ĐẠI HỌC QUỐC GIA HỒ CHÍ MINH
CUỘC THI VIETNAM DATATHON 2023

Lời mở đầu:

Trong bối cảnh thị trường bán lẻ ngày càng cạnh tranh, các doanh nghiệp cần tìm ra các cách mới để tối đa hóa doanh số bán hàng và lợi nhuận. Mô hình Machine Learning (ML) đã nổi lên như một động lực thúc đẩy trong lĩnh vực kinh doanh nói chung và lĩnh vực kinh doanh mặc hàng giày dép nói riêng, giúp các công ty đưa ra quyết định dựa trên dữ liệu thúc đẩy doanh số bán hàng và tối ưu hóa quản lý hàng tồn kho. Bằng cách khai thác sức mạnh của thuật toán Machine Learning, các doanh nghiệp có thể mở khóa kho tàng thông tin chi tiết từ lượng lớn dữ liệu bán hàng và hàng tồn kho, cho phép họ xây dựng các chiến lược bán hàng hiệu quả, dự báo chính xác doanh số bán hàng trong tương lai và triển khai các phương thức quản lý hàng tồn kho thông minh.

Trong lĩnh vực gợi ý chiến lược bán hàng, các thuật toán Machine Learning có thể phân tích dữ liệu bán hàng lịch sử, mô hình hành vi của khách hàng và xu hướng thị trường để xác định thời điểm tối ưu để bán các sản phẩm cụ thể. Khả năng này giúp các doanh nghiệp tận dụng tối đa các thời điểm nhu cầu cao điểm và tối đa hóa doanh thu bằng các liên kết các nỗ lực bán hàng với những thời điểm thuận lợi nhất.

Machine Learning cũng đóng một vai trò quan trọng trong dự báo và gợi ý chiến lược bán hàng. Bằng cách phân tích dữ liệu bán hàng lịch sử, xu hướng thị trường và các chỉ số kinh tế, các thuật toán Machine Learning có thể tạo ra dự báo chính xác về xu hướng bán hàng trong tương lai. Những dự báo này cung cấp thông tin chi tiết có giá trị về hiệu suất bán hàng dự kiến, cho phép các doanh nghiệp chủ động điều chỉnh các chiến dịch tiếp thị, lịch trình sản xuất và mức độ nhân sự để đáp ứng nhu cầu dự kiến. Ngoài ra, các thuật toán Machine Learning có thể phân tích sâu hơn để giải thích các yếu tố ảnh hưởng đến hiệu suất bán hàng, giúp các doanh nghiệp xác định các lĩnh vực cần cải thiện và tối ưu hóa chiến lược của họ cho phù hợp.

Quản lý hàng tồn kho, một khía cạnh quan trọng của hiệu quả chuỗi cung ứng, đang được cách mạng hóa bởi các mô hình Machine Learning. Các thuật toán ML có thể phân tích dữ liệu hàng tồn kho lịch sử, xu hướng bán hàng và thời gian dẫn của nhà cung cấp để tạo ra dự

báo chính xác về nhu cầu hàng tồn kho trong tương lai. Những dự báo này cho phép các doanh nghiệp duy trì mức tồn kho tối ưu, giảm rủi ro thiếu hàng và hàng tồn kho dư thừa, cả hai đều có thể ảnh hưởng tiêu cực đến sự hài lòng của khách hàng và lợi nhuận. Ngoài ra, các thuật toán ML có thể trực quan hóa và phân tích dữ liệu hàng tồn kho, cung cấp cho doanh nghiệp thông tin chi tiết theo thời gian thực về mức tồn kho, mô hình di chuyển sản phẩm và tình trạng thiếu hàng tiềm ẩn, cho phép họ đưa ra quyết định chủ động để duy trì hàng tồn kho đầy đủ.

Vậy câu hỏi được đặt ra ở đây là: “Liệu ứng dụng mô hình Machine Learning cho việc dự đoán các hoạt động kinh doanh của một doanh nghiệp bán lẻ lĩnh vực giày dép có thực sự hiệu quả và tối ưu hóa được doanh thu của doanh nghiệp sử dụng mô hình không?”
Để giải quyết câu hỏi trên, chúng em sẽ so sánh khả năng dự đoán của các thuật toán khác nhau cho việc giải quyết bài toán doanh thu và quản lý hàng tồn kho. Về dataset, nhóm chúng em quyết định sử dụng bộ dataset 2 mà ban tổ chức đã cung cấp và sẽ thực việc dự đoán một cách chính xác bằng việc sử dụng các bộ thuật toán trong Machine Learning. Kết quả mà nhóm chúng em nghiên cứu có thể giúp cho các doanh nghiệp ứng dụng tốt trong việc giải quyết bài toán doanh thu và quản lý hàng tồn kho.

Mục tiêu ý tưởng:

Từ ý tưởng về việc áp dụng mô hình ML vào dự án, chúng tôi kỳ vọng rằng, ý tưởng này trong sự kết hợp với các mô hình sẽ giúp đề tài có thể:

1. **Gợi ý Chiến Lược Bán Hàng:** Hãy cung cấp ví dụ về cách mô hình ML có thể phân tích lịch sử mua sắm của khách hàng, xác định xu hướng và dự đoán thời điểm lý tưởng để triển khai chiến lược khuyến mãi hoặc giảm giá để tối ưu hóa doanh số bán hàng.
2. **Dự Báo Doanh Số Bán Hàng:** Mô tả cụ thể về cách ML có thể sử dụng dữ liệu lịch sử để dự đoán doanh số bán hàng trong tương lai và làm thế nào thông tin này có thể hỗ trợ quyết định chiến lược kinh doanh.
3. **Quản Lý Hàng Tồn Kho Thông Minh:** Trình bày ví dụ về cách ML có thể phân tích xu hướng hàng tồn kho, dự đoán nhu cầu hàng tồn kho tương lai, và giúp doanh nghiệp duy trì mức tồn kho tối ưu để tránh tình trạng thiếu hụt hoặc hàng tồn kho dư thừa.
4. **Tối Ưu Hóa Chuỗi Cung Ứng:** Mô tả cách ML có thể được sử dụng để tối ưu hóa quy trình chuỗi cung ứng bằng cách dự đoán thời gian dẫn và lượng cung cấp cần thiết.
5. **Ưu Đãi Cá Nhân Hóa:** Giới thiệu cách ML có thể phân tích hành vi mua sắm cá nhân của khách hàng để tạo ra ưu đãi cá nhân hóa, tăng cường trải nghiệm mua sắm và tăng cường sự trung thành của khách hàng.

Giới thiệu thành viên nhóm:

1. **Nguyễn Trần Minh Thư:**
   a. **Thông tin cá nhân:**
      - Họ tên: Nguyễn Trần Minh Thư
      - Ngày sinh: 15/05/2004
      - Nghề nghiệp: Sinh viên
      - Nơi công tác: Đại học Khoa học Tự nhiên - ĐHQG HCM
      - Số điện thoại: 0977150504

   b. **Thành tích:**

Thời gian	Thành tích
12/2020 12/2021	Bronze Medalist of IYMC 2020 International Mathematics Competition Silver Medalist of IYMC 2021 International Mathematics Competition
8/2021	Impressive Prize of HIC competition organized by HAEC Inception Camp
10/2021	3rd Prize of Model Mathematics Competition
1/2021 - present	Co-founder of The Bridge Project Vice Leader of The Bridge Project (1/2021 - 12/2021) Leader of The Bridge Project (1/2022 - 12/2022) Advisor of The Bridge Project (1/2023 - present)
10/2022 - present	Student of Information Technology department of University of Science, Ho Chi Minh Vietnam National University
11/2022	2nd Prize in a competition called “Bản lĩnh IT” held by Information Technology department

2. **Phạm Thanh Hưng:**
   a. **Thông tin cá nhân:**
      - Họ tên: Phạm Thanh Hưng
      - Ngày sinh: 27/10/2003
      - Nghề nghiệp: Sinh viên
      - Nơi công tác: Đại học Kinh tế Luật - ĐHQG HCM
      - Số điện thoại: 0981913075

   b. **Thành tích:**

Thời gian	Thành tích

10/2022 - present	Student of Mathematical Economics department of University of Economics and Law, Ho Chi Minh Vietnam National University

3. Dương Gia Hân:

a. Thông tin cá nhân:

- Họ tên: Dương Gia Hân
- Ngày sinh:13/05/2003
- Nghề nghiệp: Sinh viên
- Nơi công tác: Đại học Khoa học Tự nhiên - ĐHQG HCM
- Số điện thoại:0397774304

b. Thành tích:

Thời gian	Thành tích
10/2021 - present	Student of Information Technology department of University of Science, Ho Chi Minh Vietnam National University

4. Nguyễn Thành Tài:

a. Thông tin cá nhân:

- Họ tên: Nguyễn Thành Tài
- Ngày sinh: 03/03/2004
- Nghề nghiệp: Sinh viên
- Nơi công tác: Đại học Khoa học Tự nhiên - ĐHQG HCM
- Số điện thoại:0763579678

b. Thành tích:

Thời gian	Thành tích
10/2022-present	Student of Mathematical - Computer Science of University of Science, Ho Chi Minh Vietnam National University

