TRƯỜNG ĐẠI HỌC NGOẠI THƯƠNG
CƠ SỞ II TẠI THÀNH PHỐ HỒ CHÍ MINH

BÁO CÁO NHÓM
MÔN: TRÍ TUỆ NHÂN TẠO
TRONG KỶ NGUYÊN CHUYỂN ĐỔI SỐ

DỰ BÁO DOANH SỐ TRANG SỨC TỪ LỊCH SỬ MUA HÀNG BẰNG THƯƠNG MẠI ĐIỆN TỬ CỦA CỬA HÀNG BẰNG MÔ HÌNH LSTM

Giảng viên hướng dẫn: ThS Nguyễn Thị Hoàng Anh

Nhóm sinh viên thực hiện: Nhóm 17

Khóa: 61 Mã lớp: 132

Thành phố Hồ Chí Minh, tháng 6 năm 2024

DANH SÁCH THÀNH VIÊN VÀ MỨC ĐỘ ĐÓNG GÓP

STT	HỌ VÀ TÊN	MSSV	ĐÓNG GÓP
1	Bùi Ngọc Bảo	2215115224	100%
2	Nguyễn Hồ Phương VY	2111113312	100%
3	Nguyễn Đoàn Đức Huy	2214115165	100%

MỤC LỤC

DANH MỤC HÌNH

TÓM TẮT (Đức Huy)

CHƯƠNG I: GIỚI THIỆU ĐỀ TÀI

1.1. Tổng quan về đề tài:

1.1.1. Lý do chọn đề tài:

Đề tài "Dự báo doanh số trang sức từ lịch sử mua hàng thương mại điện tử của cửa hàng" mang tính cấp thiết và có giá trị thực tiễn cao. Trước hết, trong bối cảnh thị trường trang sức ngày càng cạnh tranh, việc dự báo doanh số trở thành công cụ quan trọng giúp các cửa hàng định hình chiến lược kinh doanh hiệu quả. Bằng cách tận dụng dữ liệu lịch sử mua hàng từ các nền tảng thương mại điện tử, các cửa hàng có thể dự đoán xu hướng tiêu dùng, tối ưu hóa quản lý kho và chiến lược marketing, từ đó nâng cao doanh thu và lợi nhuận. Sự áp dụng công nghệ và xử lý dữ liệu lớn (Big Data) trong nghiên cứu này không chỉ giúp nâng cao độ chính xác của dự báo mà còn thể hiện khả năng ứng dụng tiến bộ công nghệ vào kinh doanh. Hơn nữa, việc dự báo chính xác doanh số giúp cửa hàng trang sức ra quyết định kinh doanh kịp thời và chính xác, từ đó tăng cường năng lực cạnh tranh. Đề tài này không chỉ có ý nghĩa thực tiễn mà còn đóng góp vào lĩnh vực nghiên cứu học thuật về dự báo doanh số trong ngành bán lẻ, đặc biệt là ngành trang sức, qua đó mở ra hướng nghiên cứu mới trong lĩnh vực này. Khi cửa hàng trang sức hoạt động hiệu quả hơn, không chỉ mang lại lợi ích kinh tế cho doanh nghiệp mà còn tạo ra nhiều cơ hội việc làm, góp phần vào sự phát triển kinh tế xã hội của địa phương.

1.1.2. Mục tiêu của đề tài:

Thu thập và phân tích dữ liệu lịch sử mua hàng để hiểu rõ xu hướng tiêu dùng, xây dựng mô hình dự báo doanh số sử dụng trí tuệ nhân tạo và xử lý dữ liệu lớn, ứng dụng mô hình này vào tối ưu hóa quản lý kho hàng, chiến lược marketing và lập kế hoạch kinh doanh nhằm cải thiện hiệu quả hoạt động và tăng trưởng doanh thu. Đồng thời, đề tài sẽ đánh giá hiệu quả của mô hình dự báo so với các phương pháp truyền thống và đề xuất các giải pháp cải tiến nhằm nâng cao độ chính xác và hiệu quả kinh doanh của cửa hàng trang sức, qua đó đóng góp vào nghiên cứu học thuật trong lĩnh vực dự báo doanh số bán lẻ.

1.1.3. Đóng góp của đề tài:

Đề tài "Dự báo doanh số trang sức từ lịch sử mua hàng thương mại điện tử của cửa hàng" đóng góp vào việc tối ưu hóa chiến lược kinh doanh và quản lý kho hàng, giúp cửa hàng trang sức cải thiện hiệu quả hoạt động và tăng trưởng doanh thu. Nó cung cấp mô hình dự báo chính xác hơn so với phương pháp truyền thống, giúp ra quyết định kinh doanh kịp thời và chính xác. Đồng thời, đề tài góp phần vào nghiên cứu học thuật về dự báo doanh số trong ngành bán lẻ, mở ra hướng nghiên cứu mới và cung cấp giải pháp ứng dụng thực tiễn trong kinh doanh trang sức.

1.2. Giới thiệu bộ dữ liệu:

1.2.1. Nguồn gốc, mục đích của bộ dữ liệu:

Bộ dữ liệu "E-commerce Purchase History from Jewelry Store" được chia sẻ trên nền tảng Kaggle bởi tác giả Maksim Kechinov, bộ dữ liệu này chứa dữ liệu mua hàng từ

tháng 12 năm 2018 đến tháng 12 năm 2021 từ một cửa hàng trực tuyến trang sức cỡ trung bình, được thu thập bởi dự án Open CDP.

Kaggle là một nền tảng trực tuyến hàng đầu dành cho các nhà khoa học dữ liệu, nhà phân tích và nhà phát triển học máy có thể chia sẻ các bộ dữ liệu, thảo luận về các phương pháp phân tích tiên tiến và tham gia vào các cuộc thi thách thức kỹ năng phân tích dữ liệu và học máy. Nền tảng này thúc đẩy sự hợp tác và học hỏi liên tục thông qua các bộ dữ liệu thực tế từ nhiều ngành công nghiệp khác nhau, tạo điều kiện cho các nhà nghiên cứu và nhà phát triển khám phá, phân tích và cải thiện các kỹ thuật và mô hình của mình.

Bộ dữ liệu "E-commerce Purchase History from Jewelry Store" được cung cấp công khai để hỗ trợ cộng đồng khoa học dữ liệu trong việc nghiên cứu và phát triển các mô hình phân tích dữ liệu và học máy, học sâu liên quan đến thương mại điện tử, đặc biệt trong ngành bán lẻ trang sức. Bộ dữ liệu này chứa thông tin chi tiết về lịch sử mua hàng của khách hàng tại một cửa hàng trang sức trực tuyến, bao gồm tổng cộng 13 cột và 22.844 hàng. Các cột dữ liệu quan trọng bao gồm ID của khách hàng, thời gian mua hàng, loại sản phẩm, giá trị đơn hàng, số lượng sản phẩm mua, phương thức thanh toán và thông tin giảm giá.

Mục tiêu của bộ dữ liệu này là cung cấp một cơ sở dữ liệu phong phú cho việc phân tích hành vi mua sắm trực tuyến của khách hàng, nhằm giúp các nhà bán lẻ hiểu rõ hơn về xu hướng tiêu dùng, mô hình mua sắm và các yếu tố ảnh hưởng đến quyết định mua hàng trong lĩnh vực trang sức. Việc phân tích bộ dữ liệu có thể giúp các nhà bán lẻ trang sức trực tuyến nhận diện các yếu tố thúc đẩy khách hàng mua hàng, đánh giá hiệu quả của các chương trình khuyến mãi, và tối ưu hóa chiến lược giá cả để cải thiện doanh số bán hàng.

Bằng cách phân tích chi tiết các giao dịch mua bán, nhóm nghiên cứu có thể phát hiện ra các xu hướng tiêu dùng, như loại trang sức phổ biến theo mùa hoặc các dịp đặc biệt, thói quen chi tiêu của các nhóm khách hàng khác nhau, và tác động của các chương trình khuyến mãi hoặc giảm giá đối với hành vi mua sắm. Những thông tin này có thể được sử dụng để cá nhân hóa trải nghiệm mua sắm, xây dựng các chương trình khách hàng thân thiết hiệu quả hơn, và tối ưu hóa các chiến dịch tiếp thị.

Hơn nữa, bộ dữ liệu này còn mở ra nhiều cơ hội để phát triển các mô hình học máy tiên tiến nhằm dự đoán hành vi mua sắm trong tương lai và đề xuất các sản phẩm phù hợp với từng khách hàng. Các kỹ thuật như phân tích giỏ hàng (market basket analysis), phân cụm khách hàng (customer segmentation), và các hệ thống gợi ý (recommendation systems) có thể được áp dụng để mang lại những trải nghiệm mua sắm cá nhân hóa hơn và tăng cường sự hài lòng của khách hàng.

