**NhómBỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

✩✩✩✩✩



**MÔN HỌC: KHAI PHÁ DỮ LIỆU**

**BÁO CÁO CUỐI KỲ**

**ĐỀ TÀI: MINING E-COMMERCE BUSINESS TRANSACTION DATA**

**GVHD: Ths. Trần Trọng Bình**

**SVTH:**   **MSSV**

Đỗ Minh Trọng 21133111

Phan Công Danh 21133014

Nguyễn Duy Hải 21133101

Trần Thị Ngọc Trang 21133109

**Lớp thứ 7\_Tiết 11-14**

**Tp.Hồ Chí Minh, tháng 5 năm 2024**

**ĐÁNH GIÁ & CHẤM ĐIỂM**

| **STT** | **THÀNH VIÊN** | **ĐÁNH GIÁ** | **ĐIỂM** | **KÝ TÊN** |
| --- | --- | --- | --- | --- |
| 1 | Đỗ Minh Trọng |  |  |  |
| 2 | Phan Công Danh |  |  |  |
| 3 | Nguyễn Duy Hải |  |  |  |
| 4 | Trần Thị Ngọc Trang |  |  |  |

**NHẬN XÉT**

*Ký tên*

**Ths. Trần Trọng Bình**

**LỜI CẢM ƠN**

*Đầu tiên, nhóm em xin chân thành gửi lời cảm ơn với Thầy – Ths. Trần Trọng Bình – Giảng viên phụ trách môn Khai Phá Dữ Liệu – Trường Đại học Sư phạm kỹ thuật TP.HCM.*

*Trong thời gian học tập, nhóm em nhận được sự nhiều sự giúp đỡ từ Thầy. Thầy đã cung cấp cho chúng em đầy đủ kiến thức, chỉ bảo và đóng góp những ý kiến quý báu giúp chúng em có thêm hiểu biết để hoàn thành được báo cáo cuối kỳ cho môn học. Thầy đã cung cấp tài liệu và hướng dẫn tận tình cho chúng em trong suốt quá trình học tập. Thầy luôn nhiệt tình vui vẻ chỉ dạy, chia sẻ giúp chúng em trong môn học. Chúng em rất trân quý sự tâm huyết và trách nhiệm của Thầy trong công việc giảng dạy và truyền đạt kiến thức.*

*Trong quá trình thực hiện đồ án, dựa trên kiến thức được Thầy cung cấp qua các buổi học lý thuyết cũng như thực hành trên lớp, kết hợp với việc tự tìm hiểu những công cụ và kiến thức mới, nhóm đã cố gắng thực hiện đồ án một cách tốt nhất .Tuy nhiên, đồ án còn chưa được hoàn hiện và có thể còn có sai sót.*

*Nhóm rất mong nhận được sự góp ý từ Thầy nhằm rút ra những kinh nghiệm quý báu và hoàn thiện vốn kiến thức để nhóm có thể hoàn thành những đồ án, dự án khác trong tương lai .*

***Nhóm chúng em xin chân thành cảm ơn Thầy!***

**PHÂN CÔNG NHIỆM VỤ**

| **STT** | **THÀNH VIÊN** | **NHIỆM VỤ** | **MỨC ĐỘ HOÀN THÀNH** |
| --- | --- | --- | --- |
| 1 | Đỗ Minh Trọng | * Phân khúc khách hàng bằng RFM và Kmean | 100% |
| 2 | Phan Công Danh | Khai phá luật kết hợp (Apriori)  Khai phá luật kết hợp (FP-Growth) | 100% |
| 3 | Nguyễn Duy Hải | EDA (Phân bố giao dịch)  EDA (Hiệu suất bán hàng) | 100% |
| 4 | Trần Thị Ngọc Trang | Tiền Xử lý  Dự báo doanh thu (ARIMA, LSTM) | 100% |

**MỤC LỤC**

[**CHƯƠNG 1: LÝ DO CHỌN ĐỀ TÀI VÀ GIỚI THIỆU VỀ DATASET 1**](#_heading=h.30j0zll)

[**1.1. Lý do chọn đề tài 1**](#_heading=h.1fob9te)

[**1.2. Giới thiệu về dataset 2**](#_heading=h.3znysh7)

[1.2.1. Nguồn lấy dữ liệu 2](#_heading=h.2et92p0)

[1.2.2. Mô tả về tập dữ liệu đã chọn 2](#_heading=h.tyjcwt)

[**1.3. Tiền xử lý dữ liệu 3**](#_heading=h.1t3h5sf)

[1.3.1. Các kỹ thuật tiền xử lý dữ liệu 3](#_heading=h.4d34og8)

[1.3.2. Các bước tiền xử lý dữ liệu của nhóm trong đồ án 4](#_heading=h.2s8eyo1)

[**CHƯƠNG 2: CÁC PHƯƠNG PHÁP SỬ DỤNG 6**](#_heading=h.fjo0j8cvq45u)

[**2.1. Thuật Toán K-Mean 6**](#_heading=h.26in1rg)

[**2.2. Thuật toán Apriori 7**](#_heading=h.3rdcrjn)

[**2.3. Thuật toán FP-Growth 9**](#_heading=h.kb7j4ysss0br)

[**2.4. Thuật toán ARIMA 11**](#_heading=h.wmhxp77aei3t)

[**2.5. Thuật toán LSTM 1**](#_heading=h.oh5fr64r957j)3

[**CHƯƠNG 3: XÂY DỰNG MÔ HÌNH HUẤN LUYỆN 16**](#_heading=h.35nkun2)

[**3.1. Phân khúc khách hàng bằng mô hình RFM và K-Mean 16**](#_heading=h.1ksv4uv)

[3.1.1. RFM Data 16](#_heading=h.y9bnpmvh4k4h)