5. Huỳnh Hà Phương Linh:

a. Thông tin cá nhân:

- Họ tên: Huỳnh Hà Phương Linh
- Ngày sinh:
- Nghề nghiệp: Sinh viên
- Nơi công tác: Đại học Khoa học Tự nhiên - ĐHQG HCM
- Số điện thoại:

b. Thành tích:

Thời gian	Thành tích
2019	Gold Medalist, National Internet Olympiads of English
2021	Honorable Mention, Vietnam National Olympiad in English Issued by Vietnam Ministry of Education and Training
4/ 2022	Third Prize, Vietnam National Olympiad in English Issued by Vietnam Ministry of Education and Training
8/2022	Odon Vallet Fellowship Issued by Rencontres du Vietnam
10/2022 - present	Student of Information Technology department of University of Science, Ho Chi Minh Vietnam National University

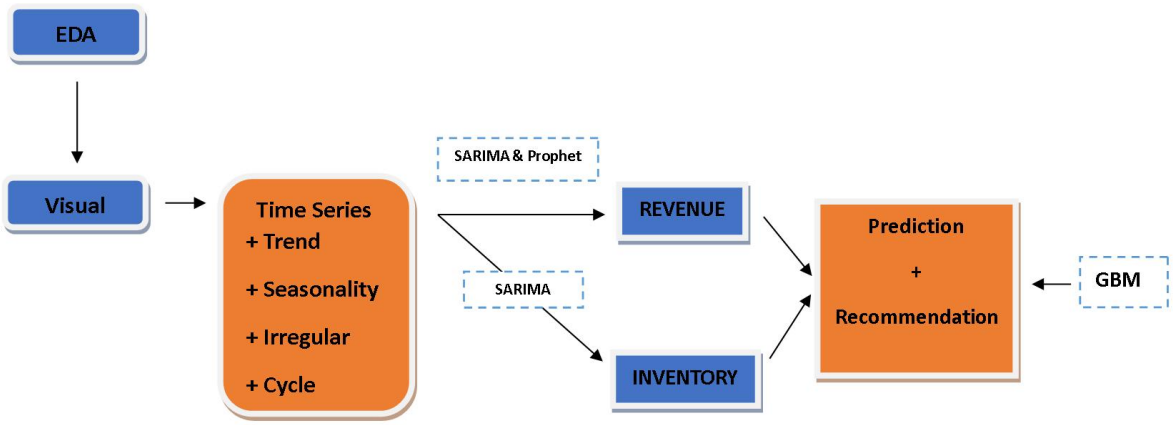
Khái quát ý tưởng thực hiện:

1. **Lựa chọn Dataset:**
   - Dataset nhóm quyết định lựa chọn là dataset 2 của chương trình Vietnam Datathon đã đưa ra.
2. **Khám phá dữ liệu (EDA):**
   - Thực hiện việc xử lý các dữ liệu bị thiếu, bị điền sai trong dataset:
     Việc dataset bị noise hoặc bị thiếu trong quá trình thu thập là một điều không thể tránh khỏi và để xử lý các dữ liệu bị noise và bị thiếu, nhóm sẽ thực hiện bằng cách xóa hoặc thêm dữ liệu thay thế (fillna bằng dữ liệu có chỉ số yếu vị cao nhất) vào. Ở đây nhóm đã quyết định thêm dữ liệu thay thế vào để tránh việc bộ dataset bị thiếu, ảnh hưởng đến việc dự đoán của mô hình.
   - Data Visualization:
     Ở bước này, nhóm sẽ thực hiện vẽ các biểu đồ để tìm hiểu xem rằng các số liệu liên quan như giá vốn và giá bán. Trong bước này nhóm cũng sẽ xem rằng các biến trong dataset tương quan với nhau như thế nào bằng cách vẽ biểu đồ Scatter Plot (biểu đồ phân tán) và biểu đồ Heatmap (biểu đồ nhiệt) để trực quan hóa mối quan hệ giữa các biến càng xét trong dataset.
3. **Gộp dữ liệu:**
   - Sơ lược cách làm: Gộp các file dữ liệu liên quan đến sales và inventory thành một file sales và một file inventory.
   - Mục tiêu: Dễ thống kê và đánh giá sự thay đổi về doanh số và về lượng hàng tồn kho.
4. **Phân tích doanh số bán hàng:**
   - Sơ lược cách làm: Sử dụng mô hình SARIMA và mô hình Prophet để dự đoán doanh số bán hàng theo mùa và sử dụng Prophet để phân tích doanh số vào những ngày nghỉ.
   - Mục tiêu: Xác định xu hướng và ảnh hưởng của yếu tố mùa vụ cũng như ngày nghỉ với doanh thu bán hàng.
5. **Phân tích lượng hàng tồn kho:**
   - Sơ lược cách làm: Áp dụng mô hình (ARIMA, SARIMA) để dự đoán lượng tồn kho trong tương lai.
   - Mục tiêu: Kiểm tra mối liên hệ giữa doanh số bán hàng và lượng hàng tồn kho.
6. **Đánh giá mối liên hệ giữa các yếu tố:**
   - Sơ lược cách làm: Sử dụng hệ số tương quan để đánh giá doanh số bán hàng và lượng hàng tồn kho, đồng thời sử dụng biểu đồ scatter để minh họa.

- Mục tiêu: Làm nền tảng, tạo tiền đề để xây dựng chiến lược kinh doanh phù hợp.

7. **Xây dựng chiến lược kinh doanh hiệu quả:**
   - Sơ lược về cách làm: Sử dụng phương pháp GBM (Gradient Boosting) để đưa ra các quyết định liên quan đến chiến lược quản lý tồn kho và tối ưu hóa doanh số bán hàng.
   - Mục tiêu: Đặt mức tồn kho an toàn và tăng doanh số bán hàng dựa trên xu hướng và thị hiếu thị trường.
8. **Theo dõi và đánh giá sự khả thi:**
   - Sơ lược về cách làm: Theo dõi thực tế so với dự đoán, lấy phản hồi từ những người có liên quan.
   - Mục tiêu: Đảm bảo kế hoạch đưa ra hiệu quả và khả thi, giúp ích được cho sự phát triển bền vững của doanh nghiệp nói chung cũng như các doanh nghiệp hoạt động trong ngành giày dép nói riêng.

Các bước thực hiện:



Hình 1: Sơ đồ các bước thực hiện

1. **EDA**
    a. **Công cụ sử dụng**:

- Jupyter notebook: là một nền tảng tính toán khoa học mã nguồn mở, có thể sử dụng để tạo và chia sẻ các tài liệu có chứa code trực tiếp, phương trình, trực quan hóa dữ liệu và văn bản tường thuật. Với Jupyter Notebook, người dùng có thể đưa dữ liệu, code, hình ảnh, công thức, video,.. vào trong cùng một file, giúp cho việc trình bày trở nên dễ dàng hơn.
- Các thư viện sử dụng trong jupyter notebook:
    - Pandas
    - Matplotlib
    - Scikit learn
    - Seaborn
    - Numpy

    b. **Các bước thực hiện:**
    Kết quả thực hiện có thể xem qua đường dẫn github này:
    https://github.com/Amature123/sale_inv/blob/master/Sale.ipynb

    i. Quan sát và xử lý data

- Trong dataset 2 được cho, gồm 3 tệp nhỏ bao gồm sale_data, inventory_data, master_data. Sau khi xem sơ

lược qua các tệp thì chúng tôi quyết định gộp các file nhỏ trong sale_data và inventory_data lại thành 1 file duy nhất chứa toàn bộ dữ liệu. Riêng master_data do không cùng feature nên sẽ phân tích riêng từng dataset.

