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

DỰ ĐOÁN GIÁ LAPTOP CŨ TRÊN TRANG THƯƠNG MẠI ĐIỆN TỬ NEWEGG

Sinh viên thực hiện:			
STT	Họ tên	MSSV	Ngành
1	Lương Lý Công Thịnh	20521960	KHMT
2	Lê Hoàng Thoại	20521976	KHMT
3	Nguyễn Ngọc Tín	20522015	KHMT

TP. HỒ CHÍ MINH – 12/2023

1. GIỚI THIỆU

1.1. Giới thiệu đề tài

Nhóm sẽ thực hiện việc phân tích và đánh giá đề tài này thông qua việc sử dụng các mô hình để dự đoán giá của các laptop cũ được bán trên trang thương mại điện tử Newegg.

Ban đầu, nhóm đã khảo sát và thu thập dữ liệu từ trang Newegg. Tiếp theo, nhóm tiến hành một quá trình phân tích cẩn thận để lựa chọn những đặc tính quan trọng nhất. Cuối cùng, nhóm ứng dụng các mô hình máy học để dự đoán giá của từng sản phẩm.

Các công cụ, giải pháp, thuật toán áp dụng cho từng giai đoạn của đề tài mà nhóm sử dụng như sau:

- Giai đoạn 1: Thu thập dữ liệu: sử dụng thư viện Selenium, BeautifulSoup.
- Giai đoạn 2: Phân tích, đánh giá dữ liệu: sử dụng thư viện: seaborn, pandas, numpy, matplotlib.
- Giai đoạn 3: Áp dụng mô hình để dự đoán giá: sử dụng mô hình Linear Regression, Support Vector Machine, Random Forest Regression.

Nhóm đã thực hiện một quá trình phân tích chi tiết và kỹ lưỡng để xác định các đặc tính quan trọng ảnh hưởng đến giá của các laptop cũ từ trang Newegg. Kết quả của việc phân tích dữ liệu đã cung cấp cái nhìn sâu sắc về yếu tố nào có tác động đáng kể đến giá của sản phẩm.

1.2. Tính minh bạch của đề tài

- Bộ dữ liệu được nhóm tự thu thập tại trang thương mại điện tử Newegg [1]
- Bộ dữ liệu và đề tài do nhóm tự phân tích thiết kế và không dựa trên đề tài nào khác.
- Bộ dữ liệu và đề tài của nhóm được lựa chọn dựa trên ý tưởng của trang web thương mại điện tử Chotot [2] do thầy gợi ý. Nhóm đã quyết định sử dụng trang Newegg vì trang web cung cấp nhiều thông tin hơn về cùng một sản phẩm so với trang Chotot.
- Bộ dữ liệu này được dùng làm đồ án môn học môn Phân tích dữ liệu vào học kì I năm học 2023-2024 và chưa sử dụng cho môn học nào khác.
- Nhóm đã tiến hành thu thập dữ liệu từ trang Newegg dự trên seminar [3] của một nhóm về thu thập dữ liệu trên lớp

2. MÔ TẢ BỘ DỮ LỆU

Bộ dữ liệu này chứa thông tin về các laptop được bán trên trang thương mại điện tử Newegg.

Bộ dữ liệu được nhóm tự thu thập tại trang thương mại điện tử Newegg [1] và được tham khảo cách thu thập dữ liệu từ trang Newegg dự trên seminar [3] của một nhóm về thu thập dữ liệu trên lớp.