1.2.2. Các thông tin bộ dữ liệu cung cấp:

Bộ dữ liệu "E-commerce Purchase History from Jewelry Store" cung cấp một cái nhìn chi tiết về lịch sử mua sắm trang sức trực tuyến của khách hàng tại một cửa hàng trang sức trong 3 năm. Dưới đây là các thông tin cụ thể mà bộ dữ liệu này cung cấp cho nhóm nghiên cứu:

- Thông tin về đơn hàng:

+ *Order ID:* Mã số đơn hàng duy nhất, đại diện cho mỗi giao dịch mua hàng. Giúp dễ dàng theo dõi lịch sử mua sắm, kiểm tra trạng thái đơn hàng, và quản lý các giao dịch.
+ *Order Datetime:* Ngày đặt hàng, cho biết thời điểm mà khách hàng đã đặt mua sản phẩm. Thông tin này cho biết chính xác thời điểm mà khách hàng đã thực hiện mua sản phẩm, bao gồm cả ngày và giờ cụ thể.
+ *Purchase product ID:* Mã số sản phẩm duy nhất cho từng mặt hàng trang sức trong đơn hàng. Giúp phân tích chi tiết về loại sản phẩm được mua, theo dõi doanh số của từng sản phẩm cụ thể, và quản lý hàng tồn kho hiệu quả
+ *Quantity of SKU in the order:* Số lượng sản phẩm đã mua trong mỗi giao dịch. SKU (Stock Keeping Unit) là đơn vị lưu kho cho sản phẩm, và quantity đề cập đến số lượng của SKU đó được bao gồm trong đơn hàng. Thông tin về số lượng giúp đánh giá nhu cầu sản phẩm, xác định các mặt hàng bán chạy, và tối ưu hóa quy trình sản xuất cũng như cung ứng.

- Thông tin về sản phẩm:

+ *Category ID:* Danh mục sản phẩm, cung cấp thông tin về loại sản phẩm như nhẫn, vòng cổ, hoa tai,...
+ *Category Alias:* Tên rút gọn hoặc viết tắt của các danh mục sản phẩm. Đây là một dạng nhãn (label) hoặc tên thay thế được sử dụng để biểu diễn một category trong bộ dữ liệu.
+ *Price in USD:* Giá của từng sản phẩm, tính theo đồng đô la Mỹ (USD), đại diện cho giá bán lẻ của mỗi mặt hàng tại thời điểm giao dịch.
+ *Color:* Thuộc tính mô tả màu sắc của sản phẩm trang sức.
+ *Metal types per Brand ID:* Các loại kim loại mà mỗi thương hiệu sử dụng cho sản phẩm của mình như vàng, bạc, bạch kim, hoặc các hợp kim khác mà thương hiệu đó sử dụng trong sản xuất trang sức.
+ *Gem:* Thuộc tính mô tả loại đá quý được gắn trong sản phẩm trang sức. Thường được phân loại dựa trên loại đá như kim cương, ngọc bích, ngọc trai, hồng ngọc, ngọc lục bảo, và các loại đá quý khác.

- Thông tin về khách hàng:

+ *Used ID:* Mã số nhận dạng khách hàng duy nhất, sử dụng để xác định danh tính của người mua trong hệ thống. Mỗi khách hàng có một Used ID riêng, giúp phân biệt họ với những khách hàng khác và cho phép theo dõi các giao dịch của từng người.

+ *Gender:* Chỉ định giới tính của khách hàng, thường được ghi nhận bằng các giá trị như "Nam", "Nữ", hoặc các nhãn khác tùy thuộc vào dữ liệu cụ thể. Thông tin này hỗ trợ việc phân tích và hiểu sâu hơn về sự khác biệt trong hành vi mua sắm giữa các nhóm giới tính.

Bộ dữ liệu giúp nhóm nghiên cứu phân tích hành vi mua sắm, xác định xu hướng tiêu dùng, và tối ưu hóa chiến lược kinh doanh như quản lý hàng tồn kho và phân tích giá trị đơn hàng. Thông qua việc phân tích dữ liệu này, nhóm nghiên cứu có thể hiểu rõ hơn về sở thích khách hàng và đưa ra các giải pháp cải thiện hiệu suất kinh doanh.

CHƯƠNG II: PHÂN TÍCH BỘ DỮ LIỆU

2.1. Kỹ thuật làm sạch dữ liệu:

Trong quá trình xử lý và phân tích dữ liệu, nhóm tác giả đã thực hiện một số bước làm sạch dữ liệu nhằm đảm bảo dữ liệu sử dụng cho mô hình phân tích và dự đoán là chất lượng và đáng tin cậy. Dưới đây là các bước cụ thể đã được thực hiện:

- Bước 1: Kiểm tra thông tin dữ liệu ban đầu:

+ Mục đích: Xác định cấu trúc và đặc điểm của dữ liệu, bao gồm số lượng bản ghi và cột, cùng với thông tin về các giá trị thiếu.
+ Kết quả: Dữ liệu bao gồm 95,911 bản ghi và 13 cột, với một số cột chứa các giá trị thiếu.

- Bước 2: Kiểm tra giá trị thiếu:

+ Mục đích: Xác định số lượng giá trị thiếu trong từng cột.
+ Kết quả: Một số cột có giá trị thiếu đáng kể như Category Code, BrandID, USD Price, UserID, Gender, Color, Metal, và Gem.

- Bước 3: Xử lý giá trị thiếu cho cột giới tính (Gender):

+ Mục đích: Thay thế các giá trị thiếu trong cột Gender bằng giá trị phổ biến nhất.
+ Cách thực hiện: Dựa trên phân tích cho thấy phần lớn khách hàng là nữ (f), các giá trị thiếu sẽ được thay thế bằng f.
+ Kết quả: Giá trị thiếu trong cột Gender được thay thế thành công.

- Bước 4: Loại bỏ cột không cần thiết (Quantity)

+ Mục đích: Loại bỏ cột Quantity vì nó chỉ chứa một giá trị duy nhất, không cung cấp thông tin hữu ích.
+ Cách thực hiện: Dùng lệnh df.drop để loại bỏ cột này.
+ Kết quả: Cột Quantity đã bị loại bỏ.

- Bước 5: Xử lý giá trị thiếu cho USD Price và UserID

+ Mục đích: Loại bỏ các hàng có giá trị thiếu trong USD Price và UserID.
+ Cách thực hiện: Lọc các hàng có giá trị USD Price và UserID không bị thiếu.
+ Kết quả: Các hàng có giá trị thiếu trong USD Price và UserID đã bị loại bỏ.

- Bước 6: Xử lý giá trị thiếu cho Category Code:

+ Mục đích: Thay thế các giá trị thiếu trong Category Code bằng other.
+ Cách thực hiện: Sử dụng new_df['CategoryCode'].fillna('other', inplace=True) để thay thế các giá trị thiếu.
+ Kết quả: Giá trị thiếu trong Category Code được thay thế bằng other.

- Bước 7: Xử lý giá trị thiếu cho BrandID:

+ Mục đích: Thay thế các giá trị thiếu trong BrandID bằng -1.

+ Cách thực hiện: Sử dụng new_df['BrandID'].fillna('-1', inplace=True) để thay thế các giá trị thiếu.
+ Kết quả: Giá trị thiếu trong BrandID được thay thế bằng -1.

- Bước 8: Xử lý giá trị thiếu cho Gem, Metal, và Color:

+ Mục đích: Thay thế các giá trị thiếu trong các cột Gem, Metal, và Color bằng unknown.
+ Cách thực hiện: Dùng vòng lặp để thay thế các giá trị thiếu bằng unknown.
+ Kết quả: Các giá trị thiếu trong Gem, Metal, và Color được thay thế thành công.

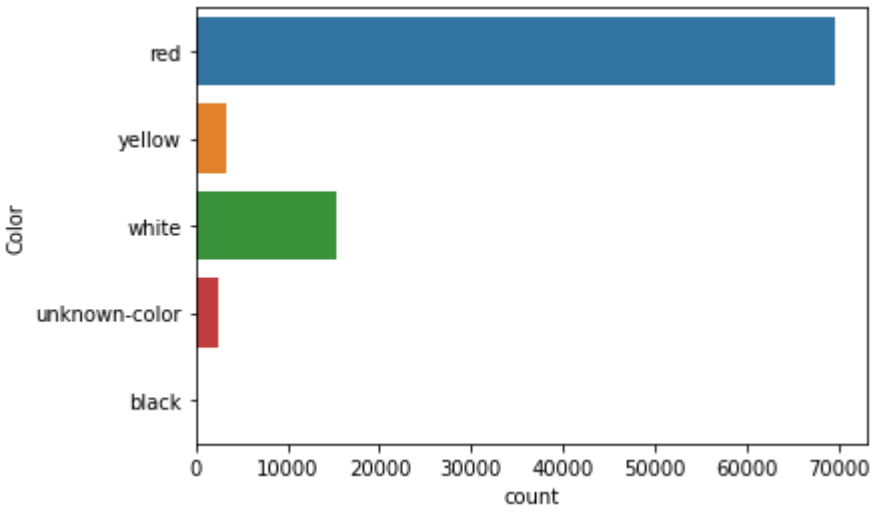
- Bước 9: Kiểm tra lại thông tin dữ liệu sau làm sạch:

+ Mục đích: Xác nhận rằng tất cả các bước làm sạch dữ liệu đã được thực hiện thành công.
+ Cách thực hiện: Sử dụng lại lệnh new_df.info() để kiểm tra thông tin dữ liệu.
+ Kết quả: Dữ liệu đã được làm sạch với 95,959 bản ghi và 12 cột, không còn giá trị thiếu.