[3.1.2. Xử lý Outlier 17](#_heading=h.h08daje9oeim)

[3.1.3. Scale 19](#_heading=h.qgz94ngpc4s5)

[3.1.4. PCA 20](#_heading=h.mqgyuige2ofk)

[3.1.5. Phân cụm với K-Mean 21](#_heading=h.ms9c0j8htm57)

[**3.2. Mô hình khai phá luật kết hợp sử dụng thuật toán Apriori và FP-Growth 30**](#_heading=h.44sinio)

[3.2.1. Sử dụng Thuật toán Apriori 30](#_heading=h.cmbrpp7l1lzk)

[3.2.2. Sử dụng thuật toán FP-Growth 33](#_heading=h.nw8on8q1akgu)

[**3.3. Dự báo doanh thu sử dụng LSTM và ARIMA 35**](#_heading=h.93ag1btef3xp)

[3.3.1. Tùy chỉnh lại bộ dữ liệu 35](#_heading=h.e38eeypduxcu)

[3.3.2. LSTM 36](#_heading=h.ewjnyonb6yg9)

[3.3.3. ARIMA 40](#_heading=h.o42bkarxf8qk)

[3.3.4. So sánh hiệu suất mô hình 48](#_heading=h.oax97k6msh7t)

[**KẾT LUẬN 50**](#_heading=h.l57lnd7p85ig)

[**TÀI LIỆU THAM KHẢO 51**](#_heading=h.4i7ojhp)

# CHƯƠNG 1: LÝ DO CHỌN ĐỀ TÀI VÀ GIỚI THIỆU VỀ DATASET

## Lý do chọn đề tài

Thương mại điện tử đã trở thành một kênh mới hỗ trợ doanh nghiệp phát triển. Thông qua thương mại điện tử, doanh nghiệp có thể tiếp cận và thiết lập sự hiện diện trên thị trường rộng hơn bằng cách cung cấp các kênh phân phối rẻ hơn và hiệu quả hơn cho sản phẩm hoặc dịch vụ của họ. Thương mại điện tử cũng đã thay đổi cách mọi người mua sắm và tiêu thụ sản phẩm và dịch vụ. Nhiều người đang chuyển sang sử dụng máy tính hoặc thiết bị thông minh để đặt hàng, những mặt hàng có thể dễ dàng được giao đến tận nhà.

Thông tin là tài sản quan trọng nhất của doanh nghiệp hiện nay. Sự thành công của một doanh nghiệp trong môi trường cạnh tranh phụ thuộc vào khả năng thu thập, lưu trữ và sử dụng thông tin. Dữ liệu là một trong những nguồn thông tin chính. Vì vậy nhóm chúng em quyết định sử dụng kiến thức đã được học để khai phá dữ liệu về giao dịch bán hàng của thương mại điện tử. Và cụ thể là dùng thuật toán K-Mean và luật kết hợp (Apriori, FP-Growth). Ngoài ra, nhóm còn sử dụng thuật toán ARIMA và LSTM để có thể xác định việc vận hành có hiệu quả hay không. Việc dự đoán doanh thu trong tương lai có thể giúp doanh nghiệp xem xét các yếu tố để có thể so sánh nó với thực tế, đánh giá hiệu quả của các chiến dịch trong tương lai.

## Giới thiệu về dataset

### Nguồn lấy dữ liệu

Nhóm lấy nguồn dữ liệu từ kaggle: [E-commerce Business Transaction](https://www.kaggle.com/datasets/gabrielramos87/an-online-shop-business)

### Mô tả về tập dữ liệu đã chọn

E-commerce Business Transaction: Tập dữ liệu là giao dịch bán hàng của một thương mại điện tử (bán lẻ trực tuyến) có trụ sở tại Vương quốc Anh trong một năm.

Đây là bộ dữ liệu giao dịch bán hàng của thương mại điện tử (bán lẻ trực tuyến) có trụ sở tại Vương quốc Anh trong một năm. Cửa hàng có trụ sở tại London này đã bán quà tặng và đồ gia dụng cho người lớn và trẻ em thông qua trang web từ năm 2007. Khách hàng của họ đến từ khắp nơi trên thế giới và thường mua hàng trực tiếp cho chính họ. Ngoài ra còn có các doanh nghiệp nhỏ mua số lượng lớn và bán cho khách hàng khác thông qua kênh cửa hàng bán lẻ.

Kích thước: Bộ dữ liệu này bao gồm 536,350 dòng (điểm dữ liệu) và 8 thuộc tính.

Tập dữ liệu bao gồm:

| **Tên cột** | **Mô tả** |  |
| --- | --- | --- |
| Date (Ngày) | ngày mà mỗi giao dịch được tạo | object |
| Price (Giá) | giá của mỗi sản phẩm trên một đơn vị tính bằng bảng Anh (£) | float |
| Quantity (Số lượng) | số lượng mỗi sản phẩm trên một giao dịch. Giá trị âm liên quan đến giao dịch bị hủy | int |
| TransactionNo (Mã giao dịch) | một số duy nhất gồm sáu chữ số xác định mỗi giao dịch. Chữ “C” trong mã biểu thị việc hủy bỏ | object |
| ProductNo (Mã sản phẩm) | ký tự duy nhất gồm năm hoặc sáu chữ số được sử dụng để xác định một sản phẩm cụ thể. | object |
| Product (Sản phẩm) | tên sản phẩm/mặt hàng | object |
| CustomerNo (Mã khách hàng) | một số duy nhất có năm chữ số xác định từng khách hàng | float |
| Country (Quốc gia) | tên quốc gia nơi khách hàng cư trú | object |