Phương pháp thu thập dữ liệu của nhóm được tiến hành như sau:

- Lựa chọn trang thương mại điện tử để thu thập dữ liệu:
  - + Nhóm của bạn đã thực hiện nghiên cứu và thu thập dữ liệu về sản phẩm laptop từ một số trang thương mại điện tử ở Việt Nam cũng như quốc tế như Chotot, Cellphones [4], eBay [5], và Amazon [6]. Tuy nhiên, kết quả thu được không đạt mong đợi do một số vấn đề như thiếu thông tin hoặc sự không đồng bộ giữa các thông tin của các sản phẩm.
  - + Sau quá trình đánh giá, nhóm của bạn đã quyết định chọn trang web Newegg làm nguồn thông tin chính để thu thập dữ liệu về sản phẩm laptop. Lựa chọn này được đánh giá là mang lại dữ liệu chất lượng và đáng tin cậy nhất cho nghiên cứu của nhóm.
- Thu thập dữ liệu:
  - + Cài đặt các thư viện cần thiết để tiến hành thu thập dữ liệu: Selenium [7], BeautifulSoup, time.sleep.
  - + Sử dụng WebDriver của Selenium để mở trang web Newegg và lặp qua các trang mong muốn
  - + Dùng BeautifulSoup để phân tích mã nguồn HTML của trang đã tải.
  - + Tìm tất cả các container chứa thông tin sản phẩm trên trang.Với mỗi container, trích xuất thông tin như tiêu đề, link, giá cả và các thông số kỹ thuật.
  - + Tạo DataFrame từ dữ liệu thu thập được. Lưu dữ liệu vào file CSV để sử dụng và phân tích sau này.

Mô tả bộ dữ liệu:

- Bộ dữ liệu bao gồm 1452 dòng và 19 cột. Mỗi dòng tương ứng với một sản phẩm, mỗi cột tương ứng với thông tin như giá, thương hiệu, ….

STT	Thuộc tính	Kiểu dữ liệu	Mô tả
1	Price	float64	Giá của sản phẩm, tính bằng USD
2	Brand	object	Thương hiệu của laptop

3	Screen Size	float64	Kích thước màn hình của laptop, được đo bằng inch.
4	CPU type	object	Loại bộ vi xử lý (CPU) của laptop
5	Memory	int64	Dung lượng bộ nhớ RAM của laptop, được tính bằng GB
6	Storage	int64	Dung lượng lưu trữ của laptop, được tính bằng GB
7	GPU	object	Card đồ họa của laptop
8	Resolution	object	Độ phân giải màn hình của laptop
9	Weight	float64	Trọng lượng của laptop, được tính bằng lbs
10	Backlit Keyboard	object	Tính năng có bàn phím có đèn nền hay không
11	Touchscreen	object	Tính năng màn hình cảm ứng có được hỗ trợ hay không
12	Graphic Type	object	Loại đồ họa của laptop, có thể là tích hợp (Integrated) hoặc có card đồ họa rời (Dedicated).
13	Operating System	object	Hệ điều hành của laptop
14	Webcam	object	Tính năng có webcam tích hợp hay không
15	Card Reader	object	Tính năng có đầu đọc thẻ nhớ tích hợp hay không
16	Thunderbolt	object	Tính năng có cổng Thunderbolt hay không
17	CPU model	object	Mô hình cụ thể của bộ vi xử lý (CPU)
18	title	object	Tiêu đề sản phẩm
19	link	object	Đường dẫn đến trang sản phẩm trên Newegg

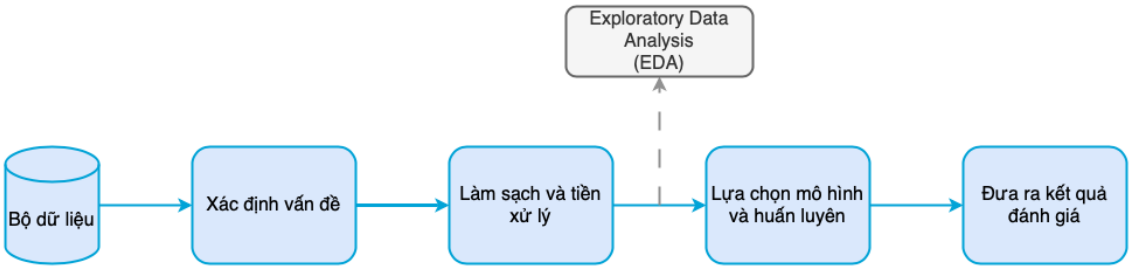
Bộ dữ liệu này chứa thông tin về các sản phẩm laptop được bán trên Newegg. Mỗi hàng trong dataframe đại diện cho một sản phẩm cụ thể. Tuy nhiên, do mỗi sản phẩm trên Newegg có các thông tin khác nhau, nên có thể có một số cột trong mỗi hàng được gán giá trị là N/A để thể hiện sự thiếu thông tin.

