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

ĐỀ TÀI: PHÂN TÍCH DỮ LIỆU THIẾT BỊ DI ĐỘNG TRÊN WEBSITE GSM ARENA

Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Phạm Lê Thành Phát	21521262
2	Phạm Huỳnh Thiên Phú	21521278
3	Ngô Gia Lâm	21521054

TP. HỒ CHÍ MINH – 12/2023

1. GIỚI THIỆU

Trong thời đại hiện nay, việc phổ cập thiết bị di động hay điện thoại thông minh hiện nay đang được triển khai rất mạnh. Các sản phẩm điện tử thông minh như điện thoại hay máy tính bảng đang được sản xuất một các cực kì mạnh mẽ, cùng với một số lượng khổng lồ lên đến hàng tỷ chiếc mỗi năm. Với số lượng sản phẩm như vậy thì website công nghệ GSM Arena nổi tiếng ở Mỹ đã và đang thu thập những sản phẩm được ra mắt hằng ngày, như các sản phẩm thiết bị đến từ Apple hay Samsung, máy tính bảng, laptop…

Trong bài báo cáo này, chúng tôi đã thu thập và phân tích bộ dữ liệu GSM Arena Device Spec Dataset, với mục tiêu là phân tích được thông tin của các sản phẩm thông minh đồng thời tìm hiểu được mức độ tăng trưởng của một số hãng có số lượng sản phẩm được bán ra nhiều đồng thời biết được thị hiếu người tiêu dùng thông qua các sản phẩm. Đồng thời dự đoán Bộ dữ liệu phân tích tự thu thập tại website GSM Arena [1] bằng cách crawl thông tin các sản phẩm hiện đang được đưa lên website. Bộ dữ liệu và đề tài do nhóm tự phân tích thiết kế và không dựa trên đề tài nào khác.

2. MÔ TẢ BỘ DỮ LIỆU

2.1. Mô tả bộ dữ liệu:

Bước đầu tiên của quá trình phân tích dữ liệu là xác định vấn đề cần giải quyết. Từ vấn đề được đặt ra, nhóm có thể xác định được nguồn dữ liệu và phương pháp phân tích dữ liệu phù hợp.Vấn đề đặt ra cho đồ án cuối kì lần này là tìm ra những "insight" từ việc phân tích bộ dữ liệu thô đã thu thập được. Để làm được điều này, nhóm cần nắm rõ về những thông tin mà bộ dữ liệu cung cấp.

Bộ dữ liệu nhóm đang thực hiện tập trung vào các sản phẩm thiết bị di động, chủ yếu là điện thoại thông minh, máy tính bảng và laptop. Bộ dữ liệu bao gồm các thông tin như:

+ Thông tin thiết yếu của thiết bị: hãng sản xuất, giá, ngày ra mắt,...

+ Thông số kĩ thuật: độ phân giải màn hình, độ phân giải camera, công nghệ mạng,...

Mục tiêu của đồ án là dự đoán giá cả của các thiết bị, biểu diễn trực quan các số liệu của thiết bị, nhận xét số liệu và tạo dashboard biểu diễn dữ liệu. Từ bộ dữ liệu gốc sau khi thu thập gồm 52 cột. Nhóm đã khảo sát, tách và xử lý dữ liệu và thu được bộ dự liệu còn lại còn 28 cột và 12513 dòng. Trong đó:

+ Name: Mỗi thiết bị sẽ có tên khác nhau, nên do đó có thể dùng cột này làm index để phân biệt các dòng với nhau.

+ MISC_Price_Euro: Là giá của thiết bị đó, chúng tôi chọn đây là biến mục tiêu để dự đoán.

+ 26 biến còn lại sẽ là thông số về thiết bị đó.

Để tránh làm bài báo cáo quá dài, chúng tôi sẽ trình bày mô tả chi tiết cho từng biến trong bộ dữ liệu trước và sau khi xử lí tại Codebook này. Ngoài ra phần giải thích

chi tiết về các biến cần sử dụng chúng tôi sẽ trình bày ở phần trích xuất và tiền xử lí dữ liệu.

2.2. Các bước thu thập dữ liệu:

+ Bước 1: thu thập thông tin các hãng sản xuất thiết bị di động.

+ Bước 2: truy cập vào danh mục sản phẩm của từng hãng, từ đó tạo được danh sách đường dẫn đến các sản phẩm của những hãng đó.

+ Bước 3: truy cập vào từng đường dẫn và thu thập thông tin sản phẩm.

3. PHƯƠNG PHÁP PHÂN TÍCH

3.1. Đặt vấn đề và đề xuất giải pháp

Bước đầu tiên của quá trình phân tích dữ liệu mà nhóm chúng tôi đề xuất đó chính là xác định vấn đề mà từ đó có thể giải quyết thông qua việc khai phá và phân tích dữ liệu. Từ những vấn đề được đề ra, nhóm tôi có thể xác định được đúng nguồn và khối lượng dữ liệu cần thu thập cũng như các phương pháp phân tích dữ liệu phù hợp với đồ án lần này. Vấn đề đặt ra cho đồ án cuối kì lần này của nhóm là làm thế nào để tìm ra được những "insight" từ việc phân tích bộ dữ liệu thô đã thu thập được. Vì vậy nhóm trước tiên cần phải nắm rõ về những thông tin mà bộ dữ liệu cung cấp.