- Sau khi đã xử lý xong bắt đầu tiến hành xuất và quan sát dữ liệu sơ bộ bằng thư viện pandas. Ở đây, chúng tôi sẽ import mẫu 1 dataset sale_data sau khi đã gộp các file lại với nhau.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 578987 entries, 0 to 60223
Data columns (total 13 columns):
 #   Column                     Non-Null Count   Dtype 
---  ------                     --------------   ----- 
 0   month                      578987 non-null  int64 
 1   week                       578987 non-null  int64 
 2   site                       578987 non-null  int64 
 3   branch_id                  578987 non-null  int64 
 4   channel_id                 578987 non-null  object
 5   distribution_channel       578987 non-null  object
 6   distribution_channel_code  578987 non-null  object
 7   sold_quantity              578987 non-null  int64 
 8   cost_price                 578987 non-null  int64 
 9   net_price                  578987 non-null  int64 
 10  customer_id                578986 non-null  object
 11  product_id                 578987 non-null  object
 12  Month                      578987 non-null  object
dtypes: int64(7), object(6)
memory usage: 61.8+ MB
```

Hình 2: Xem những loại dữ liệu có trong dataset

	month	week	site	branch_id	channel_id	distribution_channel	distribution_channel_code	sold_quantity	cost_price	n
0	2022001	202201	1800	1800	Online	Online	ZF2	1	495720	
1	2022001	202204	1116	1100	CHTT	Bán lẻ	FP	1	221000	
2	2022001	202201	1134	1100	CHTT	Bán lẻ	FP	1	255000	
3	2022001	202204	1612	1600	CHTT	Bán lẻ	FP	1	258400	
4	2022001	202202	1511	1500	CHTT	Bán lẻ	FP	1	272000	

Hình 3: 5 hàng dữ liệu đầu tiên trong dataset

ii. Sửa đổi tên cột và làm sạch dữ liệu

- Trong bước này, chúng tôi tiến hành thống kê các số liệu có trong dataset bằng pandas.DataFrame.describe().

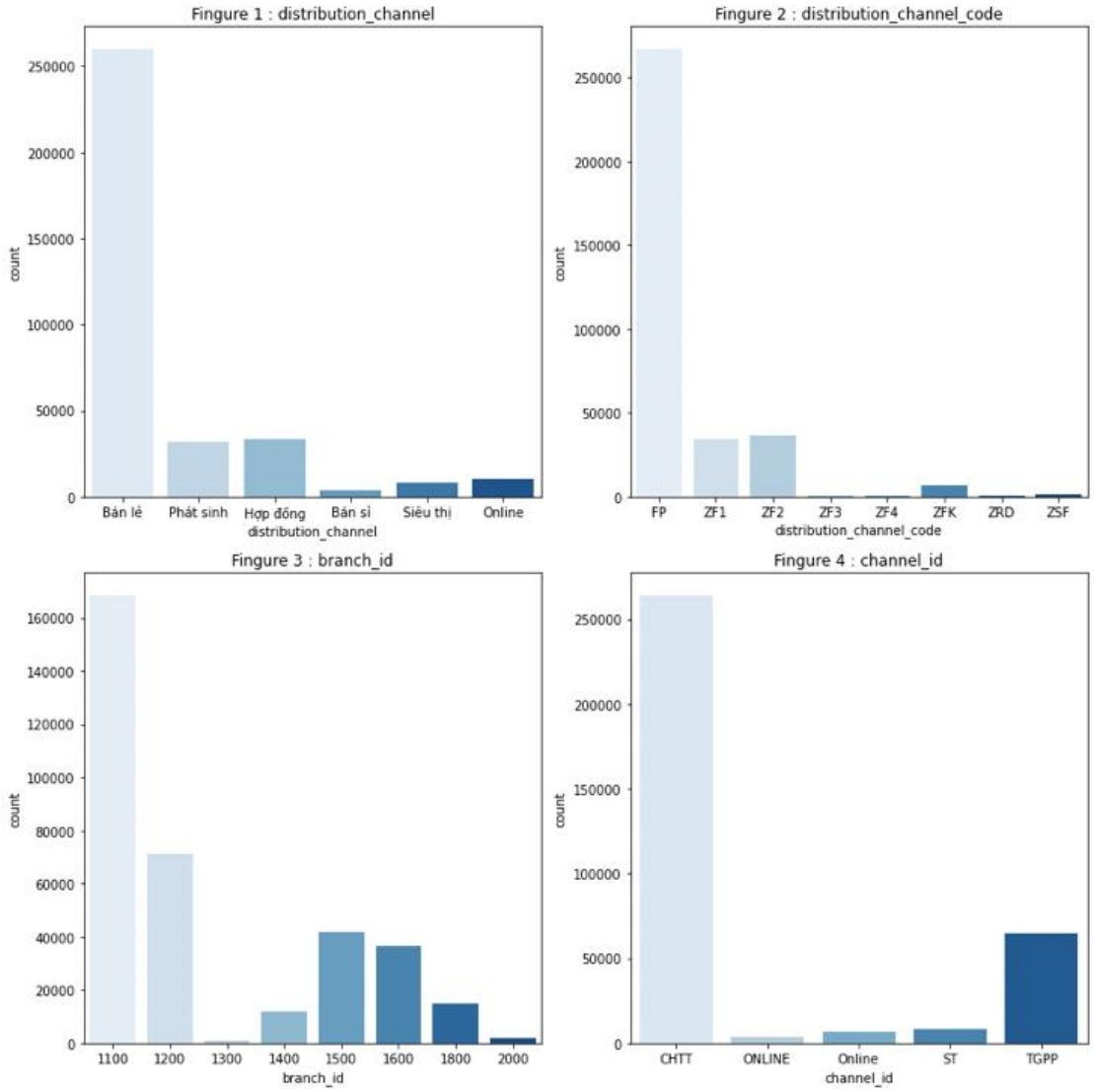
	month	week	site	branch_id	sold_quantity	cost_price	net_price
count	3.473920e+05	347392.000000	347392.000000	347392.000000	347392.000000	3.473920e+05	3.473920e+05
mean	2.022006e+06	202225.600955	1284.336035	1266.477927	1.395864	3.021302e+05	3.945433e+05
std	3.640614e+00	16.900131	214.239542	219.318489	3.028410	6.216898e+05	7.049928e+05
min	2.022001e+06	202153.000000	1100.000000	1100.000000	-12.000000	-9.846300e+06	-1.073606e+07
25%	2.022003e+06	202212.000000	1118.000000	1100.000000	1.000000	1.168360e+05	1.570000e+05
50%	2.022007e+06	202227.000000	1200.000000	1200.000000	1.000000	2.015000e+05	2.850000e+05
75%	2.022010e+06	202239.000000	1503.000000	1500.000000	1.000000	3.204660e+05	4.570000e+05
max	2.022012e+06	202252.000000	2001.000000	2000.000000	400.000000	6.082435e+07	7.424144e+07

Hình 4: Các chỉ số thống kê của các dữ liệu số trong dataset

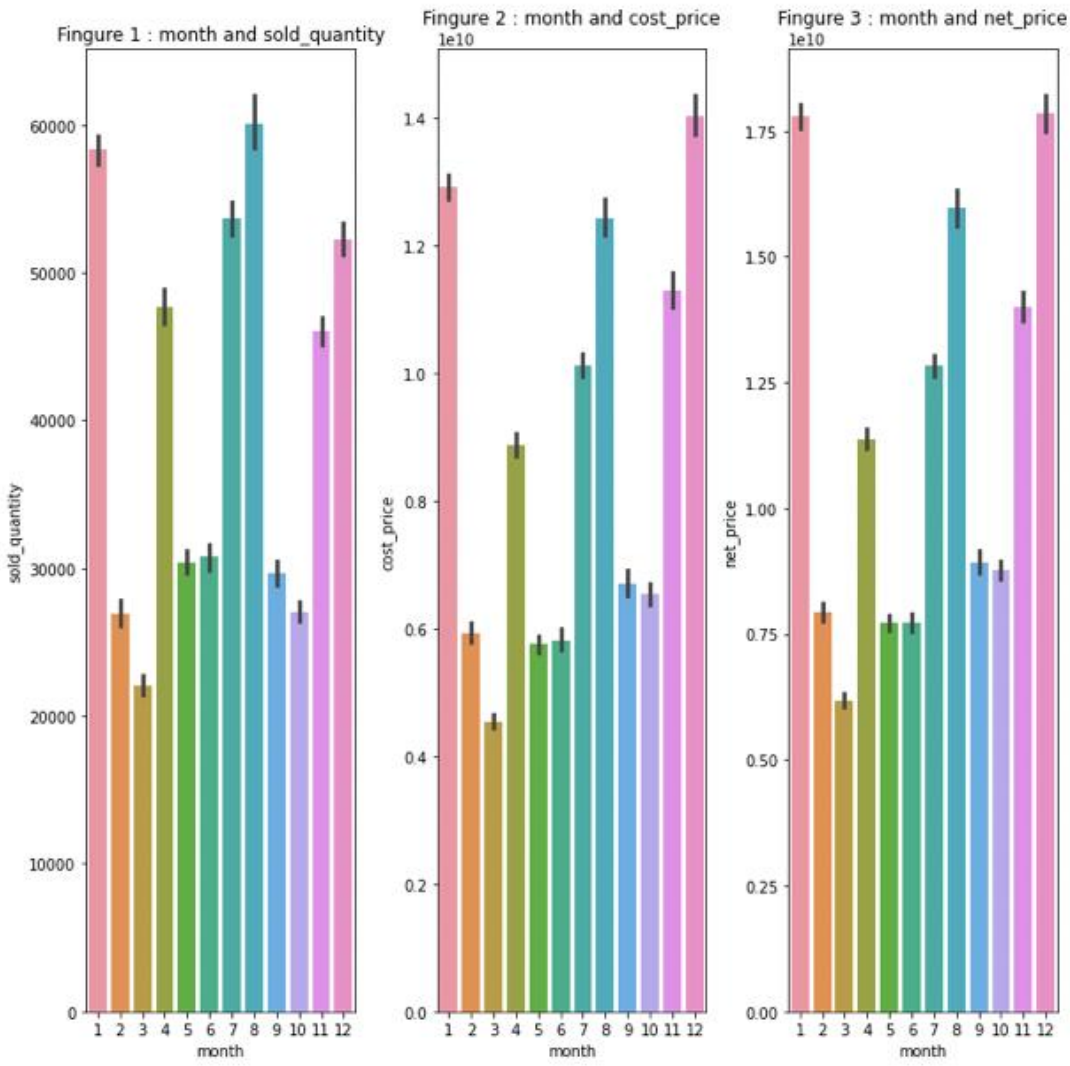
- Thông qua quá trình phân tích, tệp dữ liệu bị thiếu dữ liệu và các loại dữ liệu không đúng với bản chất của nó (ví dụ: cột month trong bài trên thay vì chỉ nên hiện tháng nhưng lại hiện dữ liệu theo dạng năm + tháng nên việc xử lý trở nên khó khăn).

iii. Trực quan hóa dữ liệu:

Sử dụng thư viện matplotlib và seaborn giúp cho việc trực quan dữ liệu dễ dàng hơn.
Dưới đây là một số hình biểu đồ mà nhóm đã tiến hành trực quan hóa:



Hình 5: Biểu đồ Histogram 1



Hình 6: Biểu đồ Histogram 2

iv. Tiền xử lý dữ liệu:

Ở bước này, chúng tôi sẽ tiến hành label encoding cho tất cả các dữ liệu định tính, giúp cho sử dụng model trở nên tối ưu hóa hơn.

2. Phân tích doanh số bán hàng

a. Các dữ liệu về doanh số bán hàng:

- month: month in year
- week: week in year
- site: warehouse/ store ID
- branch_id: the brand ID
- channel_id: code of distribution channel
- distribution_channel_code: kind of distribution channel
- sold_quantity: the number of products or units that have been sold

- cost_price: the direct costs associated with producing or purchasing the goods that a company sells during a specific accounting period
- net_price: the final price of a product
- customer_id: the customer ID
- product_id: the product ID

b. Dự đoán ban đầu:

Doanh thu của công ty đối với từng sản phẩm chịu sự ảnh hưởng của nhiều yếu tố khác nhau, bao gồm cả yếu tố khách quan và chủ quan.

- Yếu tố khách quan: Các điều kiện thời tiết, mùa vụ (như năm, tháng, tuần), vị trí địa lý (như cửa hàng), độ nhận diện thương hiệu, kênh phân phối (như kênh, phân phối), giá cả (net_price) đều có ảnh hưởng đến số lượng hàng hóa bán ra. Sự tác động của chúng có thể khác nhau và sẽ được phân tích chi tiết bên dưới để minh chứng cho những dự đoán ban đầu.
- Yếu tố chủ quan: Thông tin về người mua hàng (customer_id) thể hiện sự ảnh hưởng ở nhiều khía cạnh, như tình hình tài chính và khả năng chi trả của họ, cũng như sở thích và phong cách mà họ tự định vị.

Thông qua việc phân loại các yếu tố thành khách quan và chủ quan, chúng ta có thể tiếp cận mỗi khía cạnh một cách rõ ràng và phân tích chúng để có cái nhìn tổng thể về ảnh hưởng của chúng đối với doanh số bán hàng.

c. Phương pháp kiểm chứng:

- **Kiến thức nền tảng:** Sử dụng mô hình SARIMA và mô hình Prophet để phân tích, dự đoán doanh số kinh doanh.
- **Ưu nhược điểm của từng mô hình:**
  + Đối với mô hình SARIMA: SARIMA được thiết kế đặc biệt để xử lý yếu tố mùa vụ trong dữ liệu thời gian của dữ liệu cần xử lý, đặc biệt quan trọng khi doanh số kinh doanh có xu hướng biến động theo chu kỳ theo mùa. Bên cạnh đó, mô hình này thường thích hợp cho các loại dữ liệu thời gian có xu hướng và yếu tố mùa vụ.
    → Bài toán hiện tại chúng tôi cần xử lý là về lĩnh vực thời trang, trên thực tế, thời trang hay bất kỳ một