Bộ dữ liệu chứa cả biến số và biến phân loại:

- Biến số: Price, Screen Size, Memory, Storage, Resolution , Weight
- Biến phân loại: Brand, CPU type, GPU, Backlit Keyboard, Touchscreen, Graphic Type, Operating System, Webcam, Card Reader, Thunderbolt, CPU model, title, link

Bằng cách trình bày mô tả bộ dữ liệu này trước khi tiến hành phân tích, người dùng sẽ hiểu rõ hơn về cấu trúc và ý nghĩa của các biến và cột dữ liệu, giúp chuẩn bị tốt cho quá trình phân tích dữ liệu và trích xuất thông tin hữu ích từ bộ dữ liệu.

3. PHƯƠNG PHÁP PHÂN TÍCH



Hình 1: Quy trình PTDL

Về phương pháp phân tích dữ liệu, nhóm tiếp cận theo cách sau:

3.1. Xác định vấn đề

Từ bộ dữ liệu đã thu thập ở trên nhóm tiến hành xác định các vấn đề cho đề tài như sau:

- Xác định biến mục tiêu: Biến mục tiêu chính là giá của laptop cũ. Dự đoán giá dựa trên thông tin khác về sản phẩm.
- Đánh giá độ sạch của dữ liệu: Dữ liệu ít thiếu và không có nhiều giá trị trùng lặp. Cần xử lý giá trị thiếu để đảm bảo thông tin đầy đủ và chính xác.
- Các biến ảnh hưởng đến giá: Thương hiệu, kích thước màn hình, loại CPU, bộ nhớ, dung lượng lưu trữ, card đồ họa, hệ điều hành và các thông số kỹ thuật có thể ảnh hưởng đến giá của laptop.
- Lựa chọn mô hình: Sử dụng mô hình hồi quy tuyến tính, hồi quy phi tuyến hoặc mạng nơ-ron để dự đoán giá của laptop dựa trên thông tin từ bộ dữ liệu.
- Các độ đo đánh giá mô hình: Sử dụng các độ đo như Mean Squared Error (MSE), Root Mean Squared Error (RMSE), hoặc R-squared để đo lường độ chính xác và hiệu suất của mô hình dự đoán.

3.2. Tiền xử lý và làm sạch dữ liệu

Quá trình làm sạch bộ dữ liệu như sau:

- Xác định và loại bỏ các dòng trùng lặp để giữ lại chỉ một bản ghi duy nhất trong trường hợp có dữ liệu trùng nhau.
- Đếm số lượng giá trị null (missing value) trong mỗi cột của DataFrame.
    + Kết quả cho thấy có 3 cột bị missing value là “Resolution”, “Weight”, “CPU model”

+ Số lượng missing của Resolution là tương đối thấp nên nhóm điền các giá trị null trong cột "Resolution" bằng giá trị mode (giá trị xuất hiện nhiều nhất) của cột đó.
+ Với 2 cột “Weight”, “CPU model” nhóm tiến hành lược bỏ khỏi bộ dữ liệu