Bộ dữ liệu nhóm đang thực hiện tập trung vào các sản phẩm thiết bị di động, chủ yếu là điện thoại thông minh, máy tính bảng và laptop. Trong đó bao gồm toàn bộ thông tin thiết yếu của thiết bị như hãng sản xuất, giá, ngày ra mắt… và toàn bộ những thông số kĩ thuật như độ phân giải màn hình, độ phân giải camera, công nghệ mạng… với mục tiêu là dự đoán được giá cả của các thiết bị cũng như biểu diễn được trực quan được các số liệu của thiết bị, nhận xét số liệu và tạo được dashboard biểu diễn dữ liệu.

3.2. Trích xuất và tiền xử lí dữ liệu

- **MAIN_CAMERA**: Camera chính

1. Đối với MAIN_CAM_1_Module: thể hiện rằng thiết bị có mô đun camera hay không. Khi kiểm tra thì chúng tôi thấy rằng có hai giá trị là 0 và 1, thể hiện 1 là thiết bị có camera sau, còn 0 là thiết bị không có camera sau.
2. Đối vưới MAIN_CAM_1_Features: là các chức năng của camera. Ở đây có rất nhiều chức năng như LED-flash, HDR… tuy nhiên giá trị null chiếm đến 35% nên nhóm quyết định không sử dụng trường này.
3. Đối với MAIN_CAM_1_Video: cho biết độ phân giải video mà thiết bị có thể quay bằng camera sau. Chúng tôi đã xử lí bằng cách xếp những thiết bị có độ phân giải cao nhất vào một nhóm từ trên cao xuống, vì các thiết bị càng hỗ trợ quay nét thì giá cả càng cao, đồng thời những thiết bị đã hỗ trợ quay ở độ phân giải cao thì đều có thể quay ở độ phân giải thấp hơn. Ví dụ thiết bị có thể quay 8K@30fps, 4K@60fps, 1080p@60fps thì chúng tôi sẽ xếp vào nhóm 8K.

- **MISC_PRICE**: Giá của sản phẩm

1. Thống nhất ngoại tệ: trong cột dữ liệu bao gồm những giá trị như: "**AROUND $giá trị$ `ngoại tệ` ($, £, ₹)"** hoặc những giá trị như "**`ngoại tệ_1` $giá trị_1$**

/ `ngoại tệ_2` $giá trị_2$ / `ngoại tệ_3` $giá trị_3$". Do không thống nhất ngoại tệ, nên chúng tôi sử dụng euro làm ngoại tệ chung cho price, tạo thành những giá trị cho cột MISC_Price_Euro.

2. Chọn giá trị cho những sample mang giá trị **AROUND $giá trị$ `ngoại tệ` ($, £, ₹)**: do giá trị ở đây là around, không rõ ràng nên chúng tôi tiến hành sử dụng hàm random để chọn giá trị cho những sample này. MISC_Price_Euro. = $giá trị$ + round(random.uniform(0, 9), 2).

- **BRAND**: Tên nhà sản xuất của thiết bị

1. Do một số hãng với những yếu tố vô hình có thể tạo nên giá trị riêng cho mình, vì vậy nếu chỉ để tên hãng rồi phân loại thì sẽ không tốt do không có thể biểu hiện mức độ ảnh hưởng tới giá một cách rõ ràng. Vì thể việc encode các giá trị categorical ở Brand đòi hỏi phải sử dụng biến mục tiêu (MISC_Price_Euro) làm chuẩn để tránh vấn đề trên, qua đó ta tiến hành sử dụng Target encode. Ngoài ra One hot encode cũng có thể được sử dụng, nhưng nó sẽ gây ra nhiều chiều cho dữ liệu, nên chúng tôi không sử dụng One hot encode.
2. Target Encode thì lại gây ra vấn đề data leakage và overfitting khi sử dụng trước khi chia thành các tập train hoặc test (hoặc đem dữ liệu huấn luyện ra thực tế). Để tránh vấn đề trên, chúng tôi sử dụng CatBoost encode, đây là phương pháp thường được sử dụng khi muốn sử dụng giá trị của target như làm một chuẩn cho categorical những vẫn tránh được một phần vấn đề data leakage và overfitting.

- **FEATURES_Sensors**: Những cảm biến có trong sản phẩm

1. Tiến hành phân tích những giá trị riêng biệt nằm bên trong giá trị của từng sample, ta thu được 110 sensor khác nhau nằm trong 306 tổ hợp của những giá trị trong thuộc tính này.
2. Tiến hành tính mean giá của sensor mà các sản phẩm có, tạo ra 1 dataframe gồm 120 sensor với giá trị là mean giá của các loại sensor đó.
3. Với mỗi tổ hợp sensor, ta tính giá trị của chúng bằng mean của giá trị từng sensor mà nó sở hữu được lưu trong dataframe ở bước 2.

- **NETWORK**: Các thuộc tính liên quan tới mạng

1. NETWORK_Technology: do thuộc tính năng miêu tả những công nghệ sử dụng trong mạng, những giá trị này cũng được miêu tả bên trong những cột còn lại của NETWORK nên nhóm tiến hành loại bỏ do trùng lặp dữ liệu, nhóm tiến hành xử lý giống cột FEATURES_Sensors được nêu ở trên.
2. NETWORK_G_bands: thuộc tính năng chỉ công nghệ sử dụng và băng tần trong các loại mạng 2G, 3G, 4G, 5G của sản phẩm. Những thuộc tính này cũng bao gồm những categorical quan trọng, rất nhiều giá trị khác nhau, mỗi giá trị tập trung vào một mean giá khác nhau nên chúng tôi sử dụng CatBoost Encode.