## Tiền xử lý dữ liệu

### Các kỹ thuật tiền xử lý dữ liệu

Các kỹ thuật tiền xử lý dữ liệu là các bước chuẩn bị dữ liệu trước khi áp dụng các phương pháp khai phá dữ liệu. Chúng là những bước quan trọng để đảm bảo chất lượng của dữ liệu và tăng hiệu quả của quá trình khai phá. Dưới đây là một số kỹ thuật tiền xử lý dữ liệu phổ biến

* *Xử lý giá trị khuyết*: Điền giá trị cho các trường dữ liệu bị thiếu hoặc loại bỏ các bản ghi có giá trị khuyết.
* *Xử lý ngoại lệ*: Loại bỏ các bản ghi ngoại lệ hoặc điều chỉnh các giá trị ngoại lệ sao cho phù hợp với các giá trị còn lại.
* *Chuẩn hóa dữ liệu*: Đưa các giá trị dữ liệu về cùng một đơn vị hoặc phạm vi để tránh sự chênh lệch giữa các giá trị dữ liệu.
* *Tách và kết hợp các trường dữ liệu*: Tách các trường dữ liệu có định dạng đa giá trị thành các trường dữ liệu riêng biệt hoặc kết hợp các trường dữ liệu để tạo ra các thông tin mới.
* *Chuyển đổi định dạng dữ liệu*: Chuyển đổi định dạng dữ liệu để phù hợp với các phương pháp khai phá dữ liệu hoặc các công cụ phân tích dữ liệu khác.
* *Giảm số chiều dữ liệu*: Dùng PCA để giảm kích thước của tập dữ liệu.

Các kỹ thuật tiền xử lý dữ liệu có thể thực hiện trên một tập dữ liệu hoặc áp dụng theo chuỗi các bước để tạo ra một tập dữ liệu mới được tiền xử lý hoàn chỉnh. Các kỹ thuật này đóng vai trò quan trọng trong quá trình khai phá dữ liệu và ảnh hưởng đến kết quả cuối cùng của quá trình này.

### Các bước tiền xử lý dữ liệu của nhóm trong đồ án

**Xử lý dữ liệu bị thiếu (missing value):**

| null\_quantity = df.isnull().sum()  null\_percentages = df.isnull().mean() \* 100  duplicated\_null\_stats = pd.DataFrame({  'Số lượng giá trị null': null\_quantity,  'Tỷ lệ giá trị null (%)': null\_percentages  }).sort\_values(by='Tỷ lệ giá trị null (%)', ascending=False)  print("Thông tin về missing value:")  print ("-"\*80)  display(duplicated\_null\_stats |
| --- |

| cleaned\_df = df.dropna() |
| --- |

**Thực hiện chuẩn hóa kiểu dữ liệu**

| cleaned\_df['CustomerNo'] = cleaned\_df['CustomerNo'].astype('int').astype('object')  cleaned\_df['Date'] = pd.to\_datetime(cleaned\_df['Date']) |
| --- |

**Xử lý nhiễu**

Ở bài này chỉ tập trung vào các giao dịch thành công do đó sẽ lại bỏ những giao dịch bị hủy.

Xóa những giao dịch có Mã Giao Dịch bắt đầu bằng 'C'. Ngoài ra cũng loại bỏ các giao dịch có số lượng sản phẩm nhỏ hơn hoặc bằng 0.

| cleaned\_df = cleaned\_df[~cleaned\_df['TransactionNo'].str.startswith('C')]  cleaned\_df = cleaned\_df[df['Quantity'] > 0] |
| --- |

**Tạo và trích xuất thuộc tính**

* Tạo thêm cột Renevue được tính bằng Price \* Quantity

| cleaned\_df['Revenue'] = cleaned\_df['Price'] \* cleaned\_df['Quantity'] |
| --- |

* Tạo thêm các biến về thời gian

| cleaned\_df['Year'] = cleaned\_df['Date'].dt.year.astype('category')  cleaned\_df['Month'] = cleaned\_df['Date'].dt.month.astype('category')  cleaned\_df['Day'] = cleaned\_df['Date'].dt.day.astype('category')  cleaned\_df['DayOfWeek'] = cleaned\_df['Date'].dt.dayofweek.astype('category') |
| --- |

# 

# 

# CHƯƠNG 2: CÁC PHƯƠNG PHÁP SỬ DỤNG

## Thuật Toán K-Mean

Trong thuật toán k-Means mỗi cụm dữ liệu được đặc trưng bởi một tâm (centroid). tâm là điểm đại diện nhất cho một cụm và có giá trị bằng trung bình của toàn bộ các quan sát nằm trong cụm. Chúng ta sẽ dựa vào khoảng cách từ mỗi quan sát tới các tâm để xác định nhãn cho chúng trùng thuộc về tâm gần nhất. Ban đầu thuật toán sẽ khởi tạo ngẫu nhiên một số lượng xác định trước tâm cụm. Sau đó tiến hành xác định nhãn cho từng điểm dữ liệu và tiếp tục cập nhật lại tâm cụm. Thuật toán sẽ dừng cho tới khi toàn bộ các điểm dữ liệu được phân về đúng cụm hoặc số lượt cập nhật tâm chạm ngưỡng.

*Ý tưởng chính của thuật toán K-means*

Thuật toán K-means chia tập dữ liệu thành K cụm sao cho sự khác biệt giữa các điểm dữ liệu trong cùng một cụm là nhỏ nhất và sự khác biệt giữa các cụm là lớn nhất. Mỗi cụm được đại diện bởi một centroid (tâm cụm), là trung bình của tất cả các điểm dữ liệu trong cụm đó.