Quy trình làm sạch dữ liệu của nhóm tác giả bao gồm kiểm tra và xử lý các giá trị thiếu, loại bỏ các cột không cần thiết, và thay thế các giá trị thiếu bằng cách điền giá trị phổ biến hoặc một giá trị đại diện. Việc làm sạch dữ liệu giúp chuẩn bị dữ liệu cho các bước phân tích và mô hình hóa tiếp theo, đảm bảo rằng dữ liệu là nhất quán và không có lỗi.

2.2. Trực quan dữ liệu:

2.2.1. Màu sắc trang sức phổ biến:

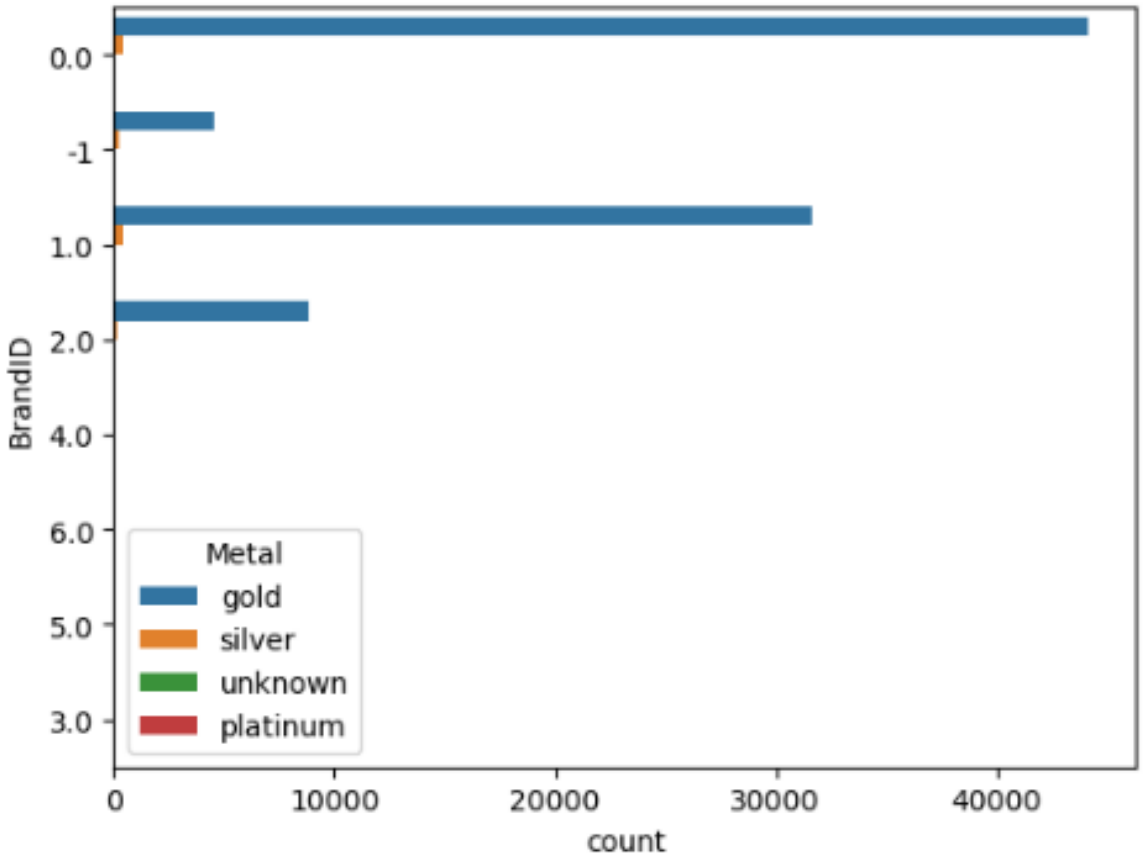


Hình 2.2.1. Màu sắc trang sức phổ biến

Màu đỏ là màu sắc phổ biến nhất, với số lượng vượt trội so với các màu khác, đạt tới gần 70,000 đơn vị. Màu trắng là màu sắc phổ biến thứ hai với khoảng 20,000 đơn vị, tiếp theo là màu vàng với số lượng khoảng 10,000 đơn vị. Các màu sắc khác như đen và

màu không xác định có số lượng ít hơn đáng kể, dưới 5,000 đơn vị. Kết quả này cho thấy sự ưa chuộng mạnh mẽ của khách hàng đối với các trang sức màu đỏ, trong khi các màu sắc khác có mức độ phổ biến thấp hơn

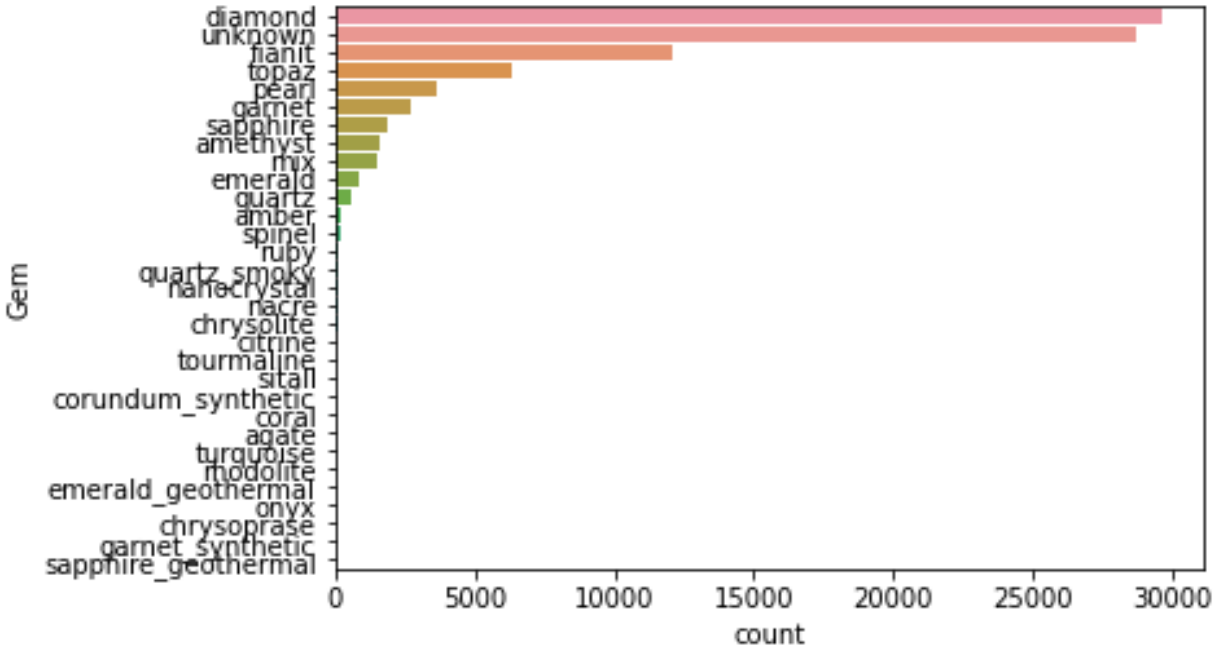
2.2.2. Các loại kim loại trên mỗi ID nhãn hiệu (Brand ID):



Hình 2.2.2. Kim loại phổ biến trên mỗi ID nhãn hiệu (Brand ID)

Vàng (gold) là kim loại phổ biến nhất, với số lượng lớn nhất được phân bố chủ yếu ở Brand ID 1 và Brand ID 2, lần lượt là khoảng 35,000 và 20,000 đơn vị. Các kim loại khác như bạc (silver) và bạch kim (platinum) có số lượng rất nhỏ và xuất hiện không đáng kể so với vàng.