3. NETWORK_GPRS: Thuộc tính này chứa những giá trị categorical như yes/no, class hoặc kèm theo chú thích về tốc độ, thông số hoặc liên quan tới công nghệ, nhóm tiến hành drop vì missing data khá nhiều.
4. NETWORK_EDGE: Thuộc tính này cũng tương tự với NETWORK_GPRS.
5. NETWORK_Speed: Thuộc tính này mang những giá trị liên quan tới tốc độ của mạng kèm theo công nghệ sử dụng, thông qua EDA nhóm nhận thấy một vài giá trị phân loại này liên quan tới mean của giá khá nhiều, những công nghệ với tốc độ cao mean sẽ cao hơn tốc độ thấp, nhóm tiến hành sử dụng CatBoost Encode.

- **LAUNCH_Announced**: Thời gian công bố

1. Tiến hành trích xuất ra năm công bố, do năm của lúc bán với lúc công bố thường cùng 1 năm nên nhóm tiến hành lấy năm công bố làm đại diện (do công bố nhiều dữ liệu hơn)
2. Sau khi sử dụng năm, nhóm nhận thấy số lượng thiết bị những năm về sau có xu hướng có giá tăng dần nên tiến hành sử dụng CatBoost Encode.

- **BODY:** Những thuộc tính liên quan tới cấu tạo bên ngoài của sản phẩm

1. BODY_Weight: Tiến hành trích xuất dữ liệu về khối lượng sử dụng đơn vị (g) và sử dụng nó để làm dữ liệu cho thuộc tính này. Những dữ liệu thiếu, nhóm tiến hành điền vào sử dụng mean vì mean cho correlation cao nhất đối với biến mục tiêu.
2. BODY_Dimensions: Thuộc tính này miêu tả kích thước của sản phẩm, nhóm tiến hành sử dụng thuộc tính này để tạo ra 3 thuộc tính mới mô tả các thành phần của kích thước như chiều dài, chiều rộng và độ sâu của sản phẩm "BODY_Length", "BODY_Width", "BODY_Thickness". Một số cột sẽ bị thiếu dữ liệu nên nhóm tiến hành điền vào sử dụng mode vì sau khi kiểm định thì mode cho correlation cao nhất đối với biến mục tiêu.

- **DISPLAY**: Thể hiện mô tả về màn hình của thiết bị

1. DISPLAY_Type: Thể hiện các công nghệ của màn hình thiết bị đó sử dụng, số lượng màu, tần số. Tuy nhiên các mô tả không đồng nhất, nên chúng tôi chỉ trích xuất lấy mô tả công nghệ màn hình ví dụ như: "TFT", "IPS", … Và sau khi trích xuất còn lại 17 giá trị khác nhau.
2. DISPLAY_SIZE: Chiều dài đường chéo của màn hình, đo theo đơn vị inch và cm. Chúng tôi quyết định sử dụng thông số đo bằng inch vì thông số này thường được sử dụng nhiều hơn, bằng cách tách số trước đơn vị inch.
3. DISPLAY_Resolution: Độ phân giải của màn hình. Chúng tôi tách ra thành 2 cột mới là resolution width và resolution height theo độ đo pixel.

- **MEMORY**: Mô tả về bộ nhớ của thiết bị

1. MEMORY_Card_slot: Khe nhớ thẻ nhớ thiết bị, chúng tôi tách theo loại thẻ nhớ gồm microsdhc, microsdxc, microsd, … Gồm 12 giá trị khác nhau.

2. MEMORY_Internal: Thể hiện thông số rom và ram của thiết bị, do đó chúng tôi tách cột đó ra thành 2 cột rom và ram, và do đơn vị đo là GB, MB và KB khác nhau nên để thống nhất, chúng tôi chuyển về đơn vị KB.

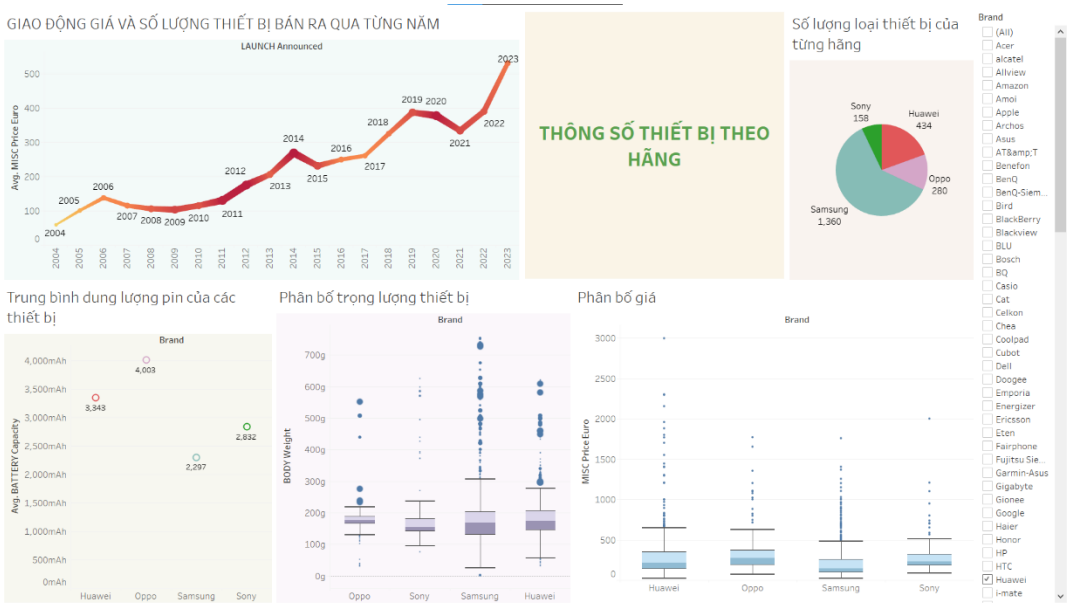
- **BATTERY**: Mô tả về pin của thiết bị