*Các bước thực hiện của thuật toán K-means:*

* Khởi tạo:Chọn K centroid ban đầu (có thể chọn ngẫu nhiên hoặc sử dụng một phương pháp khởi tạo nào đó).
* Phân cụm:Gán mỗi điểm dữ liệu vào cụm có centroid gần nhất.
* Cập nhật centroid:Tính toán lại vị trí của mỗi centroid bằng cách lấy trung bình các điểm dữ liệu trong cụm tương ứng.
* Lặp lại:Lặp lại bước phân cụm và cập nhật centroid cho đến khi các centroid không thay đổi nhiều hoặc đạt đến một số lần lặp tối đa.

*Đánh giá kết quả phân cụm*

Để đánh giá kết quả của thuật toán K-means, ta có thể sử dụng một số phương pháp như:

Elbow Method: Phương pháp này dựa trên việc tính toán tổng bình phương khoảng cách từ các điểm dữ liệu đến centroid của chúng (Within-Cluster Sum of Squares - WCSS) và tìm điểm gấp khúc (elbow) trong đồ thị WCSS theo số lượng cụm K. Điểm gấp khúc này được cho là số lượng cụm tối ưu.

Silhouette Score: Chỉ số này đo lường mức độ tương đồng của một điểm dữ liệu với cụm của nó so với các cụm khác. Silhouette Score dao động từ -1 đến 1, với giá trị gần 1 cho thấy điểm dữ liệu nằm gần trung tâm cụm của nó và xa các cụm khác.

*Ưu và nhược điểm của thuật toán K-means*

Ưu điểm: Dễ hiểu và dễ triển khai: K-means là một trong những thuật toán phân cụm đơn giản và dễ hiểu nhất. Tính toán nhanh: Với các tập dữ liệu vừa và nhỏ, K-means có thể thực thi nhanh chóng.

Nhược điểm: Nhạy cảm với giá trị khởi tạo: Kết quả của thuật toán K-means có thể khác nhau tùy thuộc vào việc chọn centroid ban đầu

Không thích hợp cho các cụm có hình dạng phức tạp: K-means hoạt động tốt với các cụm hình cầu nhưng không hiệu quả với các cụm có hình dạng phức tạp.

Đòi hỏi xác định trước số lượng cụm K: Người dùng phải xác định số lượng cụm K trước khi chạy thuật toán, điều này có thể khó khăn nếu không có thông tin về cấu trúc dữ liệu.

Thuật toán K-means được sử dụng rộng rãi trong nhiều lĩnh vực như: phân khúc khách hàng, phân tích thị trường, phát hiện bất thường, giảm chiều dữ liệu.

Thuật toán K-means là một công cụ mạnh mẽ và linh hoạt, giúp các nhà phân tích dữ liệu và nhà khoa học dữ liệu khám phá và hiểu sâu hơn về cấu trúc ẩn trong dữ liệu của họ.

## Thuật toán Apriori

Thuật toán Apriori là một thuật toán khai phá luật kết hợp (Association Rule Mining) được sử dụng để tìm kiếm các mối quan hệ kết hợp giữa các mặt hàng trong một tập dữ liệu lớn. Thuật toán này được phát triển bởi Agrawal và Srikant vào năm 1994.

Ý tưởng chính của thuật toán Apriori là sử dụng tính chất Apriori: một tập con của một tập hợp phổ biến cũng phải phổ biến. Cụ thể, thuật toán sẽ liệt kê tất cả các tập con của các tập hợp phổ biến và kiểm tra xem chúng có phổ biến không. Nếu một tập con không phổ biến, thì các tập con con của nó sẽ không phổ biến.

Thuật toán Apriori có các bước thực hiện như sau:

* Tìm tất cả các mặt hàng xuất hiện trong tập dữ liệu và xây dựng các tập hợp 1- phần tử.
* Lặp lại các bước sau cho đến khi không thể tạo ra tập hợp mới:
  + Tạo ra các tập hợp k- phần tử từ các tập hợp (k-1)- phần tử phổ biến.
  + Tính toán độ hỗ trợ của các tập hợp k- phần tử.
  + Lọc bỏ các tập hợp k- phần tử không phổ biến.

Sau khi tìm ra tất cả các tập hợp phổ biến, ta có thể sử dụng chúng để tạo ra các quy tắc kết hợp (association rules). Một quy tắc kết hợp là một tuyên bố về mối quan hệ giữa các mặt hàng, ví dụ như "nếu mua A và B, thì cũng sẽ mua C". Các quy tắc này có thể giúp các nhà bán lẻ, chủ cửa hàng hay nhà quản lý cửa hàng tối ưu hóa việc trưng bày sản phẩm, quảng cáo hay giá cả để tăng doanh số bán hàng.

Để xác định các quy tắc kết hợp từ các tập hợp phổ biến, ta cần xác định hai giá trị thống kê quan trọng: độ hỗ trợ (support) và độ tin cậy (confidence).

Độ hỗ trợ của một tập hợp là tỷ lệ giữa số lần xuất hiện của tập hợp đó trong tập dữ liệu và tổng số mẫu dữ liệu. Giá trị độ hỗ trợ cao cho thấy rằng tập hợp đó xuất hiện thường xuyên trong tập dữ liệu, và có khả năng là một phần quan trọng trong các quy tắc kết hợp.

Độ tin cậy của một quy tắc là tỷ lệ giữa số lần quy tắc xuất hiện và số lần tập hợp điều kiện (antecedent) xuất hiện. Nó cho biết mức độ tin tưởng của quy tắc kết hợp, tức là xác suất rằng tập hợp kết quả (consequent) sẽ xuất hiện nếu tập hợp điều kiện (antecedent) xuất hiện.