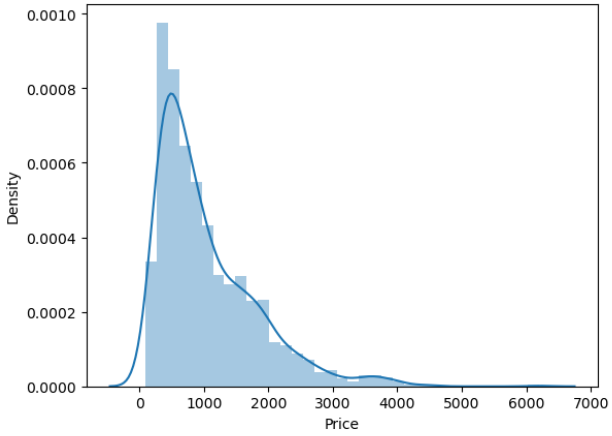
– Kiểm tra lại sau khi xử lý, đảm bảo không còn cột nào chứa giá trị null.

Quá trình tiền xử lý dữ liệu được thực hiện qua từng cột như sau

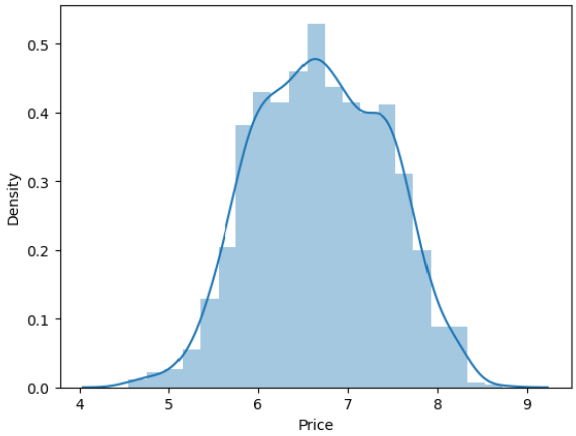
– “Price”

+ Sử dụng phương pháp IQR [7] để xác định và loại bỏ các giá trị outlier trong cột “Price” của dataframe.
+ Tính toán giá trị lower_bound và upper_bound dựa trên IQR.
+ Tạo dataframe mới chỉ chứa các dòng dữ liệu mà giá “Price” nằm trong khoảng giữa lower_bound và upper_bound.

Lí do sử dụng IQR bởi vì trước khi sử dụng thì phân phối có xu hướng lệch trái, sau khi sử dụng thì phân phối đã đồng đều hơn có thể giúp cải thiện hiệu suất mô hình.



before



after

– “Backlit Keyboard”, “Thunderbolt”, “Card Reader”, “Touchscreen”

+ Sau quá trình xem xét nhóm quyết định gom dữ liệu thành 2 nhóm là “yes” và “no” để giảm độ phức tạp của biến do mỗi nhà bán hàng có các cách ghi khác nhau và không có thì hầu như là họ để trống.
+ Các cột này có 2 giá trị là “yes”, “no” vì thế nhóm sẽ tiến hành thay thế giá trị bằng “0” cho “no” và “1” cho yes

- "Resolution"
  - Tách "Resolution" thành chiều rộng và chiều cao.
  - Chuyển đổi kiểu dữ liệu thành số nguyên
  - Tính toán độ phân giải điểm ảnh trên mỗi inch (PPI [8]):
  - Áp dụng công thức PPI = căn bậc hai của tổng bình phương chiều rộng và chiều cao / kích thước màn hình.
  - Sử dụng giá trị chiều rộng, chiều cao từ hai cột mới và kích thước màn hình từ cột "Screen Size" để tính toán PPI cho từng sản phẩm.
  - Thêm cột "ppi" vào bộ dữ liệu
- "CPU" và "GPU":
  - Xác định danh sách các loại CPU và GPU
  - Tạo cột mới ("CPU_series" từ "CPU type" và "GPU_brand" từ "GPU""")
  - Phân loại các giá trị thành các nhóm như Intel Core, AMD Ryzen, NVIDIA, AMD, hoặc Other nếu không thuộc danh sách đã xác định.