khía cạnh nào khác trong cuộc sống (ví dụ như du lịch - mùa hè người dân sẽ đi du lịch nhiều hơn những giai đoạn khác hay ngành nông nghiệp - ngành nghề mà chúng ta có thể thấy rõ nhất rằng là mỗi một mùa sẽ thích hợp với một loài cây khác nhau, và mỗi mùa thì sẽ là mùa thu hoạch hoặc mùa gieo gặt của từng loại nông sản thích hợp với nó). Trong khi đó, ARIMA tuy có thể phân tích và dự đoán xu hướng, song lại không thể đảm nhận công việc này.

+ Đối với mô hình Prophet: Mô hình này được xây dựng để xử lý các ngày nghỉ và sự biến động hàng năm một cách hiệu quả, giúp dự đoán có thể phản ánh đúng sự biến động theo chu kỳ thời gian ngắn hạn và dài hạn. Ngoài ra, nó cung cấp giao diện đơn giản, dễ sử dụng và có khả năng cấu hình linh hoạt, làm cho nó trở thành công cụ hữu ích cho người mới vào lĩnh vực dự báo thời gian.
  → Bên cạnh mô hình Prophet, về cơ bản cũng có nhiều mô hình có thể dự đoán được với lượng nhiều hơn rất nhiều. Song nếu xem xét về quy mô của bảng dữ liệu gồm có 12 bảng dữ liệu, mỗi bảng gồm có khoảng hơn 60000 dữ liệu với 11 cột là 11 tiêu chí như đã trình bày ở trên, việc sử dụng Prophet phù hợp về mặt chi phí, thời gian cũng như giải thích cho việc áp dụng mô hình Long Short-Term Memory (LSTM) với các mô hình Neural Networks khác.
  → Mô hình Linear Regression chưa phù hợp trong tình huống này vì lượng dữ liệu mà chúng tôi cần phải phân tích khá lớn (như chúng tôi đã đề cập bên trên), song song với đó, việc sử dụng mô hình tuyến tính đều cần lập các công thức tuyến tính (tức là có sự liên hệ và lập được công thức biểu thị sự liên hệ đó). Điều này là chưa thể làm được với dữ liệu mà chúng tôi cần phân tích khi không có bất kỳ sự tăng hay giảm liên tục nào diễn ra đối với từng sản phẩm hoặc đối với bất cứ yếu tố nào có ở trong tập dữ liệu.

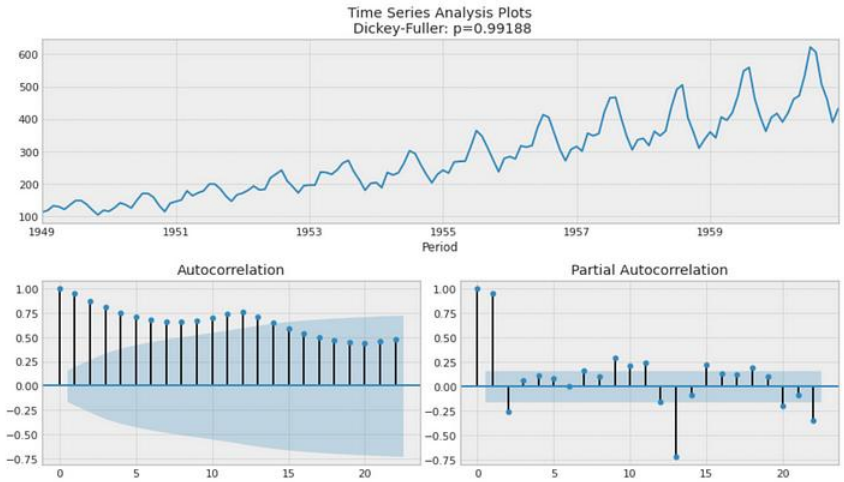
- **Vì sao nên áp dụng mô hình SARIMA và Prophet kết hợp?**
  → *Bài toán phân tích dữ liệu ngành hàng không*

+ **Mục tiêu:** Phân tích tổng số hành khách hàng không quốc tế hàng tháng (đơn vị nghìn người) từ năm 1949 - 1960 và dự báo tổng số hành khách tương lai.

SARIMAX Results

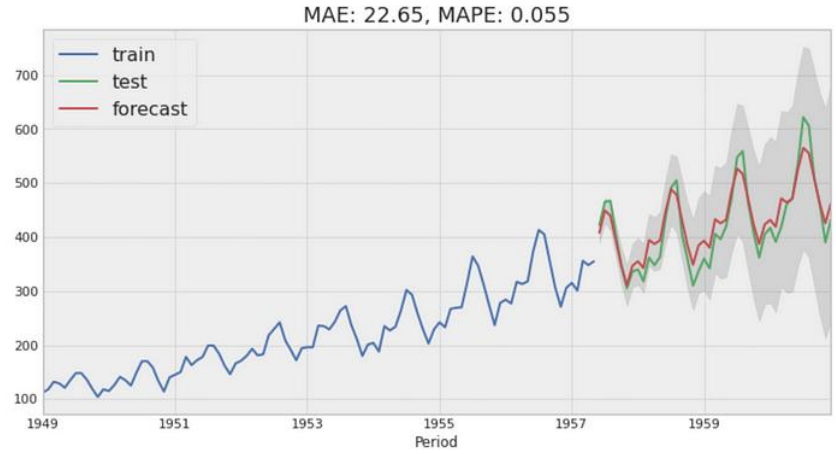
Dep. Variable:	y	**No. Observations:**	101
Model:	SARIMAX(1, 1, 0)x(1, 1, 0, 12)	**Log Likelihood**	-322.277
Date:	Sun, 04 Jul 2021	**AIC**	650.553
Time:	03:22:02	**BIC**	657.985
Sample:	0	**HQIC**	653.548
	- 101		
Covariance Type:	opg		

Hình 7: Các dữ liệu input của bài toán

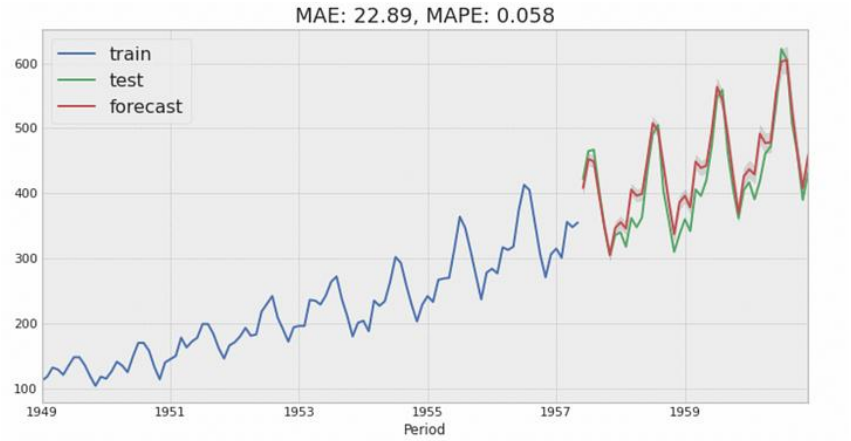


Hình 8: Tập hợp dữ liệu sau khi được phân tích ra đồ thị

+ **Kết quả:**



Hình 9: Kết quả dự đoán xu hướng sau khi áp dụng mô hình SARIMA

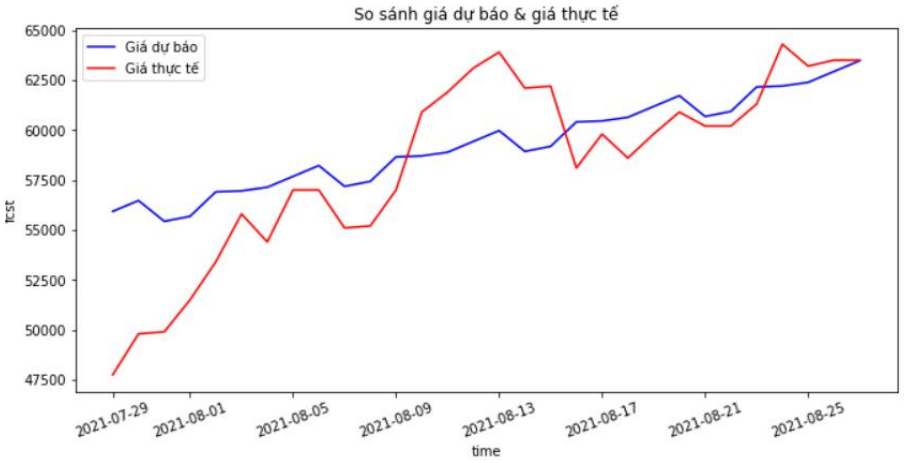


Hình 10: Kết quả dự đoán xu hướng sau khi áp dụng mô hình Prophet

+ **Nhận xét:**
  → Trong một ngữ cảnh và điều kiện cụ thể, việc sử dụng cả hai mô hình cùng một lúc mang lại một ưu thế quan trọng, giảm bớt nhược điểm của mỗi mô hình đơn lẻ. Điều này có thể dẫn đến việc tăng cường độ chính xác của dự đoán.
  → Điều này đặt ra cơ hội kết hợp sức mạnh và ưu điểm của cả hai mô hình, tận dụng khả năng của mỗi mô hình trong các khía cạnh cụ thể của dữ liệu hoặc môi trường dự báo. Sự kết hợp này có thể giúp tạo ra một hệ thống dự đoán mạnh mẽ và linh hoạt hơn, giảm thiểu rủi ro và cung cấp một cái nhìn toàn diện hơn về dự báo trong mọi tình huống.

- **Kết luận về mô hình:**
  Việc kết hợp cả hai mô hình này mang lại lợi ích ở mức tương đối cho việc khai thác và phân tích dữ liệu với nhiều mốc thời gian và có sự chuyển biến bất thường theo từng giai đoạn cũng như từng yếu tố khách quan, chủ quan của người tiêu dùng. Sự kết hợp này về mặt lý thuyết sẽ giúp:
  → *Xử Lý Cả Mùa Vụ và Sự Biến Động Không Đồng Nhất:*
  Kết hợp SARIMA và Prophet giúp xử lý cả yếu tố mùa vụ và sự biến động không đồng nhất trong dữ liệu. SARIMA chủ yếu chịu trách nhiệm về yếu tố mùa vụ, trong khi Prophet có thể xử lý sự biến động không đồng nhất và các yếu tố đặc biệt.

Hình 11: Dự đoán giá cổ phiếu bằng mô hình SARIMA



Hình 12: Dự đoán giá cổ phiếu bằng mô hình Prophet

→ *Tăng Cường Độ Chính Xác:* Khi kết hợp, cả hai mô hình có thể tăng cường độ chính xác của dự đoán, đặc biệt là trong những tình huống mà mỗi mô hình một mình có thể không đủ mạnh mẽ. ARIMA không đặt ra giả định về sự tuần hoàn hàng năm, nên nó có thể phù hợp với nhiều loại dữ liệu khác nhau; trong khi Prophet giả định rằng chuỗi thời gian có chu kỳ tuần hoàn hàng năm, điều này có thể không phù hợp cho tất cả các loại dữ liệu.

→ *Sự Linh Hoạt và Thí Nghiệm:* Việc kết hợp cả hai mô hình mang lại sự linh hoạt để thử nghiệm và tinh chỉnh mô hình dựa trên kết quả thực tế. Bạn có thể điều chỉnh trọng số của mỗi mô hình để xem cách chúng tương tác và đóng góp vào dự đoán cuối cùng. SARIMA có thể điều chỉnh và tối ưu hóa được thông qua việc điều chỉnh các biến số p, q, d; trong khi Prophet khi dùng các tập dữ liệu lớn sẽ tốn rất nhiều thời gian và tài nguyên tính toán.

3. **Phân tích hàng tồn kho**
    a. **Các dữ liệu về hàng tồn kho:**
        - plant: the ID of plant/ stock
        - calendar_year: year
        - calendar_year_week: week in year
        - sloc: site location
        - quantity: represents the overall inventory quantity of products
        - total_amount: a total of monetary values
        - product_id: represents the product ID

    b. **Dự đoán ban đầu:**
    Lượng hàng tồn kho chịu ảnh hưởng từ nhiều yếu tố, bao gồm:
        - Vị trí cửa hàng (Site location): Vị trí cửa hàng có ảnh hưởng đến lượng tồn kho, đặc biệt là do yếu tố thời tiết và khả năng lưu trữ hàng hóa. Các khu vực có điều kiện thời tiết thuận lợi hơn có thể lưu trữ nhiều hàng hóa hơn, và kích thước của khu vực kho cũng quyết định khả năng chứa lượng lớn hàng hóa.
        - Năm và Tuần trong năm (Year và Year_week): Thời gian trong năm ảnh hưởng đến doanh số bán hàng và, từ đó, ảnh hưởng đến lượng hàng tồn kho của cửa hàng.

    c. **Phương pháp kiểm chứng:**
        - **Kiến thức nền tảng:** Sử dụng mô hình SARIMA để xử lý lượng hàng tồn kho, quản lý và tối ưu hóa chuỗi cung ứng
        - **Lý do lựa chọn mô hình:**
        Trong việc phân tích lượng hàng tồn kho, SARIMA và ETS là hai mô hình được sử dụng phổ biến nhất, tuy nhiên, chúng tôi lựa chọn mô hình SARIMA vì nhiều lý do:
            + Xử Lý Yếu Tố Mùa Vụ: SARIMA được thiết kế để xử lý các yếu tố mùa vụ, điều này quan trọng khi làm việc với dữ liệu hàng tồn kho, nơi có thể xuất hiện các yếu tố chu kỳ theo thời gian. Yếu tố mùa vụ có tác động lớn đối với hàng tồn kho và có thể ảnh hưởng đến nhiều khía cạnh trong quản lý hàng tồn kho. Dưới đây là một số tác động quan trọng của yếu tố mùa vụ (chỉ xem xét đối với sản phẩm về thời trang mà chúng tôi cần phân tích).

→ Biến Động Theo Chu Kỳ Thời Gian: Yếu tố mùa vụ tạo ra sự biến động theo chu kỳ trong lượng tồn kho. Các mùa vụ như lễ hội, mùa mua sắm, hay các sự kiện đặc biệt có thể tăng cường nhu cầu và dẫn đến sự thay đổi trong lượng tồn kho.
→ Dự Đoán Nhu Cầu: Hiểu rõ yếu tố mùa vụ giúp doanh nghiệp dự đoán nhu cầu của sản phẩm trong các giai đoạn cụ thể trong năm. Điều này làm cho quản lý hàng tồn kho trở nên hiệu quả hơn, tránh tình trạng thiếu hụt hoặc dư thừa.
→ Quản Lý Sự Kiện Đặc Biệt: Những sự kiện đặc biệt như giảm giá, khuyến mãi, hay các chương trình quảng cáo thường được lên kế hoạch theo chu kỳ mùa vụ để tối ưu hóa hiệu suất. Điều này có thể tạo ra những thay đổi lớn trong lượng tồn kho.
→ Điều Chỉnh Nguồn Cung: Do sự biến động của yếu tố mùa vụ, việc điều chỉnh nguồn cung là quan trọng. Các nhà sản xuất và nhà cung cấp cần lên kế hoạch để đáp ứng nhu cầu tăng cao trong các giai đoạn đặc biệt.
→ Chiến Lược Giá: Yếu tố mùa vụ có thể ảnh hưởng đến chiến lược giá của doanh nghiệp. Việc điều chỉnh giá dựa trên mùa vụ có thể tạo ra sự khuyến khích mua sắm hoặc giảm thiểu rủi ro tồn kho.
→ Quản Lý Rủi Ro: Sự biến động theo mùa vụ có thể tạo ra rủi ro cho doanh nghiệp, đặc biệt là nếu không thể dự đoán chính xác và quản lý tồn kho một cách linh hoạt.
Đây là một ưu điểm của SARIMA so với ETS vì dữ liệu của chúng ta có theo những tiêu chí về dòng thời gian (cụ thể là năm hay tháng), trong khi ETS không thể hoạt động tốt với những dữ liệu mang tính mùa vụ và có khả năng ứng dụng những thuật toán phức tạp kém hơn. Trong khi đó, với dữ liệu mà nhóm phải phân tích - doanh số bán hàng trong giai đoạn năm và tháng, tức là chúng ta cũng cần phải xét đến các yếu tố về ngày lễ, mùa mua sắm hoặc ít mua sắm.

+ Linh Hoạt với Biến Động Thời Gian: SARIMA cung cấp sự linh hoạt để ứng phó với biến động thời gian

và biến động trong xu hướng, giúp dự báo số lượng tồn kho dựa trên thời gian.

+ Điều Chỉnh Tham Số: SARIMA cho phép điều chỉnh các tham số như p, d, q để tối ưu hóa hiệu suất của mô hình, làm cho nó linh hoạt và có thể áp dụng cho nhiều loại dữ liệu khác nhau. Bên cạnh đó, mô hình SARIMA cho phép điều chỉnh thông số về xu hướng và chu kỳ, từ đó giúp xử lý dữ liệu chúng tôi cần một cách hiệu quả hơn, đảm bảo tính chính xác cao hơn.
+ Ứng Dụng Rộng Rãi và có tính ổn định cũng như độ tin cậy cao: SARIMA đã được ứng dụng rộng rãi trong các ngành công nghiệp và có nhiều bằng chứng về hiệu suất tích cực của nó trong dự đoán và quản lý tồn kho. Trong một số bài toán, việc sử dụng SARIMA cho ra hiệu suất MAE và MAPE thấp hơn so với một số thuật toán còn lại.
+ Tương Thích với Các Biến Động Không Đều: SARIMA có khả năng xử lý biến động không đều trong dữ liệu, điều này thường xảy ra trong quản lý hàng tồn kho khi có các yếu tố bất ngờ (ví dụ như yếu tố về đại dịch Covid, do một số dữ liệu trong các dữ liệu tồn kho có dữ liệu của năm 2021).

- **Kết luận:**

  Tóm lại, việc sử dụng SARIMA thay vì ETS hoặc ứng dụng cả hai giúp cho việc phân tích trở nên đơn giản hơn, tiết kiệm chi phí hơn, tuy nhiên vẫn đảm bảo được xử lý các dữ liệu mà chúng ta cần phải phân tích. Để rõ ràng hơn nữa, chúng tôi xin phép liệt kê những ưu điểm của SARIMA so với ETS trong bài toán phân tích hàng tồn kho mà chúng tôi nhận thấy sau quá trình tìm hiểu chuyên sâu về hai thuật toán này:

  + Chuỗi Thời Gian Có Mùa Vụ Rõ Ràng: Dữ liệu chúng tôi xử lý như đã đề cập ở trên theo giai đoạn 2 năm, mỗi năm có 12 tháng, và còn được phân chia theo các mùa, điều này chứng tỏ, dữ liệu có xu hướng thay đổi theo mùa vụ, SARIMA có khả năng mô hình hóa mối quan hệ này một cách hiệu quả hơn. SARIMA có khả năng ổn định và dự đoán các biến động theo chu kỳ thời gian, điều này là đặc điểm quan trọng khi đối mặt với dữ liệu hàng tồn kho.

+ Xử Lý Xu Hướng và Chu Kỳ Phức Tạp: SARIMA cho phép bạn điều chỉnh mô hình để phản ánh xu hướng và chu kỳ phức tạp trong dữ liệu. Điều này làm cho SARIMA trở thành một công cụ linh hoạt cho việc xử lý các tình huống nâng cao và phức tạp.
+ Khả Năng Dự Đoán Tốt Trong Các Điều Kiện Thay Đổi: SARIMA thường cho hiệu suất tốt trong việc dự đoán trong các điều kiện thay đổi, giúp dự báo lượng tồn kho một cách chính xác trong các tình huống đa dạng.
+ Ứng Dụng Rộng Rãi Trong Phân Tích Chuỗi Thời Gian: SARIMA là một trong những mô hình phổ biến được sử dụng trong phân tích chuỗi thời gian, và có nhiều tài liệu, nguồn thông tin và công cụ hỗ trợ cho việc triển khai và tối ưu hóa SARIMA.