Từ các giá trị độ hỗ trợ và độ tin cậy, ta có thể áp dụng một số tiêu chuẩn để chọn lọc các quy tắc kết hợp có giá trị nhất. Ví dụ, ta có thể lọc các quy tắc kết hợp có độ hỗ trợ và độ tin cậy cao hơn một ngưỡng nhất định, hoặc chọn các quy tắc có giá trị Lift (tỉ lệ giữa độ hỗ trợ của quy tắc và tích độ hỗ trợ của antecedent và consequent) cao nhất.

Thuật toán Apriori là một trong những thuật toán khai phá luật kết hợp phổ biến nhất và được sử dụng trong nhiều lĩnh vực khác nhau, từ bán lẻ đến khoa học máy tính và phân tích dữ liệu. Tuy nhiên, thuật toán này cũng có nhược điểm là tốc độ chạy chậm với các tập dữ liệu lớn và đòi hỏi nhiều bộ nhớ để lưu trữ tập hợp phổ biến và các quy tắc kết hợp.

## 

## Thuật toán FP-Growth

Từ các giá trị độ hỗ trợ và độ tin cậy, ta có thể áp dụng một số tiêu chuẩn để chọn lọc các quy tắc kết hợp có giá trị nhất. Ví dụ, ta có thể lọc các quy tắc kết hợp có độ hỗ trợ và độ tin cậy cao hơn một ngưỡng nhất định, hoặc chọn các quy tắc có giá trị Lift (tỉ lệ giữa độ hỗ trợ của quy tắc và tích độ hỗ trợ của antecedent và consequent) cao nhất.

Thuật toán FP-Growth (Frequent Pattern Growth) là một thuật toán khai phá luật kết hợp (Association Rule Mining) hiệu quả, được sử dụng để tìm kiếm các tập hợp phổ biến trong một tập dữ liệu lớn mà không cần phải tạo ra các tập con giống như thuật toán Apriori. Thuật toán này được Han, Pei và Yin đề xuất vào năm 2000.

Ý tưởng chính của thuật toán FP-Growth là sử dụng cấu trúc cây gọi là FP-Tree (Frequent Pattern Tree) để lưu trữ và biểu diễn các tập hợp phổ biến. Nén cơ sở dữ liệu vào cây FP-tree, chỉ giữ lại thông tin liên kết (kết hợp) của các hạng mục (tập phổ biến). Chia cơ sở dữ liệu nén thành các cơ sở dữ liệu có điều kiện, mỗi cơ sở dữ liệu được chia ra ứng với một hạng mục phổ biến và ta sẽ khai thác các cơ sở dữ liệu này một cách độc lập.

*Các bước thực hiện của thuật toán FP-Growth*

Xây dựng FP-Tree:

* Quét cơ sở dữ liệu: Tính toán độ hỗ trợ của từng mặt hàng và lọc ra các mặt hàng không phổ biến (có độ hỗ trợ nhỏ hơn ngưỡng cho trước).
* Sắp xếp mặt hàng: Sắp xếp các mặt hàng theo thứ tự giảm dần của độ hỗ trợ.
* Xây dựng FP-Tree: Xây dựng cây FP-Tree bằng cách quét lại cơ sở dữ liệu và thêm các giao dịch vào cây theo thứ tự đã sắp xếp.

Khai phá FP-Tree:

* Tạo FP-Tree con: Tạo ra các cây FP-Tree con cho từng mặt hàng phổ biến.
* Khai phá đệ quy: Áp dụng khai phá đệ quy trên các FP-Tree con để tìm các tập hợp phổ biến.

Ưu điểm của thuật toán tăng trưởng FP Dưới đây là những ưu điểm sau của thuật toán tăng trưởng FP, chẳng hạn như: Việc ghép nối các mục không được thực hiện trong thuật toán này, khiến nó nhanh hơn. Cơ sở dữ liệu được lưu trữ dưới dạng phiên bản nhỏ gọn trong bộ nhớ. Nó hiệu quả và có thể mở rộng để khai thác cả các mẫu phổ biến dài và ngắn. Nhược điểm của thuật toán FP-Growth Thuật toán này cũng có một số nhược điểm như: Cây FP cồng kềnh và khó xây dựng hơn Apriori. Nó có thể đắt tiền. Thuật toán có thể không vừa với bộ nhớ dùng chung khi cơ sở dữ liệu lớn.

Nhược điểm thuật toán này cũng có một số nhược điểm như: FP Tree cồng kềnh và khó xây dựng hơn Apriori. Nó có thể đắt tiền. Thuật toán có thể không vừa với bộ nhớ dùng chung khi cơ sở dữ liệu lớn.

Thuật toán FP-Growth là một công cụ mạnh mẽ và hiệu quả trong việc khai phá luật kết hợp, giúp khám phá các tập hợp phổ biến trong tập dữ liệu lớn mà không gặp phải các hạn chế về hiệu suất như thuật toán Apriori. Với cấu trúc FP-Tree, FP-Growth cung cấp một phương pháp tối ưu để phân tích các mẫu dữ liệu và đưa ra các quyết định kinh doanh thông minh.

## Thuật toán ARIMA

Thuật toán ARIMA (AutoRegressive Integrated Moving Average) là một mô hình thống kê được sử dụng rộng rãi trong dự báo chuỗi thời gian. ARIMA kết hợp ba thành phần chính: tự hồi quy (AR), sai phân (I), và trung bình trượt (MA). ARIMA là một trong những công cụ mạnh mẽ nhất trong dự báo chuỗi thời gian, đặc biệt khi dữ liệu có xu hướng và mùa vụ.