3.3. Lựa chọn và huấn luyện mô hình

Quá trình huấn luyện mô hình bao gồm các bước sau:

- Chọn biến đầu vào và đầu ra:
  - Chọn các biến đầu vào ("features") và biến đầu ra ("target") từ dữ liệu.
- Chia dữ liệu:
  - Sử dụng hàm train_test_split để chia dữ liệu thành tập huấn luyện (X_train, y_train) và tập kiểm tra (X_test, y_test) với tỷ lệ 80-20%.
- Xây dựng pipeline tiền xử lý:
  - Xác định các đặc trưng số ("numeric_features") và đặc trưng phân loại ("categorical_features").
  - Sử dụng ColumnTransformer để thực hiện các bước tiền xử lý như chuẩn hóa đặc trưng số và mã hóa đặc trưng phân loại.
- Xây dựng mô hình:
  - Lựa chọn mô hình hồi quy RandomForestRegressor với các siêu tham số đã cài đặt trước.

- Xây dựng pipeline với mô hình:
  + Tạo pipeline kết hợp quá trình tiền xử lý và mô hình hồi quy.
- Huấn luyện mô hình:
  + Sử dụng dữ liệu huấn luyện (X_train, y_train) để huấn luyện mô hình thông qua pipeline đã xây dựng.

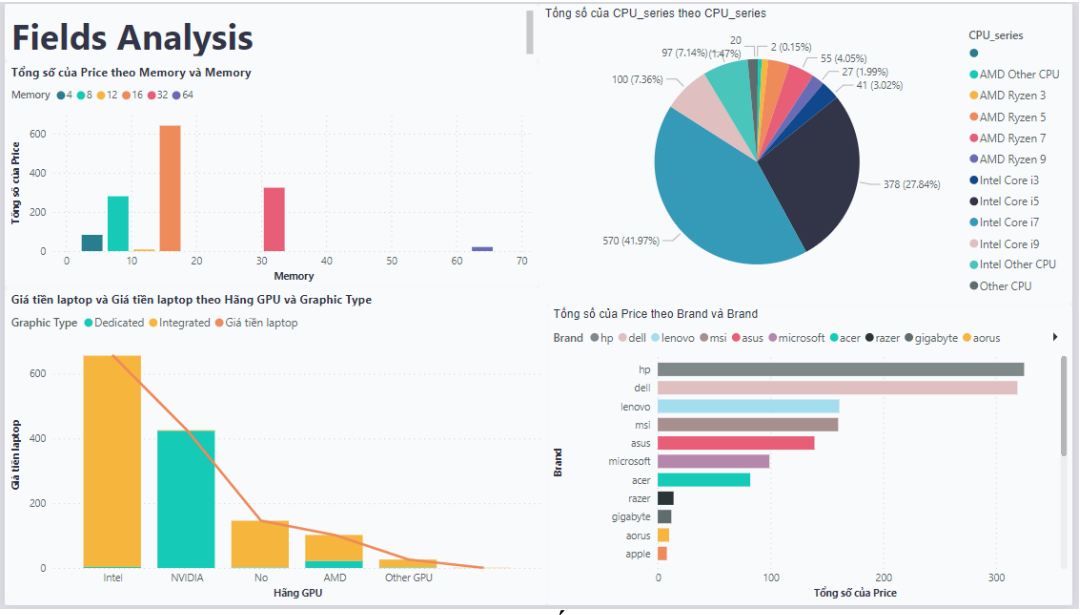
Đây là quá trình cơ bản để huấn luyện mô hình hồi quy RandomForestRegressor sử dụng các biến được chọn và tiền xử lý dữ liệu