Tuy nhiên các cột khác chỉnh là mô tả về biên nên chúng tôi chỉ sử dụng cột type của battery. Gồm các thông tin về loại pin, khả năng tháo rời và dung lượng pin. Chúng tôi tách cột trên thành type và capacity thể hiện loại pin như Li-Po, Li-ion, … và dung lượng pin. Đối với dụng pin thì có 2 loại số lượng đo mAh và Wh, chúng tôi chọn số liệu với đơn vị mAh và chuyển Wh sang mAh bằng công thức Wh*1000/5 (vì đa số pin dùng loại 5V).

3.3. Phân tích và trực quan dữ liệu

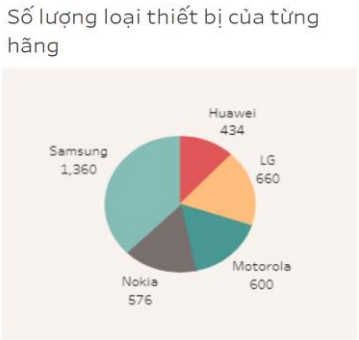
Trực quan: Trực quan thông số thiết bị | Tableau Public

Chúng tôi đã sử dụng các dữ liệu sau khi đã tiền xử lí và xây dựng một dashboard chứa thông tin như sau bằng công cụ Tableau Public [2]:



Hình 1. Giao diện Dashboard

Chi tiết về các hãng khác có thể truy cập tại website trên, ây giờ chúng tôi sẽ tiến hành giải thích các biểu đồ cũng như phân tích dữ liệu của một số nhà sản xuất:



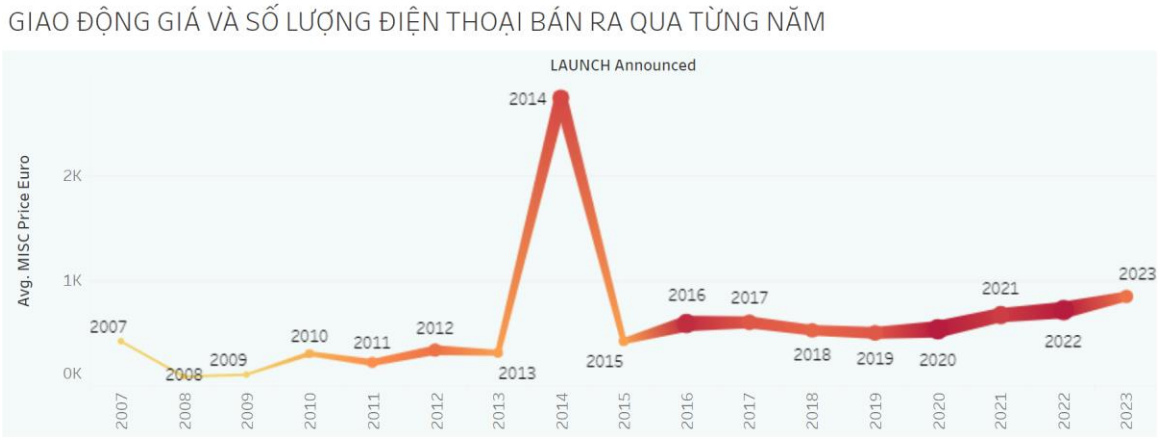
Hình 2. Số lượng thiết bị từng hãng

1. Số lượng thiết bị từng hãng

Trên đây là top 5 nhà sản xuất các thiết bị nhiều nhất, chúng ta có thể tùy chỉnh trên dashboard cũng như cập nhật thêm. Samsung là nhà sản xuất nhiều mẫu thiết bị nhất với 1360 mẫu, tiếp theo đó các hãng còn lại cũng có số lượng mẫu thiết bị gần ngang nhau.

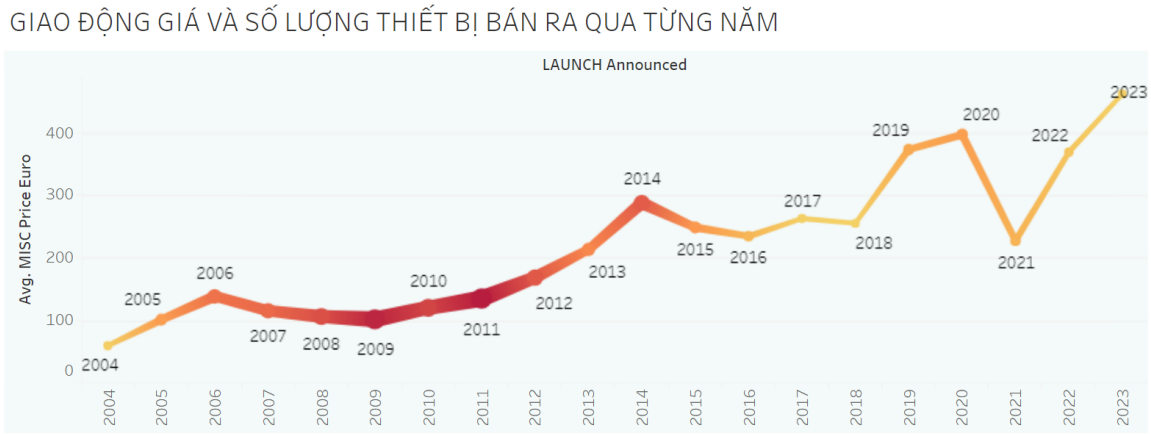
2. Biểu đồ giao động giá và số lượng thiết bị bán ra từng năm:

Cột X thể hiện giá trị trung bình của các sản phẩm di động bán theo từng năm, cột Y thể hiện năm ra mắt của sản phẩm, và độ dày của đường biểu diễn thể hiện số lượng sản phẩm được sản xuất trong những năm đó. Có thể thấy rằng số lượng thiết bị được sản xuất nhiều nhất là vào giai đoạn 2010 – 2012 và 2014, và giá sản phẩm ngày một tăng cao. Duy chỉ có năm 2014 có một số nhà sản xuất có giá thiết bị tăng cao, tiêu biểu là Apple. Nhóm đã tìm hiểu thông tin trên Internet và thấy rằng giá trung bình thiết bị smartphone năm 2014 là 287$ tại trang Business Insider [3], nếu so sánh với biểu đồ của một số hãng khác thì có vẻ như gần trùng khớp, do đó nhóm đã xem lại và phát hiện rằng một số sản phẩm có giá cao bất thường đến từ hãng Apple, cụ thể là một vài dòng sản phẩm Apple Watch phiên bản giới hạn.

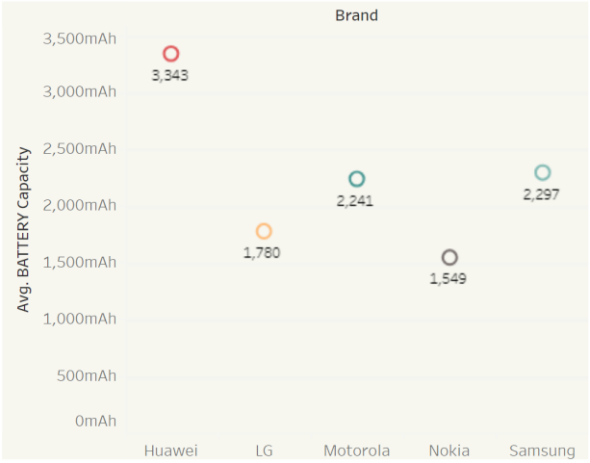


Hình 3. Biểu đồ giao động giá của hãng Apple

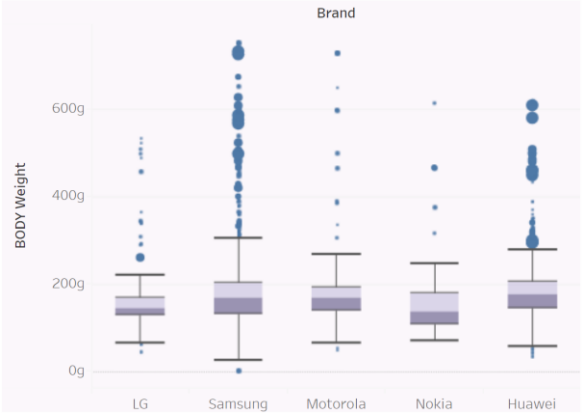
Trong khi đó đối với hãng Samsung, giá của những sản phẩm năm 2014 cũng có tăng 1 chút theo chiều hướng tương tự hãng Apple khi bắt đầu sản xuất những thiết bị "limited edition" với giá cao hơn. Tuy nhiên giá các thiết bị năm 2021 bị giảm một cách bất thường, lí do là vì năm 2021 Samsung bắt đầu đánh mạnh vào thị trường điện thoại giá rẻ, ra mắt rất nhiều dòng sản phẩm có giá tốt để cạnh tranh với các hãng khác, cũng là năm Samsung sản xuất nhiều thiết bị smartphone nhất.



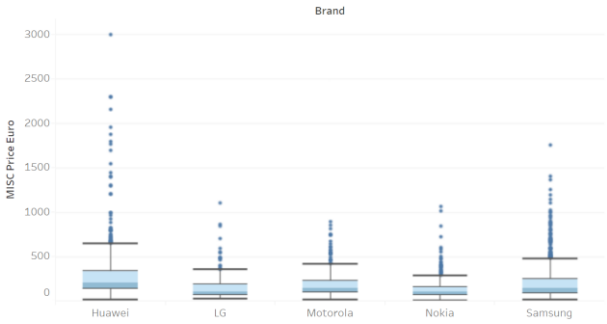
Hình 4. Biểu đồ giao động giá của hãng Samsung



Hình 5. Trung bình dung ượng pin



Hình 6. Phân bố trọng lượng thiết bị



Hình 7. Phân bố giá thiết bị

2. Trung bình dung lượng pin các thiết bị:

Cột X là trung bình dung lượng pin, cột Y là các hãng tiêu biểu có nhiều thiết bị, chúng tôi lấy top 5 các nhà sản xuất có thiết bị được sản xuất nhiều nhất. Cho thấy rằng hãng Huawei là hãng có dung lương pin trung bình cao nhất, trong khi đó Nokia có dung lượng pin trung bình thấp nhất, cũng dễ hiểu vì Nokia là nhà sản xuất lâu năm nên dung lượng pin của các sản phẩm cũ nên pin của chúng có dung lượng thấp hơn các thiết bị hiện nay.