Các thành phần của ARIMA

* AR (AutoRegressive): Thành phần tự hồi quy. Mô hình AR sử dụng mối quan hệ giữa giá trị hiện tại và các giá trị trong quá khứ để dự đoán giá trị tương lai. Một mô hình AR(p) sử dụng p giá trị trong quá khứ.
* I (Integrated): Thành phần sai phân. Thành phần này liên quan đến việc lấy sai phân của dữ liệu để loại bỏ xu hướng và biến đổi chuỗi thời gian thành chuỗi dừng. Một mô hình I(d) sử dụng d lần sai phân để đạt được tính dừng.
* MA (Moving Average): Thành phần trung bình trượt. Mô hình MA sử dụng lỗi dự báo của các giá trị trong quá khứ để dự đoán giá trị tương lai. Một mô hình MA(q) sử dụng q giá trị lỗi trong quá khứ.

Các bước xây dựng mô hình ARIMA

Xác định tính dừng của chuỗi thời gian:

* Một chuỗi thời gian được gọi là dừng nếu các đặc điểm thống kê của nó (như trung bình và phương sai) không thay đổi theo thời gian.
* Sử dụng các kiểm định thống kê như kiểm định Augmented Dickey-Fuller (ADF) để kiểm tra tính dừng.

Khử xu hướng và sai phân:

* Nếu chuỗi không dừng, ta sử dụng sai phân để biến nó thành chuỗi dừng. Điều này có thể bao gồm sai phân bậc một hoặc bậc hai tùy thuộc vào mức độ xu hướng.

Xác định các tham số p, d, q:

* Sử dụng các biểu đồ ACF (Autocorrelation Function) và PACF (Partial Autocorrelation Function) để xác định các giá trị tối ưu của p và q.
* Tham số d được xác định trong bước khử xu hướng.

Ước lượng các tham số của mô hình:

* Sử dụng phương pháp bình phương nhỏ nhất hoặc các phương pháp tối ưu hóa khác để ước lượng các tham số của mô hình ARIMA.

Chẩn đoán mô hình:

* Kiểm tra các phần dư của mô hình để đảm bảo chúng không tương quan và tuân theo phân phối chuẩn. Sử dụng các kiểm định như Ljung-Box để kiểm tra.

Dự báo:

* Sử dụng mô hình ARIMA đã hiệu chỉnh để dự báo các giá trị tương lai.

Ưu điểm của ARIMA

* Linh hoạt: ARIMA có thể xử lý nhiều loại chuỗi thời gian khác nhau, bao gồm chuỗi có xu hướng và chuỗi không dừng.
* Hiệu quả: Thường cho kết quả dự báo chính xác khi các tham số được chọn đúng cách.
* Được sử dụng rộng rãi: ARIMA là một trong những mô hình dự báo chuỗi thời gian phổ biến nhất trong kinh tế, tài chính và các lĩnh vực khác.

Nhược điểm:

* Phức tạp: Quá trình xác định và ước lượng các tham số mô hình có thể phức tạp và tốn thời gian.
* Không phù hợp cho dữ liệu phi tuyến tính: ARIMA hoạt động tốt với các chuỗi thời gian tuyến tính và có thể gặp khó khăn với các mẫu phi tuyến tính hoặc có mùa vụ phức tạp.
* Yêu cầu kiến thức chuyên sâu: Đòi hỏi người dùng có kiến thức sâu về thống kê và phân tích chuỗi thời gian để thiết lập và giải thích mô hình đúng cách.

Ứng dụng của ARIMA

* ARIMA được sử dụng rộng rãi trong nhiều lĩnh vực như:
* Dự báo kinh tế: Dự báo các chỉ số kinh tế như GDP, lãi suất, và lạm phát.
* Tài chính: Dự báo giá cổ phiếu, lợi nhuận và các biến số tài chính khác.
* Quản lý chuỗi cung ứng: Dự báo nhu cầu sản phẩm và quản lý tồn kho.
* Y tế: Dự báo xu hướng bệnh dịch và nhu cầu dịch vụ y tế.

Thuật toán ARIMA là một công cụ mạnh mẽ và linh hoạt cho việc dự báo chuỗi thời gian, giúp phân tích và dự đoán xu hướng trong dữ liệu quá khứ để đưa ra các dự báo chính xác. Mặc dù yêu cầu kiến thức chuyên sâu và có thể phức tạp, ARIMA vẫn là một trong những phương pháp dự báo chuỗi thời gian được sử dụng phổ biến nhất nhờ vào khả năng xử lý các chuỗi dữ liệu phức tạp và đưa ra dự báo chính xác.

## Thuật toán LSTM

LSTM (Long Short-Term Memory) là một loại mạng nơ-ron hồi quy (Recurrent Neural Network - RNN) được thiết kế để xử lý và dự báo chuỗi thời gian hoặc dữ liệu tuần tự dài hạn. LSTM khắc phục những hạn chế của RNN truyền thống bằng cách thêm các cơ chế lưu trữ và quên thông tin, giúp mô hình học được các mối quan hệ dài hạn trong dữ liệu.

LSTM được xây dựng từ các đơn vị nhớ (memory cells), mỗi đơn vị nhớ có ba thành phần chính:

* Cổng vào (Input Gate):Quyết định thông tin mới nào từ đầu vào sẽ được lưu trữ trong trạng thái của đơn vị nhớ.
* Cổng quên (Forget Gate): Quyết định thông tin nào sẽ bị loại bỏ khỏi trạng thái của đơn vị nhớ.
* Cổng đầu ra (Output Gate): Quyết định phần nào của trạng thái sẽ được sử dụng để tính toán đầu ra.

Cổng Quên (Forget Gate):

* Chức năng: Quyết định thông tin nào từ trạng thái trước đó cần được quên hoặc giữ lại.
* Hoạt động: Cổng quên sử dụng hàm kích hoạt sigmoid để tạo ra một giá trị giữa 0 và 1 cho mỗi phần của trạng thái. Giá trị gần 0 có nghĩa là thông tin sẽ bị lãng quên, trong khi giá trị gần 1 có nghĩa là thông tin sẽ được giữ lại.