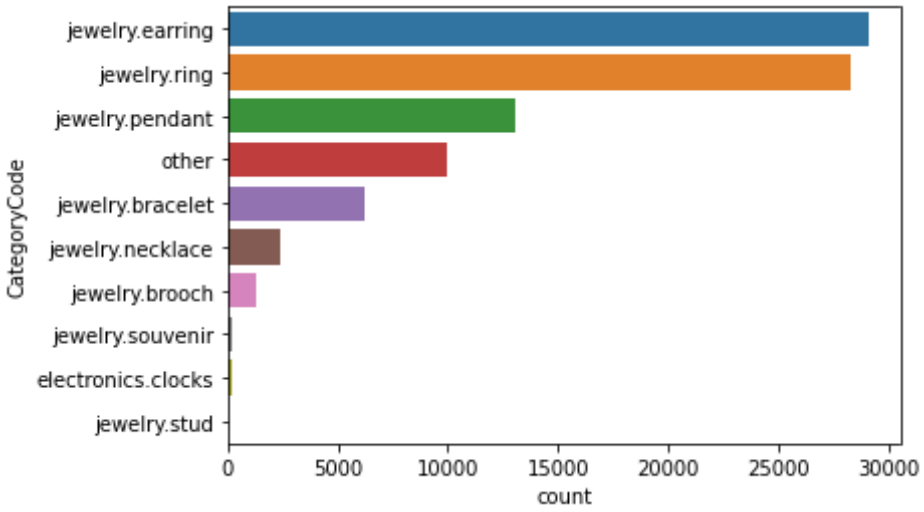
2.2.3. Loại đá trang sức phổ biến:



Hình 2.2.3. Loại đá trang sức phổ biến

Kim cương (diamond) là loại đá phổ biến nhất, với số lượng khoảng 30,000 đơn vị. Loại đá không xác định (unknown) cũng có số lượng tương đối lớn, khoảng hơn 25,000 đơn vị. Tiếp theo là topaz và garnet với số lượng lần lượt khoảng 10,000 và 7,000 đơn vị. Các loại đá khác như amethyst, sapphire, emerald, và quartz có số lượng thấp hơn, dao động từ 2,000 đến 5,000 đơn vị. Các loại đá quý còn lại như ruby, opal, và jade có số lượng ít hơn nhiều, dưới 2,000 đơn vị. Kết quả này cho thấy kim cương là loại đá được ưa chuộng nhất trong trang sức của cửa hàng, trong khi các loại đá khác như topaz và garnet cũng có sự phổ biến đáng kể

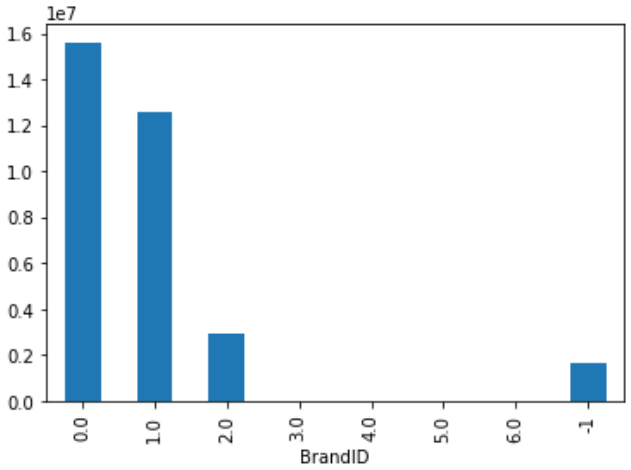
2.2.4. Phân bố Mã danh mục (Category Code):



Hình 2.2.4. Biểu đồ Phân bố Mã danh mục (Category Code)

Sản phẩm "jewelry.earring" chiếm tỉ lệ lớn nhất với hơn 30,000 lượt mua, tiếp theo là "jewelry.ring" với khoảng hơn 25,000 lượt. Các danh mục khác như "jewelry.pendant," "other," và "jewelry.bracelet" cũng có số lượng đáng kể, trong khi các danh mục còn lại có số lượng ít hơn. Dựa trên dữ liệu này, chiến lược dự báo doanh số trang sức nên tập trung vào các sản phẩm thuộc danh mục "jewelry.earring" và "jewelry.ring" do tần suất mua hàng cao, nhằm tối ưu hóa dự báo doanh thu cho các sản phẩm chủ lực này.

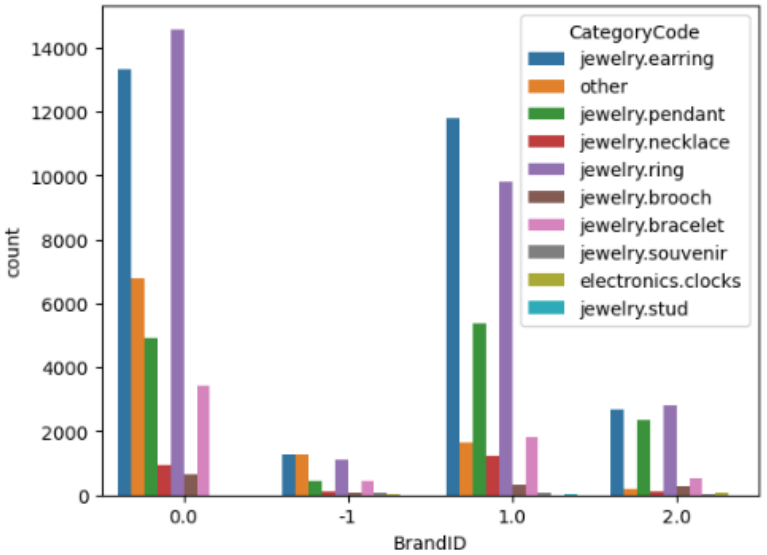
2.2.5. Phân bố ID nhãn hiệu (Brand ID):



Hình 2.2.4. Biểu đồ phân bố ID nhãn hiệu (Brand ID)

Biểu đồ phân bố ID nhãn hiệu (Brand ID) cho thấy rằng hai nhãn hiệu có mã 0 và 1 chiếm phần lớn doanh số, với hơn 1.4 triệu và 1.2 triệu lượt bán tương ứng. Nhãn hiệu có mã 2 chiếm một phần nhỏ hơn đáng kể, trong khi các nhãn hiệu khác gần như không đáng kể.

2.2.6. Phân bố của các loại trang sức phổ biến theo mã danh mục (CategoryCode) và ID nhãn hiệu (BrandID):



Hình 2.2.6. Biểu đồ phân bố của các loại trang sức phổ biến theo mã danh mục (CategoryCode) và ID nhãn hiệu (BrandID)

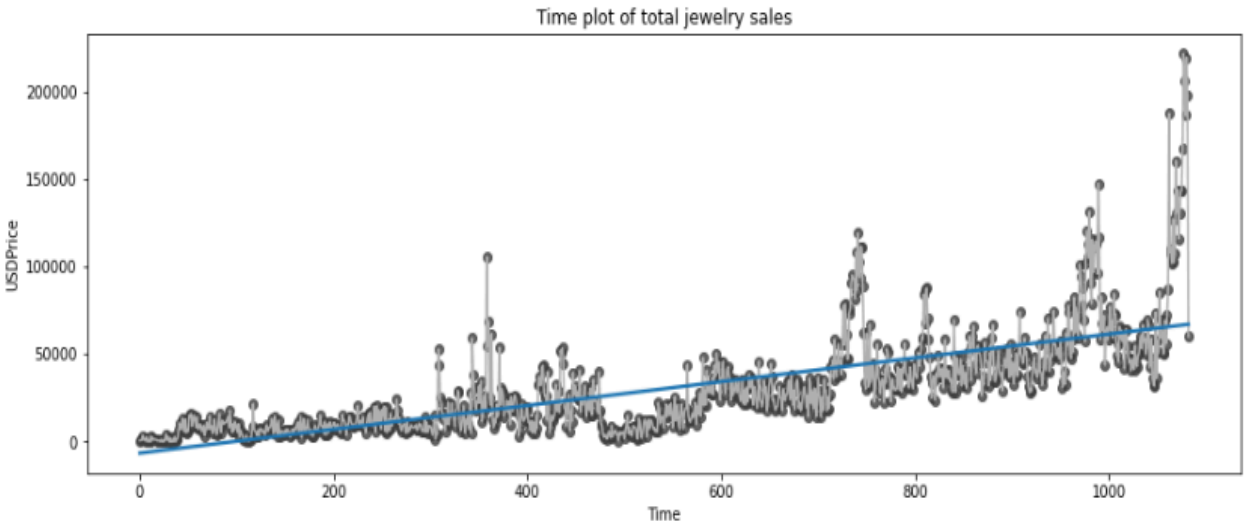
Đối với BrandID 0, hoa tai (jewelry.earring) và nhẫn (jewelry.ring) là hai loại trang sức phổ biến nhất, với số lượng lần lượt khoảng 14,000 và 12,000 đơn vị. Các loại khác như dây chuyền (jewelry.necklace), mặt dây chuyền (jewelry.pendant), và vòng tay (jewelry.bracelet) cũng có số lượng đáng kể, dao động từ 4,000 đến 8,000 đơn vị.

Đối với BrandID 1, hoa tai và nhẫn vẫn chiếm ưu thế, với số lượng khoảng 10,000 và 8,000 đơn vị, trong khi mặt dây chuyền và dây chuyền tiếp tục giữ vị trí quan trọng với số lượng khoảng 5,000 đến 7,000 đơn vị.

Đối với BrandID 2, số lượng các loại trang sức giảm đáng kể, nhưng hoa tai và nhẫn vẫn là các sản phẩm phổ biến nhất, với số lượng lần lượt khoảng 6,000 và 4,000 đơn vị.