3. Phân bố trọng lượng thiết bị:

Cột X là phân bố trọng lượng của của các sản phẩm, cột Y là top 5 hãng có số lượng thiết bị sản xuất nhiều nhất. Ta có thể thấy rằng trọng lượng trung bình của các sản phẩm của các hãng LG, Samsung, Motorola và Huawei gần tương tự nhau, chỉ có Nokia là thấp hơn một chút, cũng như min thấp hơn, vì sản phẩm Nokia có nhiều sản phẩm "feature phone" có trọng lượng thấp hơn nhiều so với các thiết bị hiện nay.

4. Phân bố giá thiết bị

Cho thấy rằng các thiết bị Huawei có nhiều thiết bị có giá trải dài từ thấp đến rất cao, còn Samsung là hãng có dãy sản phẩm có mật độ phủ giá dày đặc nhất, dễ hiểu vì Samsung là nhà sản xuất nhiều dòng sản phẩm nhất trải dài từ cao cấp đến giá rẻ.

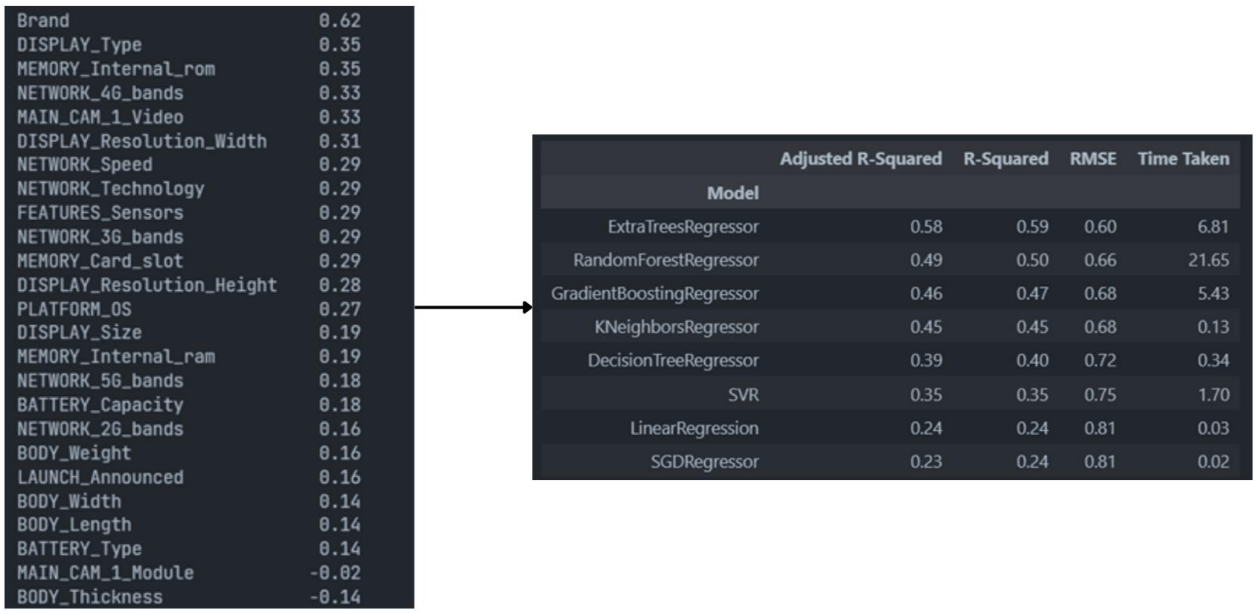
3.4. Lựa chọn mô hình

Đầu tiên, nhóm tiến hành chia tập dữ liệu ra thành Train và Test theo tỉ lệ 7/3, quá trình này sẽ giúp việc đánh giá kết quả của mô hình sau huấn luyện tốt hơn. Tập Train và Test sau khi tiến hành preprocess của nhóm sẽ được sử dụng 25 thuộc tính để dự đoán, với 8438 mẫu dữ liệu. Sau đó sử dụng KNN Imputer để điền khuyết vào các mẫu bị thiếu dữ liệu.

Nhóm tiến hành sử dụng Lazy Predict, một thư viện của python giúp thử nghiệm và kiểm tra nhanh những mô hình học máy của thư viện scikit-learn để tiến hành lựa

chọn ra mô hình có hiệu suất cao nhất. Chúng tôi đã tiến hành chọn những mô hình đại diện với tùy chỉnh như sau:

Đầu tiên chúng tôi sẽ sử dụng các trường dữ liệu dưới đây để tiến hành dự đoán, kết quả được thể hiện ở hình dưới đây:

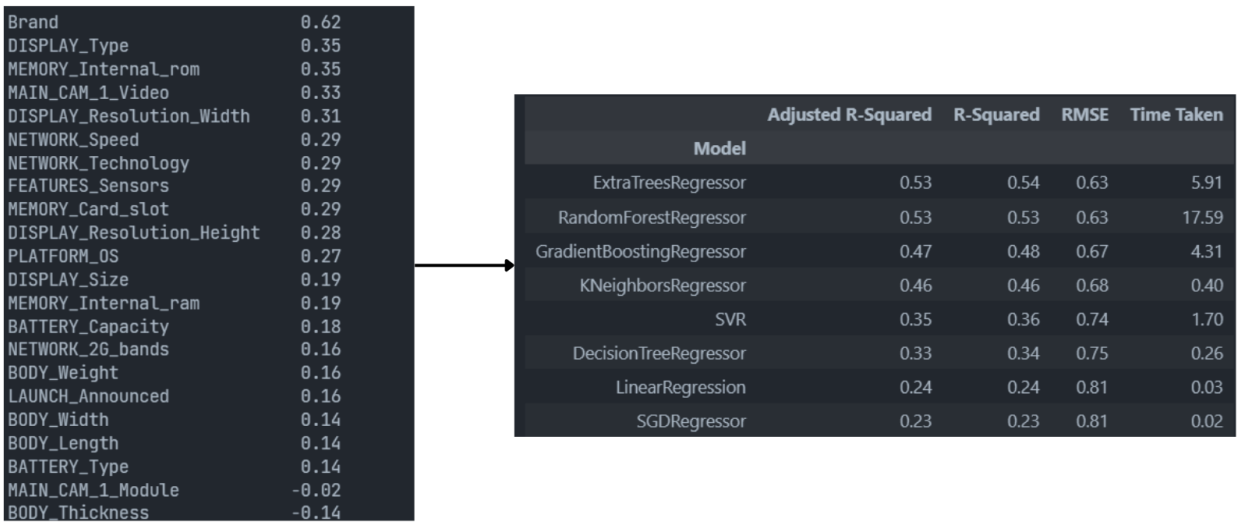


Brand	0.62
DISPLAY_Type	0.35
MEMORY_Internal_rom	0.35
NETWORK_4G_bands	0.33
MAIN_CAM_1_Video	0.33
DISPLAY_Resolution_Width	0.31
NETWORK_Speed	0.29
NETWORK_Technology	0.29
FEATURES_Sensors	0.29
NETWORK_3G_bands	0.29
MEMORY_Card_slot	0.29
DISPLAY_Resolution_Height	0.28
PLATFORM_OS	0.27
DISPLAY_Size	0.19
MEMORY_Internal_ram	0.19
NETWORK_5G_bands	0.18
BATTERY_Capacity	0.18
NETWORK_2G_bands	0.16
BODY_Weight	0.16
LAUNCH_Announced	0.16
BODY_Width	0.14
BODY_Length	0.14
BATTERY_Type	0.14
MAIN_CAM_1_Module	-0.02
BODY_Thickness	-0.14

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
ExtraTreesRegressor	0.58	0.59	0.60	6.81
RandomForestRegressor	0.49	0.50	0.66	21.65
GradientBoostingRegressor	0.46	0.47	0.68	5.43
KNeighborsRegressor	0.45	0.45	0.68	0.13
DecisionTreeRegressor	0.39	0.40	0.72	0.34
SVR	0.35	0.35	0.75	1.70
LinearRegression	0.24	0.24	0.81	0.03
SGDRegressor	0.23	0.24	0.81	0.02

Hình 8. Kết quả sau khi sử dụng Lazy Predict khi không lược bỏ feature.

Sau đó chúng tôi tiến hành dự đoán một lần nữa nhưng bỏ đi một số trường NETWORK, kết quả như sau:



Brand	0.62
DISPLAY_Type	0.35
MEMORY_Internal_rom	0.35
MAIN_CAM_1_Video	0.33
DISPLAY_Resolution_Width	0.31
NETWORK_Speed	0.29
NETWORK_Technology	0.29
FEATURES_Sensors	0.29
MEMORY_Card_slot	0.29
DISPLAY_Resolution_Height	0.28
PLATFORM_OS	0.27
DISPLAY_Size	0.19
MEMORY_Internal_ram	0.19
BATTERY_Capacity	0.18
NETWORK_2G_bands	0.16
BODY_Weight	0.16
LAUNCH_Announced	0.16
BODY_Width	0.14
BODY_Length	0.14
BATTERY_Type	0.14
MAIN_CAM_1_Module	-0.02
BODY_Thickness	-0.14

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
ExtraTreesRegressor	0.53	0.54	0.63	5.91
RandomForestRegressor	0.53	0.53	0.63	17.59
GradientBoostingRegressor	0.47	0.48	0.67	4.31
KNeighborsRegressor	0.46	0.46	0.68	0.40
SVR	0.35	0.36	0.74	1.70
DecisionTreeRegressor	0.33	0.34	0.75	0.26
LinearRegression	0.24	0.24	0.81	0.03
SGDRegressor	0.23	0.23	0.81	0.02

Hình 9. Kết quả sau khi sử dụng Lazy Predict khi lược bỏ bớt feature.

Kết quả cho thấy rằng khi lược bỏ một số feature, hiệu suất của đa số mô hình bị giảm đi, chứng tỏ rằng các feature đều có tương quan lớn đến giá trị được dự đoán. Dựa vào kết quả trên tập test ở trên với những tham số mặc định của những mô hình trên, nhóm tiến hành chọn ExtraTrees Regressor (kết quả cao nhất), KNN (kết quả cao và

thời gian chạy nhanh) và Linear Regression (mô hình tuyến tính) để tiến hành tuning và chạy lại.