4. **Phân tích mối liên hệ**
    a. **Dữ liệu phân tích:**
        - Doanh số và lượng hàng tồn kho được cung cấp
        - Doanh số dự đoán và lượng hàng tồn kho trong tương lai
    b. **Mục tiêu phân tích:**
    Bằng việc phân tích mối liên hệ giữa doanh số bán hàng và lượng hàng tồn kho, chúng tôi hướng đến việc xây dựng chiến lược quản lý, hoạt động một cách hiệu quả. Từ việc phân tích này, chúng tôi kỳ vọng có thể:
        - Dự Đoán Tương Lai: Xác định mối liên hệ giữa doanh số và lượng hàng tồn kho để có khả năng dự đoán tương lai. Mục tiêu có thể là xây dựng mô hình dự đoán có thể giúp dự báo lượng tồn kho dự kiến dựa trên doanh số bán hàng.
        - Quản Lý Tồn Kho: Hiểu rõ mối liên hệ giữa doanh số và lượng hàng tồn kho để quản lý tồn kho một cách hiệu quả. Mục tiêu là tối ưu hóa mức tồn kho, giảm rủi ro thiếu hàng hoặc hàng tồn kho dư thừa.
        - Định Hình Chiến Lược Kinh Doanh: Sử dụng thông tin từ mối liên hệ để định hình chiến lược kinh doanh. Điều này có thể bao gồm việc xác định các kênh phân phối hiệu quả, kế hoạch tiếp thị, hay quyết định về giá cả dựa trên ảnh hưởng của chúng đối với hàng tồn kho.
        - Hiểu Rõ Tác Động Của Yếu Tố Ngoại Vi: Phân tích để hiểu rõ tác động của các yếu tố ngoại vi như thời tiết, mùa vụ, vị trí địa lý đến mối liên hệ giữa doanh số và tồn kho. Mục tiêu là phát hiện các yếu tố ảnh hưởng và điều chỉnh chiến lược diễn đàn của doanh nghiệp.
        - Tối Ưu Hóa Chiến Lược Kinh Doanh: Dựa trên mối liên hệ, xác định cách tối ưu hóa chiến lược kinh doanh nhằm tăng cường hiệu suất bán hàng và quản lý hàng tồn kho. Mục tiêu là đạt được sự cân bằng giữa cung và cầu một cách hiệu quả.

5. Xây dựng kế hoạch kinh doanh

a. Dữ liệu cần sử dụng:

- Sử dụng kết quả phân tích doanh số bán hàng để xây dựng kế hoạch kinh doanh. Các thông tin về xu hướng bán hàng, yếu tố ảnh hưởng, và các chiến lược hiệu quả có thể được tích hợp vào kế hoạch.
- Sử dụng kết quả phân tích hàng tồn kho để tối ưu hóa lượng cung ứng. Đề xuất giải pháp để xây dựng mô hình quản lý hàng hóa tốt hơn, đồng thời điều chỉnh kế hoạch kinh doanh theo lượng tồn kho dự kiến.

b. Phương pháp kiểm chứng:

- **Phương pháp thực hiện:** Sử dụng mô hình Gradient Boosting Machine

  Gradient Boosting là một phương pháp máy học được sử dụng để xây dựng mô hình dự đoán, đặc biệt là trong bài toán hồi quy và phân loại. Phương pháp này là một phần của họ các mô hình ensemble, trong đó nhiều mô hình yếu được kết hợp để tạo ra một mô hình mạnh hơn.

  Cụ thể, Gradient Boosting tập trung vào việc xây dựng một loạt các cây quyết định (decision trees) theo cách tuần tự. Mỗi cây được xây dựng để sửa lỗi của cây trước đó trong chuỗi. Quy trình này được thực hiện bằng cách sử dụng đạo hàm của hàm mất mát (loss function) đối với dự đoán hiện tại và thêm cây mới để giảm độ lỗi.

  Thuật toán Gradient Boosting phổ biến nhất là Gradient Boosted Decision Trees (GBDT), còn được gọi là Gradient Boosting Machines (GBM). Xác định các siêu tham số chính như learning rate, số lượng cây, độ sâu cây, và các tham số liên quan khác là quan trọng để điều chỉnh hiệu suất của mô hình.

- **Lý do lựa chọn:**

  Ưu điểm của GBM:

  Tuy GBM có những nhược điểm riêng của nó, cụ thể nó có thể dễ bị overfitting nếu không điều chỉnh các siêu tham số một cách chặt chẽ, cũng có thể nhạy cảm với nhiễu trong dữ liệu đào tạo; song, những ưu điểm của nó phù hợp với dữ liệu mà chúng tôi cần phân tích.

+ Dữ Liệu Thời Gian: GBM thích hợp cho dữ liệu thời gian, đặc biệt là khi xu hướng có thể thay đổi theo thời gian. Việc có các file dữ liệu "TT T<month> -< year>.xlsx" và "31-10-2022_Ton Kho 1161 - 1170.xlsx" cho thấy xu hướng có thể biến động theo cả tháng và năm.
+ Khả Năng Xử Lý Nhiễu và Phức Tạp: Dữ liệu có nhiễu và độ phức tạp cao (như là thông tin về nhiều cửa hàng, kênh phân phối, và sản phẩm), GBM có khả năng xử lý tốt hơn so với mô hình đơn giản như Decision Tree. Các biến động về giá, tồn kho, và doanh số bán hàng có thể được học tốt bởi GBM.
+ Yêu Cầu Dự Đoán Chính Xác: Nếu mục tiêu là dự đoán chính xác xu hướng của mặt hàng, GBM có khả năng cung cấp dự đoán chính xác hơn do khả năng học từ sai số của mô hình trước đó.
+ Khả Năng Linh Hoạt và Đa Dạng: GBM thường sử dụng các cây quyết định nhỏ và yếu, giúp tạo ra một mô hình linh hoạt và không dễ bị overfitting. Điều này có ý nghĩa trong việc đảm bảo mô hình có thể tổng quát hóa tốt trên dữ liệu mới.
+ Tính Tương Tác giữa Biến: GBM có thể hiệu quả khi cần xem xét sự tương tác phức tạp giữa các biến, đặc biệt là khi có nhiều thông tin từ các nguồn khác nhau như giá vốn, giá bán, và thông tin sản phẩm.

So sánh GBM với các mô hình khác:

Gradient Boosting Machine (GBM)	**Decision Tree**	**K-means**