Kết quả này cho thấy hoa tai và nhẫn là những loại trang sức được ưa chuộng nhất ở các nhãn hiệu, trong khi các loại khác như mặt dây chuyền, dây chuyền, và vòng tay cũng có sự phổ biến đáng kể.

2.2.7. Biểu đồ thời gian của tổng doanh số bán đồ trang sức:



Hình 2.2.6 Biểu đồ thời gian của tổng doanh số bán đồ trang sức:

Biểu đồ thời gian của tổng doanh số bán đồ trang sức cho thấy xu hướng tăng trưởng theo thời gian, với một số dao động mạnh mẽ trong các khoảng thời gian ngắn. Đường xu hướng tăng dần, chỉ ra rằng doanh số bán hàng có xu hướng tăng trưởng đều đặn. Các đỉnh đột biến trong biểu đồ có thể phản ánh các sự kiện bán hàng đặc biệt hoặc chiến dịch quảng cáo hiệu quả. Nhìn chung, doanh số bán hàng không chỉ tăng lên mà còn bị ảnh hưởng bởi các yếu tố ngắn hạn

2.4. Mô hình học sâu: (Ngọc Bảo)

2.5.1. Xác định vấn đề:

2.5.1.1. Bối cảnh:

Dự đoán doanh thu bán hàng là một phần cốt lõi của việc quản lý kinh doanh, cho phép các nhà bán lẻ tiên liệu trước nhu cầu của thị trường. Đặc biệt trong ngành trang

sức, nơi mà thị hiếu và xu hướng có thể thay đổi nhanh chóng, việc dự đoán chính xác không chỉ giúp các nhà bán lẻ hiểu rõ hơn về nhu cầu khách hàng mà còn tạo điều kiện để tối ưu hóa quy trình cung ứng sản phẩm.

Việc dự đoán doanh thu chính xác giúp giảm thiểu rủi ro liên quan đến hàng tồn kho. Sản phẩm trang sức, thường có giá trị cao và đòi hỏi chi phí lưu kho lớn, cần được quản lý chặt chẽ để tránh tình trạng tồn đọng hoặc thiếu hụt. Một chiến lược dự đoán hiệu quả giúp cân bằng giữa cung và cầu, đảm bảo rằng hàng hóa luôn sẵn sàng khi khách hàng cần, đồng thời giảm thiểu lãng phí do hàng tồn kho quá mức.

Dự đoán doanh thu còn hỗ trợ việc xây dựng các chiến lược marketing hiệu quả. Hiểu rõ về các xu hướng mua sắm giúp các nhà bán lẻ xác định được thời điểm và cách thức tiếp cận khách hàng mục tiêu, từ đó tối ưu hóa chi phí và hiệu quả của các chiến dịch tiếp thị.

Và một trong những phương pháp dự đoán phổ biến là phân tích dữ liệu lịch sử bán hàng. Việc này bao gồm việc xem xét các xu hướng mua sắm, doanh thu theo mùa, và các dịp lễ tết để đưa ra dự đoán cho tương lai. Dữ liệu lịch sử có thể cung cấp những thông tin quý giá về hành vi tiêu dùng và giúp xác định các mô hình mua sắm theo thời gian.

Một cách tiếp cận được nhóm tác giả đánh giá hiệu quả là sử dụng các mô hình dự đoán kết hợp, trong đó bao gồm:

+ *Phân tích dữ liệu lịch sử:* Xem xét các xu hướng doanh thu theo mùa và các dịp đặc biệt dựa trên dữ liệu bán hàng trong quá khứ.
+ *Học sâu và Trí tuệ nhân tạo:* Sử dụng thuật toán học sâu để phân tích dữ liệu phức tạp, phát hiện các mẫu không rõ ràng mà các phương pháp truyền thống có thể bỏ sót.

Sự kết hợp này giúp nhóm tác giả tối ưu hóa độ chính xác của dự đoán, từ đó hỗ trợ các nhà bán lẻ trong việc lập kế hoạch chiến lược kinh doanh và cải thiện quản lý hàng tồn kho.

2.5.1.2. Lý do chọn Mô hình LSTM:

LSTM, một biến thể của mạng neural hồi quy (RNN), được thiết kế để giải quyết các vấn đề liên quan đến dữ liệu chuỗi thời gian. Điều này là nhờ khả năng của LSTM trong việc lưu giữ thông tin dài hạn, giúp mô hình nhận diện và học hỏi từ các mẫu và xu hướng trong dữ liệu lịch sử.

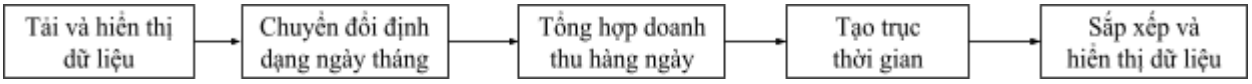
So với các mạng neural truyền thống, LSTM giúp giảm thiểu vấn đề vanishing gradient (tiêu biến độ dốc), một vấn đề phổ biến khi học các chuỗi thời gian dài. Điều này làm cho LSTM trở nên lý tưởng để học từ các chuỗi thời gian dài hạn, một yếu tố quan trọng trong việc dự đoán doanh thu, nơi các mẫu dữ liệu có thể kéo dài qua nhiều tháng hoặc năm.

Nhiều nghiên cứu và ứng dụng trong dự đoán chuỗi thời gian, bao gồm dự báo doanh số bán lẻ và phân tích tài chính, đã chứng minh LSTM có hiệu suất cao trong việc dự đoán chính xác xu hướng tương lai.

2.5.2. Quy trình thực hiện:

Quy trình thực hiện mô hình học sâu của nhóm tác giả gồm có 3 bước, cụ thể:

- Bước 1: Tiền xử lý dữ liệu:



- Bước 2: Chuyển đổi thành bài toán chuỗi thời gian: Dữ liệu được chia thành tập huấn luyện, kiểm tra và xác thực để đảm bảo mô hình được đánh giá toàn diện.

- Bước 3: Huấn luyện mô hình LSTM:

+ Xây dựng và cấu hình mô hình LSTM: Mô hình LSTM với 165 đơn vị trong lớp LSTM, kết hợp với các lớp dropout để giảm thiểu overfitting và các lớp dense để dự đoán đầu ra.
+ Huấn luyện mô hình: Mô hình được huấn luyện với dữ liệu huấn luyện và xác thực, sử dụng early stopping để ngăn chặn overfitting bằng cách dừng huấn luyện khi loss không giảm sau một số epoch nhất định.

- Bước 4: Đánh giá và dự đoán:

+ Đánh giá mô hình trên tập kiểm tra: Mô hình được đánh giá trên tập kiểm tra với RMSE (Root Mean Square Error) để đo lường độ chính xác.
+ Biểu diễn kết quả dự đoán: So sánh giữa giá trị doanh thu thực tế và giá trị dự đoán để trực quan hóa hiệu suất của mô hình.

2.5.3. Kết quả và phân tích:

2.5.3.1. Cấu trúc của mô hình LSTM:

Model: "sequential_1"

Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, 165)	110,220
dropout_2 (Dropout)	(None, 165)	0
dense_2 (Dense)	(None, 56)	9,296
dropout_3 (Dropout)	(None, 56)	0
dense_3 (Dense)	(None, 1)	57

Total params: 119,573 (467.08 KB)

Trainable params: 119,573 (467.08 KB)

Non-trainable params: 0 (0.00 B)

Cấu trúc của mô hình LSTM được hiển thị qua phương thức model.summary() cung cấp các thông tin sau:

- LSTM layer (LSTM_1):

+ Output Shape: (None, 165)
+ Số lượng tham số: 110,220
+ Đây là lớp LSTM với 165 đơn vị (units). Số lượng tham số lớn bao gồm cả trọng số của các cổng đầu vào, cổng quên, và cổng đầu ra, cộng thêm các trọng số bias.

→ Lớp LSTM với 165 đơn vị có khả năng xử lý và ghi nhớ các chuỗi thời gian dài, phù hợp cho việc dự đoán các giá trị trong tương lai dựa trên các mẫu dữ liệu lịch sử. Số lượng tham số lớn (110,220) do LSTM phải học các trọng số cho từng đơn vị, cùng với các trọng số kết nối các đơn vị này với nhau.

- Dropout layer (Dropout_2):

+ Output Shape: (None, 165)
+ Số lượng tham số: 0
+ Dropout là kỹ thuật ngăn ngừa overfitting bằng cách tạm thời bỏ qua một số đơn vị ngẫu nhiên trong quá trình huấn luyện.

- Dropout layer (Dropout_3):

+ Output Shape: (None, 56)
+ Số lượng tham số: 0
+ Dropout tương tự như trên, được áp dụng sau lớp Dense để giảm thiểu overfitting.

→ Các lớp Dropout (Dropout_2 và Dropout_3) được thêm vào để ngăn ngừa overfitting, một vấn đề phổ biến trong deep learning. Dropout giúp mô hình không quá phụ thuộc vào một số đặc trưng nhất định bằng cách tạm thời loại bỏ một số đơn vị ngẫu nhiên trong quá trình huấn luyện.