Cổng Vào (Input Gate):

* Chức năng: Quyết định thông tin mới nào từ đầu vào hiện tại cần được lưu trữ vào trạng thái.
* Hoạt động: Cổng vào cũng sử dụng hàm sigmoid để xác định những thông tin nào từ đầu vào cần được thêm vào trạng thái. Sau đó, một hàm kích hoạt tanh sẽ tạo ra một vectơ mới chứa thông tin mới để thêm vào trạng thái.

Cập Nhật Trạng Thái (Cell State Update):

* Chức năng: Tính toán trạng thái mới của đơn vị nhớ bằng cách kết hợp thông tin cũ và thông tin mới.
* Hoạt động: Trạng thái mới được tính bằng cách kết hợp trạng thái cũ đã điều chỉnh qua cổng quên và thông tin mới đã điều chỉnh qua cổng vào.

Cổng Đầu Ra (Output Gate):

* Chức năng: Quyết định phần nào của trạng thái sẽ được sử dụng để tính toán đầu ra hiện tại.
* Hoạt động: Cổng đầu ra sử dụng hàm sigmoid để quyết định thông tin nào từ trạng thái mới cần được sử dụng làm đầu ra. Hàm tanh được áp dụng lên trạng thái mới để chuẩn hóa giá trị đầu ra.

Ưu điểm của LSTM:

* Khả năng học các quan hệ dài hạn: LSTM có thể nhớ và học các mối quan hệ dài hạn trong chuỗi dữ liệu, khắc phục vấn đề vanishing gradient của RNN truyền thống.
* Linh hoạt và mạnh mẽ: Có thể áp dụng cho nhiều loại dữ liệu tuần tự khác nhau như văn bản, âm thanh, chuỗi thời gian, và video.

Nhược điểm:

* Độ phức tạp tính toán cao: LSTM yêu cầu nhiều tài nguyên tính toán hơn so với các mô hình truyền thống.
* Khó khăn trong việc điều chỉnh tham số: Cần có kiến thức chuyên sâu để điều chỉnh - các tham số và cấu trúc của mô hình để đạt hiệu suất tốt nhất.

Ứng Dụng của LSTM

* Dự báo chuỗi thời gian: Dự báo giá cổ phiếu, dự báo nhu cầu sản phẩm, dự báo thời tiết.
* Xử lý ngôn ngữ tự nhiên: Dịch máy, phân tích cảm xúc, nhận diện thực thể có tên.
* Nhận dạng giọng nói: Chuyển đổi giọng nói thành văn bản.
* Xử lý video: Nhận diện hành động trong video, phân tích video.

LSTM là một công cụ mạnh mẽ trong xử lý chuỗi thời gian và dữ liệu tuần tự nhờ vào khả năng ghi nhớ và học các mối quan hệ dài hạn. Với cấu trúc phức tạp nhưng hiệu quả, LSTM đã trở thành một trong những phương pháp tiên tiến nhất cho các ứng dụng dự báo và phân tích chuỗi thời gian, xử lý ngôn ngữ tự nhiên, và nhiều lĩnh vực khác.

# CHƯƠNG 3: XÂY DỰNG MÔ HÌNH HUẤN LUYỆN

## Phân khúc khách hàng bằng mô hình RFM và K-Mean

**Giả định:**

Một công ty cần xác định các phân khúc khách hàng tiềm năng để giúp nhắm mục tiêu thị trường một cách tốt hơn và hiệu quả hơn. Thông tin này sẽ đảm bảo rằng tất cả các chiến dịch và nỗ lực tiếp thị trong tương lai đều được hướng đến những khách hàng có khả năng thanh toán và mua hàng cao.

Để xác định các phân khúc trong cơ sở khách hàng hiện tại, chúng ta sẽ sử dụng thuật toán phân cụm k-means. Sau đó, chúng ta có thể xác định phân khúc khách hàng tốt nhất bằng cách sử dụng mô hình RFM (Recency, Frequency, Monetary). Từ đó sẽ đề xuất một số khuyến nghị kinh doanh phù hợp với các đặc điểm của phân khúc khách hàng tốt nhất.

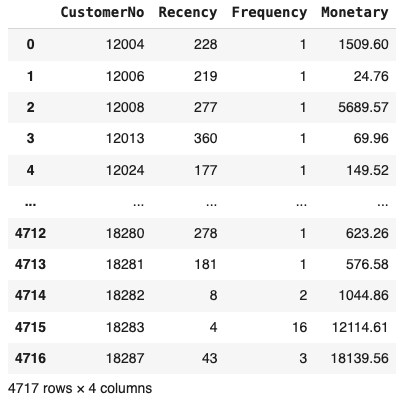
### 3.1.1. RFM Data

Đầu tiên cần gom theo CustomerNo rồi biến đổi để có được các thông tin sau:

* *Recency:* Chúng ta sẽ xác định thời gian gần đây bằng cách trừ ngày phân tích với ngày gần nhất được ghi nhận, giả sử việc phân tích được thực hiện một ngày sau ngày giao dịch cuối cùng được ghi nhận.
* *Frequency:* Đối với mỗi khách hàng, chúng ta sẽ tính chỉ số tần suất bằng cách đếm số số mã giao dịch duy nhất của họ, cho biết tổng số giao dịch riêng biệt mà họ đã thực hiện.
* *Monetary:* Chúng ta sẽ tổng hợp tổng số tiền mà mỗi khách hàng đã chi tiêu từ cột Revenue.