4. HUẤN LUYỆN MÔ HÌNH VÀ KẾT QUẢ

Nhóm tiến hành sử dụng Bayesian Optimization để làm cơ sở fine tuning cho các mô hình đã được chọn ở trên với độ đo mục tiêu cần tối ưu là r2_score. Bayesian Optimization là một thuật toán giúp tối ưu hiệu quả những hàm mục tiêu có chi phí evaluation lớn, điều này sẽ giúp việc tăng tốc việc tuning tốt hơn. Ngoài ra, sau đây là kết quả khi đã Fine Tuning của một vài mô hình:

- ExtraTrees: (n_estimators = 80, max_depth = None)
- KNN: (n_neighbors = 3, p =2)
- Linear Regression: ('fit_intercept': False), PolynomialFeatures(degree = 2)

Sau khi nhóm sử dụng mô hình trên để huấn luyện và tiến hành test lại, kết quả thu được cho thấy mô hình có tiến bộ hơn trước khi tuning, cụ thể thì đây là bảng đánh giá kết quả dựa trên 3 độ đo r2_score, MSE, MAE:

	R^2	MSE	MAE
Linear Regression	0.3939	0.5208	0.3098
KNN	0. 4543	0.4688	0.1888
Extra Trees	0.6851	0.2706	0.2043

Bảng 1. Bảng đánh giá kết quả

Kết quả cho thấy, mô hình hoạt động khá tốt đối với những model trên khi đưa vào tập test được chia theo tỉ lệ 7/3 trước đó, với r2 cao nhất thuộc về mô hình ExtraTrees với giá trị là 0.6851. Với số lượng mẫu dữ liệu là hơn 8400 mẫu, chúng tôi nhận định rằng những con số này là hoàn toàn hợp lí và có thể đưa vào dự đoán. Cho thấy việc dự đoán giá các thiết bị thông minh cần rất nhiều features, như Brand là phần ảnh hưởng đến giá các thiết bị khá nhiều ví dụ như Apple vì họ luôn làm các sản phẩm cao cấp dẫn đến trung bình giá sản phẩm của Apple luôn ở mức cao.

5. KẾT LUẬN

Trong bài toán này, chúng tôi đã tiến hành thu thập dữ liệu các thiết bị thông minh tại website GSM Arena, với các thông số của thiết bị và mục tiêu tìm hiểu về các thông số có thể ảnh hưởng đến giá của thiết bị, xử lí được raw data và các biến phân loại, đưa dữ liệu về dạng chuẩn và biểu diễn dữ liệu. Bộ dữ liệu "GSM Arena Device Specs" có chủ yếu là các biến phân loại cũng như lượng missing data lớn do tổng hợp các thiết bị có tuổi đời cao không được công bố thông số, do đó gây khó khăn cho việc dự đoán và xử lí dữ liệu.

Về phần trực quan dữ liệu, chúng tôi đã xử lí các biến gây nhiễu, quy chuẩn thông số các thiết bị sao cho mô hình hoạt động tốt nhất. Từ đó chúng tôi đã biểu diễn được thông tin của các nhà sản xuất thiết bị và tìm hiểu được lí do giá các thiết bị biến đổi

theo năm tháng, tính được mật độ phủ dãy sản phẩm của từng nhà sản xuất, cũng như có được cơ sở để tìm hiểu thong tin thị trường thiết bị di động đã biến đổi như thế nào.

Về huấn luyện mô hình và kiểm thử, nhóm cũng đã thực hiện dự đoán giá thiết bị trên các mô hình học máy bằng cách sử dụng Lazy Predict để tính toán nhanh và lựa chọn những mô hình có hiệu suất tốt nhất là Linear Regression, KNN, Extra Trees. Kết quả thu được là mô hình Extra Trees có khả năng dự đoán tốt nhất đối với bộ dữ liệu.

Trên đây là phần tổng kết đồ án của nhóm, trong tương lai nhóm có định hướng tiếp cận thêm nhiều bộ dữ liệu mới cũng như cách thức xử lí mới cho bộ dữ liệu này. Môn học "Phân tích và trực quan dữ liệu" đã giúp cho nhóm hiểu biết thêm nhiều về các loại dữ liệu, cách thức xử lí dữ liệu thô, biểu diễn dữ liệu và các mô hình dự đoán. Chúng tôi xin cảm ơn Thầy Phạm Thế Sơn đã hướng dẫn và giảng dạy chúng tôi môn học này.

TÀI LIỆU THAM KHẢO

[1] "GSMArena.com - mobile phone reviews, news, specifications and more..." Accessed: Nov. 20, 2023. [Online]. Available: https://www.gsmarena.com/

[2] "Trực quan thông số thiết bị," Tableau Software. Accessed: Dec. 19, 2023. [Online]. Available: Trực quan thông số thiết bị | Tableau Public

[3] T. Danova, "Nearly 70% Of Smartphones Sold In 2014 Cost Less Than $150," Business Insider. Accessed: Dec. 19, 2023. [Online]. Available: https://www.businessinsider.com/nearly-70-of-smartphones-sold-in-2014-cost-less-than-150-2014-12

[4] Shankarpandala, Shankarpandala/lazypredict: Lazy predict help build a lot of basic models without much code and helps understand which models works better without any parameter tuning (Github.com)

[5] Fernando Nogueira. (2014–). Bayesian Optimization: Open source constrained global optimization tool for Python. (Github.com)

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Phạm Lê Thành Phát	- Kiểm tra dữ liệu - Trích xuất dữ liệu - Phân tích kết quả biểu diễn - Viết báo cáo và soạn slide
2	Phạm Huỳnh Thiên Phú	- Thu thập dữ liệu - Mô tả bộ dữ liệu - Trích xuất dữ liệu - Tạo dashboard trực quan dữ liệu
3	Ngô Gia Lâm	- Trích xuất dữ liệu - Tiền xử lý - Kiểm thử mô hình và tuning

PHỤ LỤC HÌNH ẢNH, BẢNG BIỂU