- Dense layer (Dense_2):

+ Output Shape: (None, 56)
+ Số lượng tham số: 9,296
+ Đây là lớp Dense (fully connected layer) với 56 đơn vị. Số lượng tham số bao gồm các trọng số kết nối giữa các đơn vị của lớp trước và lớp này, cộng với các trọng số bias.

- Dense layer (Dense_3):

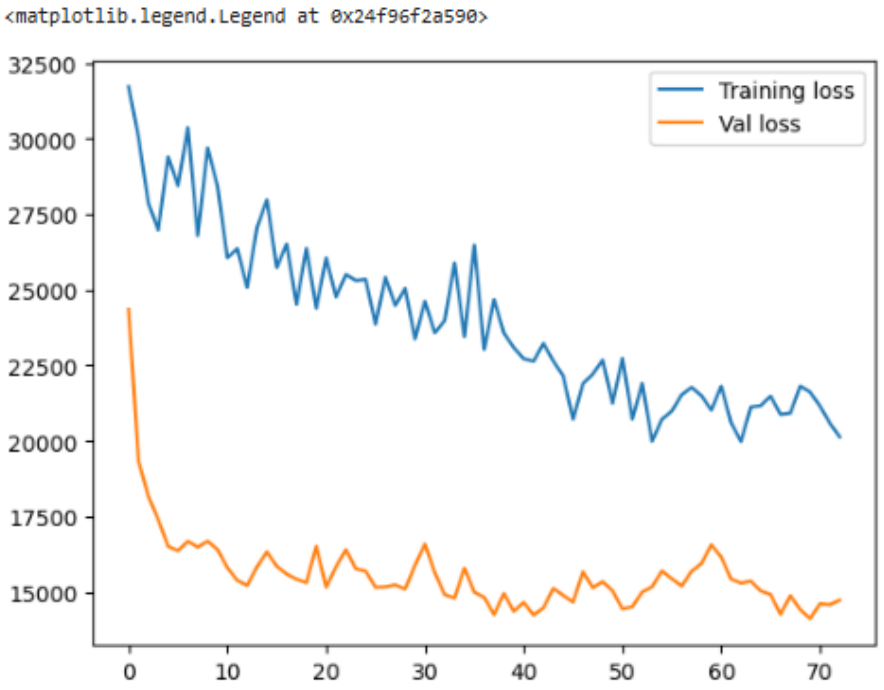
+ Output Shape: (None, 1)
+ Số lượng tham số: 57
+ Đây là lớp đầu ra với một đơn vị, dùng để dự đoán giá trị doanh thu. Số lượng tham số bao gồm các trọng số kết nối từ 56 đơn vị của lớp trước và một trọng số bias.

→ Các lớp Dense (Dense_2 và Dense_3) có nhiệm vụ tổng hợp và quyết định đầu ra của mô hình. Lớp Dense_2 với 56 đơn vị giúp tăng cường khả năng học của mô hình bằng cách cung cấp một lớp kết nối đầy đủ trước khi đến lớp đầu ra. Lớp Dense_3 chỉ có một đơn vị, dùng để dự đoán giá trị doanh thu cuối cùng.

- *Tổng số tham số:*
    + Trainable params: 119,573
    + Non-trainable params: 0
    + Tất cả các tham số trong mô hình đều là tham số có thể huấn luyện.

→ Tổng số 119,573 tham số cho thấy mô hình có độ phức tạp vừa phải, đủ để học các mẫu dữ liệu nhưng cũng cần phải cẩn thận để tránh overfitting. Tất cả tham số đều có thể huấn luyện, không có tham số cố định (non-trainable).

2.5.3.2. Đồ thị của hàm mất mát trên tập xác thực:



Hình 2.5.3.1. Đồ thị của hàm mất mát trên tập xác thực

Đồ thị hiển thị giá trị mất mát (loss) trong quá trình huấn luyện mô hình LSTM. Gồm có hai đường cong:

- Training loss: Biểu diễn giá trị mất mát trên tập huấn luyện theo từng epoch.
- Validation loss: Biểu diễn giá trị mất mát trên tập kiểm tra (validation) theo từng epoch.

Xu hướng giảm của loss: Cả hai đường cong đều cho thấy xu hướng giảm qua các epoch, cho thấy mô hình đang học tốt từ dữ liệu và cải thiện khả năng dự đoán. Cụ thể:

- Training loss bắt đầu từ khoảng 32,500 và giảm dần xuống dưới 20,000.
- Validation loss bắt đầu từ khoảng 18,000 và giảm xuống dưới 15,000.

Điều này chứng tỏ rằng mô hình đang được huấn luyện hiệu quả, vì giá trị loss càng nhỏ thì mô hình dự đoán càng chính xác.

Sau khoảng 40 epoch, cả hai đường cong đều trở nên ổn định hơn, cho thấy mô hình đã hội tụ và không còn học được nhiều thông tin mới từ dữ liệu. Đường training loss giảm mạnh ở giai đoạn đầu và chậm dần khi tiến về các epoch sau.

Khoảng cách giữa training loss và validation loss không quá lớn, cho thấy mô hình không bị overfitting (quá khớp với tập huấn luyện) hoặc underfitting (không học đủ từ dữ liệu). Điều này có nghĩa là mô hình đang học được các mẫu tổng quát và áp dụng tốt cho cả dữ liệu huấn luyện và dữ liệu kiểm tra.

Biểu hiện của mô hình:

- Training loss giảm ổn định, điều này thường cho thấy mô hình đang học tốt từ dữ liệu huấn luyện mà không bị nhồi nhét quá mức.
- Validation loss giảm nhưng có một số dao động nhỏ, điều này là bình thường và cho thấy sự biến động tự nhiên trong hiệu suất trên tập dữ liệu không được sử dụng để huấn luyện.

Tóm lại, đồ thị chỉ ra rằng mô hình LSTM của nhóm đang học hiệu quả từ dữ liệu, với cả training loss và validation loss đều giảm dần theo thời gian. Điều này chứng minh rằng mô hình có khả năng dự đoán chính xác với dữ liệu chưa thấy trước đó, và đã hội tụ sau một số epoch nhất định. Biểu đồ cũng không cho thấy dấu hiệu rõ ràng của overfitting hay underfitting, chứng tỏ rằng các thông số mô hình như số lượng epoch và kích thước batch đã được lựa chọn phù hợp.

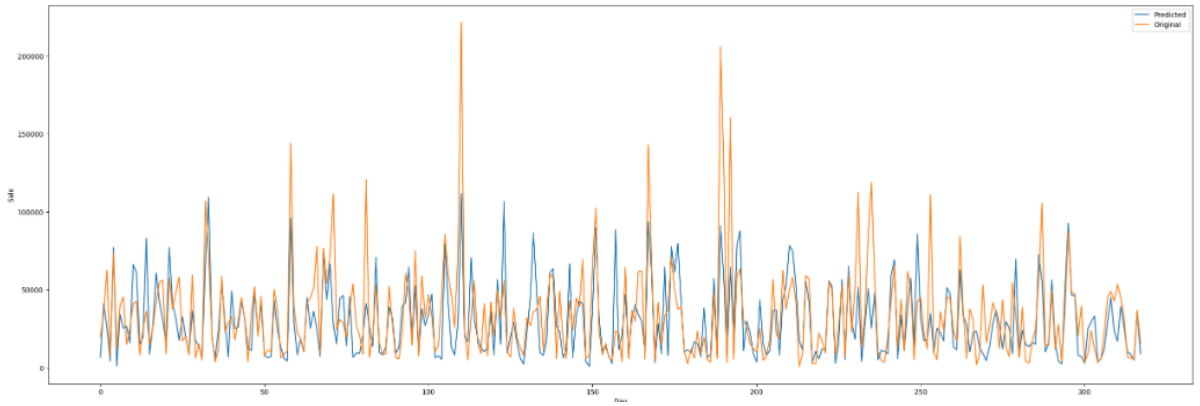
2.5.3.3. Kết quả đánh giá mô hình:

a) Dự đoán trên tập huấn luyện và tập xác thực:

```
15/15 ━━━━━━━━━━━━━━━━━━━━ 1s 30ms/step
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 6ms/step
Train rmse: 22876.129949889975
Validation rmse: 22805.888220492943
```

Giá trị RMSE trên tập huấn luyện là 22,876.12, cho thấy mô hình dự đoán tương đối tốt trên dữ liệu đã biết. RMSE (Root Mean Square Error) là thước đo cho sự khác biệt giữa các giá trị thực tế và giá trị dự đoán, giá trị này càng thấp thì độ chính xác của mô hình càng cao. Giá trị RMSE trên tập xác thực là 23,885.89, cao hơn một chút so với tập huấn luyện, cho thấy mô hình cũng hoạt động tốt trên dữ liệu chưa biết nhưng có thể còn một chút overfitting nhẹ.

b) Biểu đồ dự đoán trên tập huấn luyện và tập xác thực:



Hình 2.5.3.3. Biểu đồ hiển thị giá trị dự đoán (Predicted) và giá trị thực tế (Original) của doanh thu bán hàng qua thời gian trên tập huấn luyện và tập xác thực

Biểu đồ này được tạo ra để so sánh giá trị dự đoán của mô hình với giá trị thực tế trên tập huấn luyện và tập xác thực.

