các cột có trong dữ liệu để huấn luyện mô hình và nhận được kết quả là:

Loại	R2	MSE
Hệ Số Tốt	0.8313110597254117	104009264929825.83
Tất Cả Cột	0.8492594461514338	93996129153960.34

Mặc dù chỉ số R2 của hai mô hình này không được cao như của mô hình có sử dụng polynomial nhưng chỉ số MSE lại thấp hơn rất rất nhiều so với mô hình có sử dụng polynomial. Do đó, chúng tôi quyết định sử dụng mô hình sử dụng tất cả các cột để làm mô hình chính của dự án dự đoán giá laptop của nhóm chúng tôi

6. KẾT LUẬN

Trong đề tài này, chúng tôi đã thu thập dữ liệu về giá và các thông tin liên quan của các mẫu Laptop đã và đang được bày bán tại Thế Giới Di Động. Dựa vào những kiến thức đã được học và kiến thức thực tế, chúng tôi đã chọn lọc các yếu tố ảnh hưởng đến giá laptop. Sau quá trình làm sạch và tiền xử lý, chúng tôi đã có thể xây dựng được mô hình dự đoán giá laptop dựa trên các yếu tố đó.

Mô hình hoạt động ổn định với độ sai số MSE 93,996,129,153,960.34. Điều này chứng tỏ dự đoán của mô hình sai lệch với thực tế ở khoảng mức 9 triệu đồng. So với mức giá laptop mới trung bình ở mức 9-100 triệu đồng thì đây là một mức sai số có thể chấp nhận được. Mô hình này có giá trị sử dụng giúp người dùng dự đoán được ngân sách dự phòng trước khi mua Laptop.

TÀI LIỆU THAM KHẢO

[1] Thế Giới Di Động. Link: https://www.thegioididong.com/laptop (24/10/2024)

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Mã Kim Phát	-Tiền xử lý dữ liệu -Trực quan hóa dữ liệu bằng PowerBI và kiểm chứng bậc phù hợp -Viết báo cáo và làm slide
2	Võ Tấn Trung	-Tiền xử lý dữ liệu -Viết demo -Phân loại biến dùng để huấn luyện mô hình -Viết báo cáo và làm slide
3	Đặng Chí Nguyên	-Tiền xử lý dữ liệu -Viết báo cáo và làm slide -Kiểm chứng và đánh giá chọn bậc phù hợp
4	Nguyễn Tiến Nam	-Thu thập dữ liệu -Tiền xử lý dữ liệu -Viết demo -Viết báo cáo và làm slide

Link GitHub Demo: https://github.com/Trung1573/DS105-Project-Predict-Price-Of-Laptop