***1.Khả năng Học Tốt*:** GBM có khả năng học từ sai số của mô hình trước đó, giúp cải thiện hiệu suất dự đoán theo thời gian. Điều này	***1. Dễ Hiểu và Diễn Giải:*** Cây quyết định thường dễ hiểu và diễn giải, có thể rõ ràng hiểu cách mô hình đưa ra quyết định.	***1. Phân Nhóm Dữ Liệu:*** K-means thích hợp khi bạn muốn phân nhóm các quan sát thành các cụm (clusters) dựa trên sự tương

làm cho GBM thích hợp khi xu hướng có thể thay đổi theo thời gian và yêu cầu một mô hình linh hoạt. ***2. Khả năng Xử lý Nhiễu*:** GBM có khả năng xử lý nhiễu và các biến giả mạo trong dữ liệu, giúp tạo ra mô hình tổng quát và không dễ bị overfitting. ***3. Hiệu Suất Cao:*** GBM thường đạt được hiệu suất cao trên các tập dữ liệu lớn và phức tạp, làm cho nó là lựa chọn phổ biến trong nhiều ứng dụng. ***4. Đa dạng Hóa Cây Quyết Định:*** GBM thường sử dụng các cây quyết định nhỏ và yếu, giúp tránh được vấn đề overfitting và tăng	***2. Phù Hợp cho Dữ Liệu Đơn Giản:*** Khi dữ liệu đơn giản và có cấu trúc rõ ràng, một cây quyết định có thể là lựa chọn tốt vì nó ít phức tạp hơn và dễ triển khai.	đồng giữa chúng. ***2. Tìm Ra Các Nhóm Tự Nhiên:*** Nếu có xu hướng tự nhiên của dữ liệu tập trung vào các cụm cụ thể, K-means có thể là một phương pháp hiệu quả. ***3. Không Yêu Cầu Chuẩn Bị Dữ Liệu Cao:*** K-means không đòi hỏi chuẩn bị dữ liệu cao như GBM, và nó có thể được áp dụng trực tiếp vào dữ liệu.

tính đa dạng của mô hình.		

Kết luận: Khi dự đoán xu hướng, GBM thường được ưa chuộng do khả năng học tốt và khả năng xử lý dữ liệu phức tạp.

- **Nguyên tắc hoạt động:**
  + *Cây Quyết Định (Decision Trees):* Gradient Boosting chủ yếu sử dụng cây quyết định như là mô hình cơ bản (weak learner). Mỗi cây được xây dựng dựa trên các quy tắc quyết định để phân loại hoặc dự đoán giá trị mục tiêu.
  + *Hàm Mất Mát (Loss Function):* Thuật toán cố gắng tối thiểu hóa hàm mất mát, đại diện cho sự sai lệch giữa giá trị dự đoán và giá trị thực tế. Các hàm mất mát khác nhau được chọn tùy thuộc vào bài toán (hồi quy hoặc phân loại).
  + *Gradient Descent:* Trong mỗi bước, một cây mới được thêm vào chuỗi để giảm độ lỗi của mô hình hiện tại. Đạo hàm của hàm mất mát được sử dụng để xác định hướng và độ lớn của cập nhật.

- **Tham số chính:**
  + *Learning Rate:* Quyết định mức độ cập nhật của các trọng số trong quá trình học. Một learning rate thấp có thể đưa đến việc học chậm, nhưng ổn định hơn.
  + *Số Lượng Cây (n_estimators):* Xác định số lượng cây quyết định sẽ được xây dựng trong chuỗi.
  + *Độ Sâu Cây (max_depth):* Điều chỉnh độ sâu tối đa của mỗi cây quyết định, ảnh hưởng đến độ phức tạp của mô hình.
  + *Tham Số Cây (min_samples_split, min_samples_leaf):* Điều chỉnh số lượng mẫu yêu cầu để một nút được chia hoặc để một lá được tạo ra, ảnh hưởng đến quá trình chia cây.

- **Áp dụng mô hình GBM vào dữ liệu của chúng tôi:**
  - *Bước 1:* Chuẩn bị Dữ liệu

    - Import thư viện cần thiết và đọc dữ liệu từ các tệp "Sales Data", "Inventory Data", và "Master Data".
    - Ghép các bảng dữ liệu dựa trên các khóa chung như product_id, branch_id, và channel_id.
- *Bước 2:* Feature Engineering
    - Tạo các biến mới từ dữ liệu như tổng doanh số bán hàng, tồn kho cuối kỳ, giá trị tồn kho, và các biến phân loại từ "Master Data".
- *Bước 3:* Chia Dữ liệu
    - Tách dữ liệu thành bộ huấn luyện và bộ kiểm tra để đánh giá hiệu suất của mô hình.
- *Bước 4:* Xây dựng Mô hình Ban Đầu
    - Sử dụng một cây quyết định đơn giản làm mô hình ban đầu để dự đoán xu hướng của mặt hàng.
- *Bước 5:* Tính Toán Độ Lỗi
    - Tính toán độ lỗi bằng cách trừ giá trị dự đoán từ thực tế. Xác định sai lệch giữa dự đoán ban đầu và thực tế.
- *Bước 6:* Xây dựng Mô hình Gradient Boosting
    - Sử dụng Gradient Boosting để xây dựng mô hình dự đoán sai lệch từ mô hình ban đầu.
    - Cấu hình các tham số như số cây quyết định, tốc độ học, và độ sâu của cây.
- *Bước 7:* Kết hợp Các Mô hình
    - Kết hợp dự đoán của mô hình ban đầu và mô hình Gradient Boosting với trọng số tùy thuộc vào hiệu suất của từng mô hình.
- *Bước 8:* Lặp lại Quá Trình
    - Lặp lại các bước 5 đến 7 nhiều lần để cải thiện dự đoán xu hướng của mặt hàng trên các điểm dữ liệu mà mô hình trước đó dự đoán sai.
- *Bước 9:* Đánh Giá Hiệu Suất
    - Đánh giá hiệu suất của mô hình trên bộ kiểm tra sử dụng các độ đo như RMSE, MSE để đảm bảo khả năng dự đoán chính xác trên dữ liệu mới.
- *Bước 10:* Tinh Chỉnh và Cải Tiến

    - Nếu cần, tinh chỉnh các tham số mô hình để đạt được hiệu suất tối ưu trên dữ liệu mới.
- *Bước 11:* Hiểu Kết Quả
    - Hiểu rõ về kết quả của mô hình và đưa ra nhận xét về xu hướng mặt hàng trong ngành bán lẻ thời trang ở Việt Nam.

Theo dõi và đánh giá

Các mô hình ML nhóm chúng tôi sử dụng là các *mô hình hồi quy*. Để đánh giá được độ chính xác của các mô hình hồi quy, nhóm chúng tôi quyết định sử dụng các chỉ số đánh giá như sau:

- **RMSE** (Root mean squared error)
- **MSE** (Mean squared error)
- **MAE** (Mean absolute error)
- **R-squared** (R2 Score)

Đối với việc ứng dụng mô hình để dự đoán đạt độ chính xác, nhóm đã thu thập số liệu từ các bài nghiên cứu liên quan đến chủ đề và rút ra được bộ chỉ số đạt độ chính xác như sau:

- **RMSE, MAE ≤ 4** (Số RMSE và MAE càng nhỏ thì độ chính xác của mô hình càng cao)
- **MSE ≤ 15** (Số MSE càng nhỏ thì độ chính xác của mô hình càng cao)
- **R-squared (R2 score) ≥ 0.7** (Số R-squared càng cao thì độ chính xác của mô hình càng cao)

Kết luận:

Ứng dụng Machine Learning vào dự đoán doanh thu và tối ưu hóa hàng tồn kho là một giải pháp tiềm năng mang lại nhiều lợi ích cho doanh nghiệp.

Với MVP đã được xây dựng, mô hình Machine Learning có thể giúp doanh nghiệp xác định xu hướng, tính mùa vụ,... để xác định nhu cầu khách hàng. Từ đó, đề xuất thời gian nhập hàng và bán hàng cũng như sản phẩm trend cụ thể, giúp tối ưu chi phí kho và tăng doanh thu. Ngoài ra, mô hình cũng có thể giúp đẩy hàng tồn kho qua đánh giá hành vi khách hàng khi công ty đưa ra giảm giá.

Với tiềm năng của mình, mô hình Machine Learning có thể giúp doanh nghiệp đạt được những lợi ích sau:

- Đón đầu được xu hướng thị trường, giúp đưa ra những kế hoạch marketing, nhập hàng và bán hàng phù hợp. Điều này có thể giúp doanh nghiệp tăng doanh số và thị phần.