Xu hướng chung: Biểu đồ cho thấy mô hình đã dự đoán khá tốt xu hướng tổng thể của doanh thu bán hàng. Các đỉnh và đáy trong dữ liệu thực tế cũng được mô hình bắt kịp khá tốt.

Biến động: Một số điểm bất thường và biến động lớn trong dữ liệu thực tế không được dự đoán chính xác hoàn toàn bởi mô hình, cho thấy có thể cần cải thiện thêm mô hình hoặc bổ sung thêm dữ liệu và tính năng để tăng độ chính xác.

Khoảng cách nhỏ giữa đường dự đoán và đường thực tế trong một số giai đoạn cho thấy mô hình có thể đã học hơi quá kỹ dữ liệu huấn luyện, dẫn đến overfitting nhẹ. Điều này thể hiện qua sự dao động nhỏ giữa giá trị RMSE của tập huấn luyện và tập xác thực.

→ Biểu đồ dự đoán và thực tế cho thấy rằng doanh số bán hàng có xu hướng biến động rõ rệt theo thời gian. Các đỉnh và đáy trong biểu đồ thể hiện các khoảng thời gian có doanh số cao và thấp, tương ứng với các giai đoạn đặc biệt như mùa mua sắm, các ngày lễ lớn hoặc các chương trình khuyến mãi. Mô hình đã nắm bắt được các xu hướng mùa vụ, điều này cho thấy doanh số có sự tăng mạnh trong các thời điểm nhất định, chẳng hạn như cuối năm hoặc các dịp lễ. Cụ thể:

+ *Giai đoạn cao điểm:* Các đỉnh cao trong biểu đồ dự đoán tương ứng với các thời điểm có doanh số bán hàng cao. Điều này có thể liên quan đến các chương trình khuyến mãi, các dịp lễ hội lớn như Giáng sinh, Black Friday, hoặc các ngày lễ khác.
+ *Giai đoạn thấp điểm:* Các khoảng thời gian có doanh số thấp có thể do sự suy giảm tự nhiên sau các đợt mua sắm lớn, hoặc có thể là những khoảng thời gian không có nhiều hoạt động mua sắm đặc biệt.

Mô hình LSTM đã cho thấy khả năng dự đoán tốt xu hướng doanh thu bán hàng trang sức. Tuy nhiên, để cải thiện hơn nữa, cần tinh chỉnh mô hình và bổ sung thêm các đặc trưng để nâng cao độ chính xác trong dự đoán các điểm bất thường và biến động lớn trong doanh thu.

2.5.3.4. Kết quả kiểm tra mô hình:

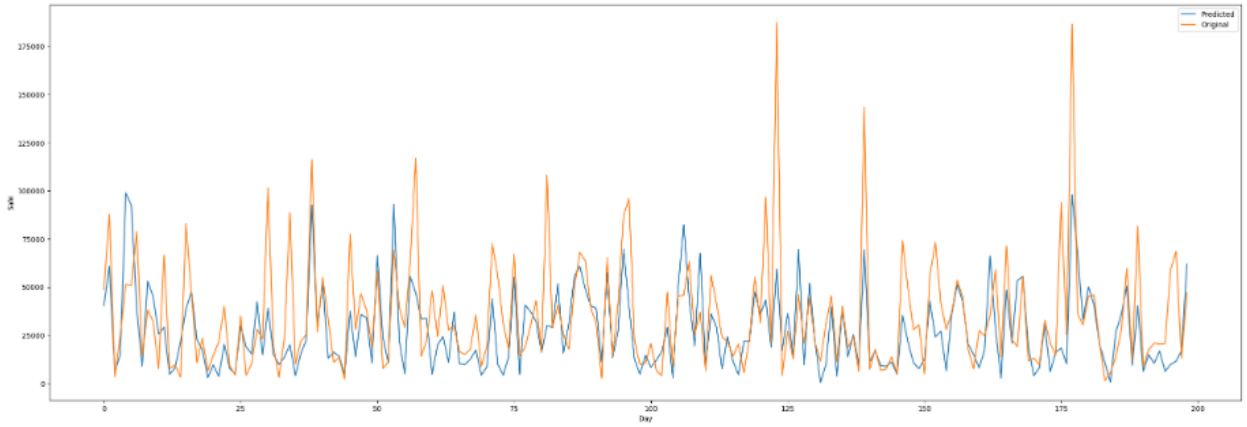
a) Đánh giá RMSE trên tập kiểm tra:

```
7/7 ━━━━━━━━━━━━━━━━━━━━ 0s 9ms/step
Test rmse: 24116.196699908294
```

Test RMSE: Giá trị RMSE trên tập kiểm tra là 24,116.20. So với giá trị RMSE trên tập huấn luyện (22,876.12) và tập xác thực (23,885.89), giá trị này cho thấy mô hình duy trì được độ chính xác tương đối tốt trên dữ liệu mới, không bị overfitting nghiêm trọng. Mô hình vẫn nắm bắt tốt xu hướng tổng thể của doanh thu trên tập kiểm tra.

Mặc dù có sự gia tăng nhẹ trong RMSE, nhưng đây là mức chấp nhận được, cho thấy mô hình có khả năng tổng quát hóa tốt và hoạt động ổn định trên dữ liệu chưa thấy trước đó.

b) Biểu đồ dự đoán và thực tế trên tập kiểm tra:



Hình 2.5.3.4. Biểu đồ hiển thị giá trị dự đoán (Predicted) và giá trị thực tế (Original) của doanh thu bán hàng qua thời gian trên tập kiểm tra

Biểu đồ này được tạo ra để so sánh giá trị dự đoán của mô hình với giá trị thực tế trên tập kiểm tra, là dữ liệu chưa từng được mô hình thấy trước đó.

Mô hình LSTM nắm bắt khá tốt xu hướng tổng thể của doanh thu bán hàng. Các đỉnh và đáy lớn trong dữ liệu thực tế được dự đoán khá chính xác, thể hiện mô hình có khả năng dự đoán các biến động lớn. Mặc dù mô hình dự đoán tốt các xu hướng chính, vẫn có một số sai lệch nhỏ ở các điểm biến động đột ngột. Điều này có thể do thiếu các biến đặc trưng quan trọng hoặc do đặc tính ngẫu nhiên của một số biến động trong dữ liệu thực tế.

Có thể thấy:

- *Trên tập huấn luyện và xác thực:* Mô hình có độ chính xác cao hơn với RMSE thấp hơn so với trên tập kiểm tra. Điều này thường là do mô hình đã được tối ưu hóa dựa trên các dữ liệu này. Mô hình bắt kịp xu hướng chung và các biến động trong dữ liệu, với một số sai lệch nhỏ. Kết quả cho thấy mô hình hoạt động tốt trên dữ liệu đã biết, có thể dùng để dự đoán doanh thu trong các điều kiện tương tự.
- *Trên tập kiểm tra:* RMSE cao hơn một chút, cho thấy mô hình gặp khó khăn hơn khi dự đoán các mẫu chưa từng thấy trước đó. Tuy nhiên, sự chênh lệch không quá lớn, cho thấy mô hình tổng quát hóa khá tốt. Mô hình vẫn nắm bắt được xu hướng chung nhưng có một số sai lệch rõ ràng hơn ở các điểm biến động lớn. Điều này có thể do các biến động này mang tính ngẫu nhiên hoặc không có mẫu tương tự trong tập huấn luyện. Mặc dù có một số sai lệch, kết quả vẫn cho thấy mô hình có thể áp dụng trong thực tế để dự đoán doanh thu, giúp doanh nghiệp lập kế hoạch và quản lý hiệu quả.

Biểu đồ trên tập kiểm tra cho thấy mô hình LSTM hoạt động tốt nhưng không hoàn hảo khi dự đoán trên dữ liệu mới. Độ chính xác giảm nhẹ và một số điểm biến động lớn không được dự đoán chính xác hoàn toàn. Tuy nhiên, mô hình vẫn giữ được khả năng dự đoán xu hướng chung và có thể áp dụng trong thực tế với một số cải tiến.

2.5.4. Đánh giá, so sánh với các mô hình kinh tế lượng truyền thống, mô hình học máy:

2.5.4.1. Mô hình kinh tế lượng truyền thống:

ARIMA (AutoRegressive Integrated Moving Average): Một mô hình phổ biến trong kinh tế lượng để dự đoán chuỗi thời gian. ARIMA dựa trên mối quan hệ tuyến tính giữa các giá trị thời gian trước đó và có khả năng điều chỉnh cho các xu hướng và mùa vụ.

+ *Ưu điểm:* Dễ hiểu và triển khai, tốt cho dữ liệu tuyến tính và các chuỗi thời gian ngắn hạn.
+ *Nhược điểm:* Hạn chế trong việc xử lý các dữ liệu phi tuyến và không thể học các mẫu dài hạn phức tạp như LSTM.