3.4. Đưa ra kết quả đánh giá

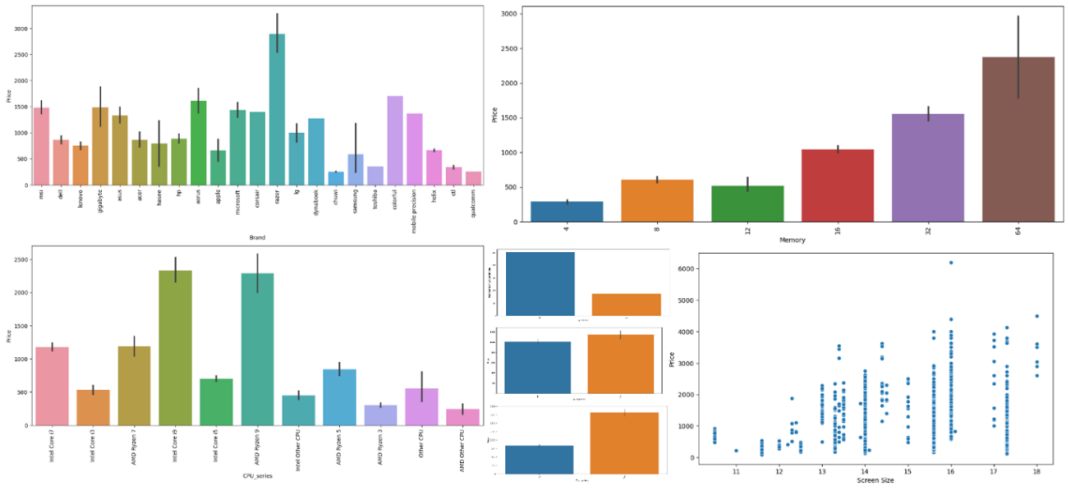
Nhóm sử dụng 3 độ đo đánh giá hiệu suất của mô hình dự đoán giá laptop cũ:

- **R^2 Score (Coefficient of Determination):** Đo lường mức độ mô hình giải thích phương sai của dữ liệu. Giá trị R^2 dao động từ 0 đến 1, với 1 là hoàn hảo. Mức độ gần 1 cho thấy mô hình giải thích tốt phương sai của dữ liệu.
- **MAE Score (Mean Absolute Error):** Tính trung bình độ lỗi tuyệt đối giữa dự đoán và giá trị thực tế. Độ đo này không bị ảnh hưởng nhiều bởi các điểm ngoại lai và thể hiện độ lớn trung bình của sai số.
- **RMSE Score (Root Mean Squared Error):** Tính căn bậc hai của độ lỗi bình phương trung bình giữa dự đoán và giá trị thực tế. Nó tập trung vào các sai số lớn hơn và nhạy cảm hơn đối với các điểm ngoại lai so với MAE.

4. PHÂN TÍCH THĂM DÒ



Hình 2: Phân tích các biến qua Power BI



Nhóm đã thực hiện việc tạo biểu đồ để phân tích các yếu tố có ảnh hưởng đến giá của laptop cũ, gồm:

- "**Brand"**: Phân phối giá laptop theo thương hiệu có sự phân tán khá lớn, với các thương hiệu cao cấp có giá trung bình cao hơn đáng kể các hang khác.
- "**Backlit Keyboard"**, "**Thunderbolt"**, "**Card Reader"**,: Hầu hết các biến có kiểu dữ liệu "yes" "no" thể hiện trên biểu đồ thì các biến có giá trị "yes" có giá trung bình cao hơn so với biến giá trị "no".
- "**CPU series**", "**GPU brand**" có ảnh hưởng đến "Price" của laptop. Các CPU, GPU có hiệu năng cao hơn thường có giá cao hơn.
- "**Memory"** có tỉnh lệ thuận với "Price", khi giá trị "Memory" càng lớn thì "Price" càng lớn
- "**Storage"** hầu như không có xu hướng với "Price", bởi vì bộ dữ liệu không phân chia giữa ssd, hdd và tốc độ đọc dữ liệu của bộ nhớ khác nhau rất nhiều.
- **"Screen Size"** Giá laptop thường tăng khi kích thước màn hình tăng, nhưng không theo quy luật tuyến tính. Một số laptop 17 inch có giá thấp hơn so với các laptop có màn hình nhỏ hơn, chỉ ra sự biến đổi không đồng đều trong mối quan hệ này.