- Giúp đẩy hàng tồn kho, giảm phí phạm không gian kho bãi bằng cách nhập hàng đúng thời điểm. Điều này có thể giúp doanh nghiệp giảm chi phí tồn kho và tăng lợi nhuận.

Để triển khai thành công giải pháp này, doanh nghiệp cần chú trọng đến các yếu tố sau:

- Chất lượng dữ liệu: Dữ liệu là yếu tố quan trọng nhất để xây dựng mô hình Machine Learning. Doanh nghiệp cần thu thập và xử lý dữ liệu một cách cẩn thận để đảm bảo dữ liệu chất lượng và phù hợp với mục đích sử dụng.
- Kiến thức về Machine Learning: Để xây dựng và triển khai mô hình Machine Learning, doanh nghiệp cần có kiến thức về Machine Learning. Doanh nghiệp có thể thuê các chuyên gia Machine Learning hoặc tham gia các khóa đào tạo về Machine Learning.
- Hệ thống công nghệ thông tin: Doanh nghiệp cần có hệ thống công nghệ thông tin đủ mạnh để hỗ trợ việc thu thập, xử lý và lưu trữ dữ liệu, cũng như xây dựng và triển khai mô hình Machine Learning.

Với sự phát triển của công nghệ Machine Learning, các doanh nghiệp có thể tận dụng công nghệ này để cải thiện hiệu quả kinh doanh và tăng lợi nhuận.

Tài liệu tham khảo:

[1] "Forecasting: principles and practice" by Rob J Hyndman and George Athanasopoulos

[2] "Introduction to Time Series Forecasting with Python" by Jason Brownlee

[3] "Time Series Analysis and Its Applications: With R Examples" by Robert H. Shumway and David S. Stoffer

[4] Bài báo nghiên cứu: "Forecasting Time Series with Complex Seasonal Patterns Using Exponential Smoothing" by Nikolaos Kourentzes và Fotios Petropoulos

[5] *Dự báo chuỗi thời gian nhiều bước với ARIMA, LightGBM và Prophet*. (2021, July 7). ICHI.PRO. https://ichi.pro/vi/du-bao-chuoi-thoi-gian-nhieu-buoc-voi-arima-lightgbm-va-prophet-224980043685808

[6] *Dự đoán giá cổ phiếu bằng cách sử dụng kỹ thuật học máy và học sâu (với mã Python)*. (2020, December 22). ICHI.PRO. https://ichi.pro/vi/du-doan-gia-co-phieu-bang-cach-su-dung-ky-thuat-hoc-may-va-hoc-sau-voi-ma-python-182728323756343

[7] Hieu, L. D. T. (2021, September 10). 10 Model Machine Learning phổ biến trong dự báo xu hướng giá cổ phiếu. *cafechungkhoan*. https://www.cafechungkhoan.com/2021/09/10-model-machine-learning-pho-bien.html

[8] MÔ HÌNH ARIMA VÀ DỰ BÁO LẠM PHÁT 6 THÁNG CUỐI NĂM 2014. (n.d.). *Nguyễn Khắc Hiếu*.

[9] *Tóm tắt: Cải tiến phương pháp học máy trong chuỗi thời gian và ứng dụng*. (n.d.). https://123docz.net/document/14951064-tom-tat-cai-tien-phuong-phap-hoc-may-trong-chuoi-thoi-gian-va-ung-dung.htm

[10] Vohungvi. (2023, October 25). *30 Project tuyệt vời về Machine Learning trong năm 2018*. THỊ GIÁC MÁY TÍNH. https://thigiacmaytinh.com/30-project-tuyet-voi-ve-machine-learning-trong-nam-2018/

[11] Danavg. (2018, July 18). *ABC analysis of active inventory*. Kaggle. https://www.kaggle.com/code/danavg/abc-analysis-of-active-inventory

[12] Claudiodavi. (2018, November 10). *Inventory/Sales - exploration*. Kaggle. https://www.kaggle.com/code/claudiodavi/inventory-sales-exploration

[13] Sohamjangra. (2023, July 16). *Tabular and Time Series Data Analysis using AI*. Kaggle. https://www.kaggle.com/code/sohamjangra/tabular-and-time-series-data-analysis-using-ai#Introduction

[14] Milanzdravkovic. (2020, January 7). *Pharma sales data analysis and forecasting*. Kaggle. https://www.kaggle.com/code/milanzdravkovic/pharma-sales-data-analysis-and-forecasting
[15] Farrasalyafi. (2020, June 30). *EDA + Modeling (SARIMA, PROPHET)*. Kaggle. https://www.kaggle.com/code/farrasalyafi/eda-modeling-sarima-prophet
[16] Fitriandri. (2021, February 28). *SCM in sales and inventory*. Kaggle. https://www.kaggle.com/code/fitriandri/scm-in-sales-and-inventory
[17] Rajjana. (2023, November 7). *Demand Forecasting And Inventory Optimization*. Kaggle. https://www.kaggle.com/code/rajjana/demand-forecasting-and-inventory-optimization
[18] Govindji. (2019, February 4). *Inventory management*. Kaggle. https://www.kaggle.com/code/govindji/inventory-management
[19] *ỨNG DỤNG MÔ HÌNH CHUỖI THỜI GIAN SARIMA VÀ MẠNG THẦN KINH NHÂN TẠO ANN DỰ BÁO LƯỢNG KHÁCH QUỐC TẾ ĐẾN VIỆT NAM*. (n.d.). http://thuvienso.bvu.edu.vn/bitstream/TVDHBRVT/19768/1/Nghiem-Phuc-Hieu.pdf
[20] L.M.L. (2021, September 9). *Exponential smoothing (Phần 1)*. GMO-Z.com Vietnam Lab Center Technology Blog. https://blog.vietnamlab.vn/exponential-smoothing/
[21] Hyndman, R., Koehler, A., Ord, K., & Snyder, R. D. (2008). Forecasting with Exponential Smoothing. In *Springer series in statistics*. https://doi.org/10.1007/978-3-540-71918-2
[22] Billah, B., King, M. L., Snyder, R. D., & Koehler, A. B. (2006). Exponential smoothing model selection for forecasting. *International Journal of Forecasting*, *22*(2), 239–247. https://doi.org/10.1016/j.ijforecast.2005.08.002
[23] SO SÁNH THUẬT TOÁN TĂNG CƯỜNG ĐỘ DỐC (XGBOOST) VỚI MỘT SỐ THUẬT TOÁN HỌC MÁY KHÁC -Trần Quý Nam
[24] Greedy Function Approximation: A Gradient Boosting Machine - Jerome H. Friedman*- February 24, 1999
[25] Wikipedia contributors. (2023, September 26). *Gradient boosting*. Wikipedia. https://en.wikipedia.org/wiki/Gradient_boosting#:~:text=Gradient%20boosting%20is%20a%20machine,are%20typically%20simple%20decision%20trees.

[26] *Thuật toán tăng cường là gì? - Giải thích về Thuật toán tăng cường trong công nghệ máy học - AWS*. (n.d.). Amazon Web Services, Inc. https://aws.amazon.com/vi/what-is/boosting/
[27] Ong H. (2019, April 8). *XGBoost: thuật toán giành chiến thắng tại nhiều cuộc thi Kaggle*. Ông Xuân Hồng. https://ongxuanhong.wordpress.com/2017/12/21/xgboost-thuat-toan-gianh-chien-thang-tai-nhieu-cuoc-thi-kaggle/