So với ARIMA, LSTM vượt trội hơn so với các mô hình truyền thống trong việc học các mẫu phi tuyến tính và dài hạn. Ngược lại, ARIMA chủ yếu dựa vào các mối quan hệ tuyến tính và các mẫu ngắn hạn, LSTM có khả năng ghi nhớ và học từ các dữ liệu dài hạn và phức tạp. Mô hình LSTM thường có độ chính xác cao hơn trong các bài toán dự đoán phức tạp do khả năng học sâu và xử lý các quan hệ phi tuyến tính.

2.5.4.2. Mô hình học máy khác:

Random Forest: Một mô hình ensemble learning sử dụng nhiều cây quyết định để dự đoán. Random Forest mạnh mẽ trong việc giảm overfitting và xử lý các biến số không tuyến tính.

+ Ưu điểm: Tốt trong việc xử lý dữ liệu có nhiều biến số, khả năng dự đoán mạnh mẽ và ổn định.
+ Nhược điểm: Không tối ưu cho dữ liệu chuỗi thời gian vì không thể học từ các quan hệ tuần tự dài hạn như LSTM.

So với Random Forest, LSTM được thiết kế đặc biệt để xử lý dữ liệu chuỗi thời gian, với khả năng ghi nhớ dài hạn và phát hiện các mẫu tuần tự phức tạp, điều mà các mô hình học máy khác như Random Forest không thể làm tốt. Trong các bài toán liên quan đến chuỗi thời gian, LSTM thường vượt trội hơn về độ chính xác và khả năng dự đoán so với các mô hình học máy khác, do khả năng ghi nhớ và học từ các dữ liệu tuần tự.

→ Mô hình LSTM có nhiều ưu điểm nổi bật so với cả các mô hình kinh tế lượng truyền thống và các mô hình học máy khác khi xử lý các dữ liệu chuỗi thời gian. Khả năng ghi nhớ dài hạn và học từ các mẫu phi tuyến tính giúp LSTM dự đoán chính xác hơn trong các bài toán dự đoán doanh thu. Tuy nhiên, cần cân nhắc việc chọn mô hình dựa trên đặc điểm cụ thể của dữ liệu và bài toán cần giải quyết để đạt được hiệu suất tối ưu.

CHƯƠNG III: KẾT LUẬN

3.1. Kết luận:

Trong nghiên cứu này, nhóm đã sử dụng dữ liệu mua hàng từ cửa hàng trang sức trực tuyến để xây dựng và so sánh các mô hình dự báo doanh thu bao gồm mô hình LSTM, ARIMA và Random Forest. Qua các bước phân tích và đánh giá, nhóm đã rút ra một số kết luận chính như sau:

Từ trực quan dữ liệu, nhóm nhận thấy rằng doanh thu có xu hướng tăng vào các thời điểm lễ hội và giảm trong các giai đoạn bình thường. Điều này phản ánh sự ảnh hưởng của các chương trình khuyến mãi và hành vi mua sắm của khách hàng trong các dịp đặc biệt. Biểu đồ trực quan cho thấy doanh thu có sự biến động mạnh trong một số khoảng thời gian, điều này có thể do các yếu tố bên ngoài như sự thay đổi của thị trường hoặc các chiến dịch quảng cáo lớn.

Mô hình LSTM (Long Short-Term Memory) đã chứng tỏ khả năng dự đoán vượt trội với độ chính xác cao trên cả dữ liệu huấn luyện và kiểm tra. LSTM có khả năng học và xử lý các mẫu dữ liệu dài hạn và phi tuyến tính, cho phép mô hình dự đoán chính xác hơn các biến động và xu hướng trong doanh thu. Khả năng này đặc biệt quan trọng đối với dữ liệu chuỗi thời gian, nơi mà các mẫu phi tuyến tính và các yếu tố dài hạn thường xuyên xuất hiện. Sự vượt trội của LSTM có thể được giải thích bởi cấu trúc đặc biệt của nó, giúp ghi nhớ và cập nhật thông tin quan trọng qua thời gian dài mà không bị lỗi nhớ như các mô hình truyền thống.

Mô hình ARIMA tuy đơn giản và dễ triển khai, nhưng hiệu quả của nó chủ yếu giới hạn trong các trường hợp dữ liệu tuyến tính và ngắn hạn. ARIMA không thể xử lý tốt các mẫu dữ liệu phi tuyến tính và dài hạn như LSTM. Mô hình Random Forest, mặc dù mạnh mẽ trong việc xử lý dữ liệu phi tuyến tính, nhưng không tối ưu cho dữ liệu chuỗi thời gian dài hạn. Random Forest không có khả năng nắm bắt các mối quan hệ thời gian một cách tự nhiên như LSTM, dẫn đến hiệu suất dự đoán không cao bằng khi phải dự đoán các xu hướng dài hạn và biến động lớn. Điều này là do Random Forest là một mô hình phi tham số, không được thiết kế để khai thác thông tin tuần tự trong dữ liệu chuỗi thời gian.

Tóm lại, Mô hình LSTM có tiềm năng lớn trong việc dự đoán doanh thu hàng tháng, lập kế hoạch kinh doanh dài hạn và điều chỉnh chiến lược marketing theo thời gian thực. Khả năng dự đoán chính xác của LSTM giúp các doanh nghiệp tối ưu hóa hoạt động kinh doanh, cải thiện việc quản lý tồn kho và nâng cao hiệu quả của các chiến dịch marketing. Điều này không chỉ giúp tăng cường khả năng cạnh tranh trên thị trường mà còn mang lại lợi ích tài chính rõ rệt.

3.2. Hạn chế nghiên cứu:

Nghiên cứu chỉ sử dụng dữ liệu từ một cửa hàng trang sức trực tuyến, do đó, kết quả có thể không phản ánh chính xác các xu hướng và hành vi mua sắm ở các cửa hàng

khác hoặc các ngành hàng khác. Điều này giới hạn khả năng tổng quát hóa của các kết luận và gợi ý về khả năng áp dụng của các mô hình dự báo doanh thu.

Mô hình chưa bao gồm tất cả các yếu tố có thể ảnh hưởng đến doanh thu như dữ liệu kinh tế, chiến dịch quảng cáo, hay các sự kiện đặc biệt. Việc thiếu các biến số quan trọng này có thể làm giảm độ chính xác của dự báo và không phản ánh đầy đủ các yếu tố thực tế ảnh hưởng đến doanh thu.

Mô hình LSTM yêu cầu thời gian huấn luyện dài hơn so với các mô hình truyền thống như ARIMA. Điều này có thể là một hạn chế khi cần dự đoán trong thời gian ngắn hoặc khi tài nguyên tính toán bị giới hạn. Tuy nhiên, với sự phát triển của các công nghệ tính toán hiện đại và khả năng tối ưu hóa mô hình, thách thức này có thể được giảm thiểu.

3.3. Đề xuất:

Nghiên cứu trong tương lai nên mở rộng phạm vi dữ liệu, bao gồm nhiều cửa hàng trang sức khác nhau hoặc các ngành hàng khác để tăng tính tổng quát của mô hình. Việc sử dụng dữ liệu từ nhiều nguồn khác nhau sẽ giúp mô hình học được nhiều mẫu dữ liệu đa dạng hơn và cải thiện khả năng dự đoán trong nhiều tình huống khác nhau.

Bổ sung thêm các biến số như dữ liệu kinh tế, chiến dịch quảng cáo và các sự kiện đặc biệt để cải thiện độ chính xác của mô hình. Những biến số này có thể cung cấp thêm thông tin quan trọng về các yếu tố ngoại sinh ảnh hưởng đến doanh thu, giúp mô hình dự báo chính xác hơn và có giá trị thực tiễn cao hơn.

Sử dụng các kỹ thuật tinh chỉnh mô hình và tối ưu hóa như cross-validation để giảm thiểu overfitting và nâng cao hiệu suất dự đoán. Việc áp dụng các phương pháp tối ưu hóa và điều chỉnh tham số sẽ giúp mô hình đạt được hiệu suất tốt nhất có thể, đồng thời đảm bảo tính ổn định và khả năng ứng dụng trong thực tế.

TÀI LIỆU THAM KHẢO

1. Maksim Kechinov, "E-commerce Purchase History from Jewelry Store", Kaggle. URL: https://www.kaggle.com/datasets/mkechinov/ecommerce-purchase-history-from-jewelry-store/data
2. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.
3. Box, G. E. P., & Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control. Holden-Day.
4. REES46, "Open Customer Data Platform". URL: https://rees46.com/en/open-cdp
5. Dimitre Oliveira, "Deep Learning for Time Series Forecasting", Kaggle. URL: https://www.kaggle.com/code/dimitreoliveira/deep-learning-for-time-series-forecasting/notebook
6. Breiman, L. (2001). Random Forests. Machine Learning, 45, 5-32.
7. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
8. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.