Đối với biến số:

Giá trung bình của laptop dựa trên các biến khác nhau. Các biến được xem xét là kích thước màn hình, bộ nhớ, dung lượng lưu trữ, bàn phím có đèn nền, đầu đọc thẻ, Thunderbolt và độ phân giải.

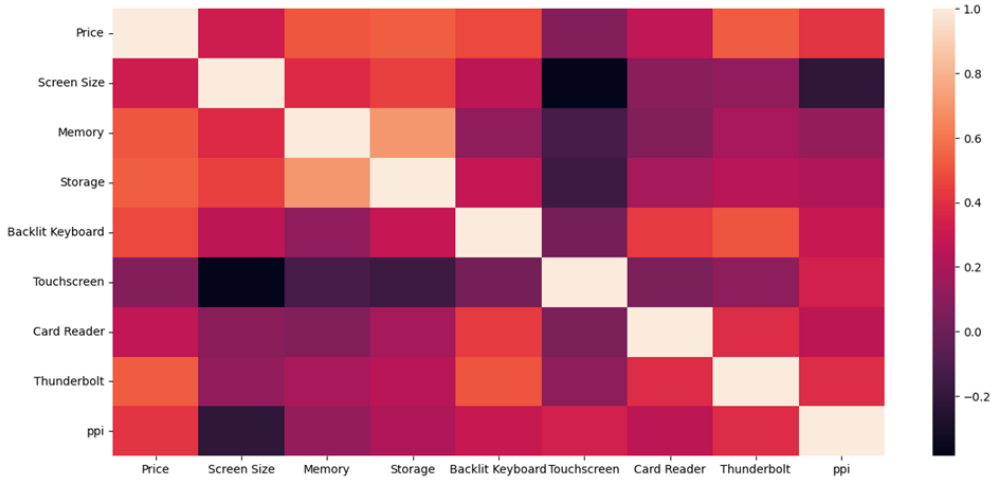
- Kết luận tổng quát:

+ Giá trung bình của laptop là khoảng 1.000 đô la.
+ Laptop có kích thước màn hình lớn hơn, bộ nhớ, có đèn nền, màn hình cảm ứng, đầu đọc thẻ và Thunderbolt nhiều hơn có xu hướng đắt hơn.
+ Tính correlation giữa các biến số và biến mục tiêu "Price"

➔ Dựa vào kết luận tổng quát kết hợp với bảng tương quan từ đó bắt đầu chọn các biến số thích hợp để đưa vào mô hình dự đoán.

Đối với biến phân loại:

Tiến hành phân tích phương sai một chiều (ANOVA [9] - Analysis of Variance) để đánh giá sự chênh lệch trung bình giữa các nhóm của biến phân loại đối với biến phụ thuộc là "Price". Mục tiêu là xác định xem có sự khác biệt đáng kể về giá trị trung bình của "Price" giữa các nhóm của biến phân loại hay không. Phương pháp này kiểm tra tác động của biến phân loại lên biến phụ thuộc, giúp nhóm hiểu rõ hơn về mức độ đồng nhất hoặc chênh lệch giữa các nhóm. Từ đó chọn được nhóm biến phân loại có mức độ đồng nhất với biến "Price" để đưa vào mô hình dự đoán.



Bài tập làm thêm: Phân tích hãng laptop nào có 'memory' và 'operating system' như thế nào có giá trung bình tốt nhất:

Nhóm thực hiện việc gom nhóm các hàng dữ liệu theo 'brand', 'memory', 'operating system' và tính giá trung bình cho từng nhóm, xong sau đó sắp xếp tăng dần, từ đó tìm được nhóm như sau:

Brand	Operating System	Memory	Average Price
hp	***Chrome OS***	***4***	***175.49***

5. KẾT QUẢ PHÂN TÍCH

Qua kết quả phân tích nhóm đã chọn lọc được các biến quan trọng ảnh hưởng tới giá trị dự đoán của biến "Price" để đưa vào mô hình huấn luyện:

- Biến số: Memory, Storage, ppi, Screen Size,
- Biến phân loại: Graphic Type, CPU_series, GPU_brand, Brand, Operating System, Backlit Keyboard, Thunderbolt.

Từ đó sử dụng ba mô hình hồi quy để đánh giá và thu được kết quả sau:

Model	R2	MAE	RMSE
Linear Regression	0.63	290.52	398.94
Support Vector Regression	0.62	266.47	402.59
Random Forest Regression	**0.73**	**231.45**	**338.97**

Thấy rõ mô hình Random Forest Regression (RFR) cho kết quả tốt nhất với R2 Score là 0.81, cao hơn hai mô hình còn lại. Mô hình RFR có thể giải thích được tốt hơn mối quan hệ giữa các biến độc lập và biến phụ thuộc trong bài toán này.

6. KẾT LUẬN

Trong bài báo cáo này, nhóm bắt đầu từ việc thu thập dữ liệu về laptop cũ từ trang Newegg, sau đó tiến hành làm sạch dữ liệu, loại bỏ các giá trị ngoại lai, xử lý các giá trị thiếu và phân tích để xác định các yếu tố ảnh hưởng đến giá. Sử dụng các mô hình hồi quy bao gồm Linear Regression, Support Vector Regression và Random Forest Regression để đánh giá và thu được kết quả.

Kết quả đánh giá cho thấy mô hình hồi quy tuyến tính có thể dự đoán giá laptop với độ chính xác tương đối cao, đặc biệt là mô hình Random Forest Regression với độ đo R2 là 0,73. Tuy nhiên mô hình RFR vẫn còn nhiều hạn chế, lỗi sai dự đoán của mô hình so với thực tế (MAE và RMSE) còn tương đối cao là 231,45 và 338,97.

Trong tương lai, có thể tiếp tục cải thiện độ chính xác của mô hình bằng cách sử dụng các mô hình phức tạp hơn như các mô hình học máy, có thể thu thập thêm các dữ liệu về các yếu tố khác có thể ảnh hưởng đến giá laptop như thời điểm ra mắt, đánh giá của người dùng,…

TÀI LIỆU THAM KHẢO

[1] Newegg. Link: https://www.newegg.com/tools/laptop-finder (Ngày truy cập 25/11/2023)

[2] Chotot. Link: https://www.chotot.com/ (Ngày truy cập 20/11/2023)

[3] Nguyễn Hoàng Phúc, Bùi Quang Phú, Silde về cách crawl data từ trang web. Link Cach_crawl_data, năm 2023

[4] Cellphones. Link https://cellphones.com.vn/ (Ngày truy cập 20/11/2023)

[5] eBay. Link https://www.ebay.com/ (Ngày truy cập 20/11/2023)

[6] Amazon. Link https://www.amazon.com/ (Ngày truy cập 20/11/2023)

[7] PhD. Cuong Sai, Sử dụng thống kê để xác định và loại bỏ dữ liệu ngoại lai cho machine learning trong R và Python, năm 2020

[8] Hồ Chúc, Mật độ điểm ảnh PPI là gì? Công thức tính mật độ điểm ảnh, năm 2023

[9] Áp dụng ANOVA cho nghiên cứu dữ liệu. Link: https://blog.vietnamlab.vn/ap-dung-anova/ (Ngày truy cập 10/12/2023)

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Lương Lý Công Thịnh	Crawl Data, Xử lí dữ liệu, Phân tích dữ liệu, Thiết kế mô hình dự đoán, Làm word
2	Lê Hoàng Thoại	Xử lí dữ liệu, Phân tích dữ liệu, Làm word, Làm ppt
3	Nguyễn Ngọc Tín	Crawl Data, Xử lí dữ liệu, Phân tích dữ liệu, Làm word