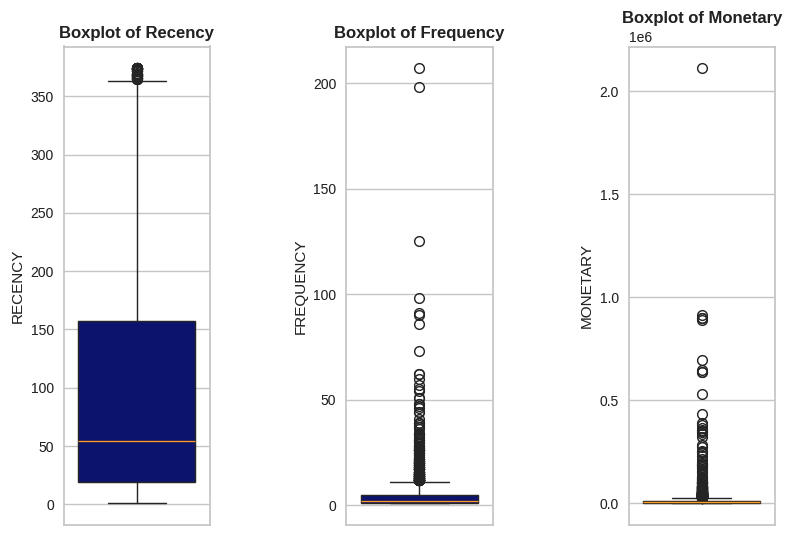
| rfm\_df = cleaned\_df.groupby('CustomerNo', as\_index = False)[['Revenue', 'TransactionNo', 'Date']]\  .agg({'Date' : 'max',  'TransactionNo' : 'nunique',  'Revenue' : 'sum',})  rfm\_df.columns = ['CustomerNo', 'Recency', 'Frequency', 'Monetary']  rfm\_df['Recency'] = pd.to\_datetime(rfm\_df['Recency'])  date = pd.to\_datetime('2019-12-10')  rfm\_df['Recency'] = date - rfm\_df['Recency']  rfm\_df['Recency'] = rfm\_df['Recency'].dt.days  rfm\_df |
| --- |

Kết quả thu được: có tổng cộng 4.717 khách hàng



### 3.1.2. Xử lý Outlier

| cols = rfm\_df.columns[1:]  fig, ax = plt.subplots(1, 3)  for col in range(len(cols)):  sns.boxplot(y = cols[col], data = rfm\_df, color = 'darkblue',  medianprops = {'color' : 'orange'}, ax = ax[col])  ax[col].set\_ylabel(cols[col].upper(),)  ax[col].set\_title(f'Boxplot of {cols[col]}', weight = 'bold')  plt.tight\_layout(w\_pad = 5) |
| --- |

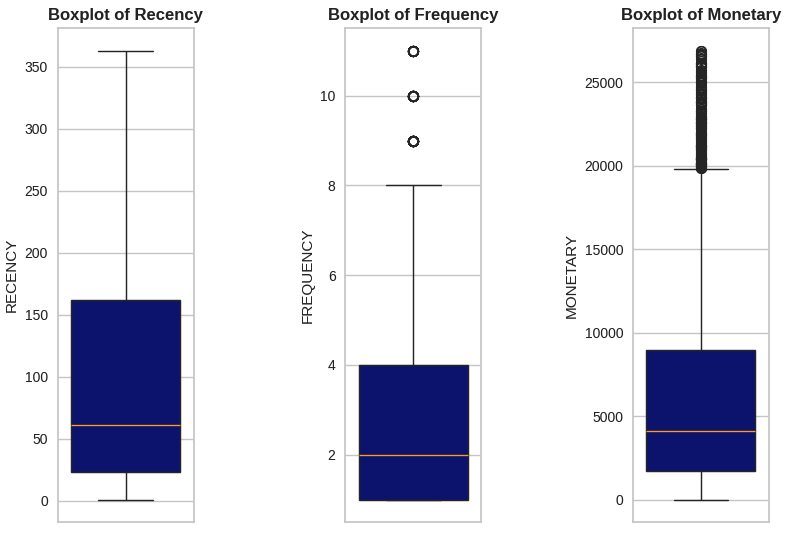
Kết quả thu được:  


**Nhận xét:**

Các biểu đồ boxplot cho thấy cả ba chỉ số Recency, Frequency và Monetary đều có sự xuất hiện của nhiều outliers (các giá trị ngoại lai). Điều này có thể ảnh hưởng đáng kể đến việc phân cụm, làm cho kết quả phân cụm không chính xác vì các outliers có thể làm sai lệch trung tâm của các cụm.

Do đó, chúng ta thực hiện xử lí theo quy tác IQR (Interquartile Range - Khoảng tứ phân vị)

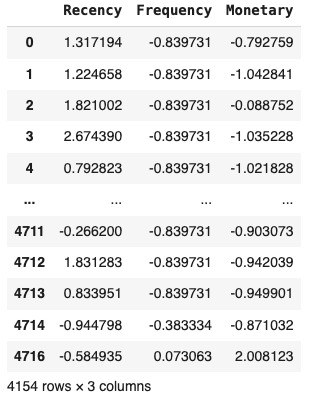
| Q1 = rfm\_df.quantile(0.25)  Q3 = rfm\_df.quantile(0.75)  IQR = Q3 - Q1  rfm\_df = rfm\_df[~((rfm\_df < (Q1 - 1.5 \* IQR)) |(rfm\_df > (Q3 + 1.5 \* IQR))).any(axis=1)] |
| --- |



### 3.1.3. Scale

| scaler = StandardScaler().set\_output(transform = 'pandas')  scaler.fit(rfm\_df.iloc[:, 1:])  rfm\_df\_std = scaler.transform(rfm\_df.iloc[:, 1:])  rfm\_df\_std |
| --- |

Kết quả thu được:



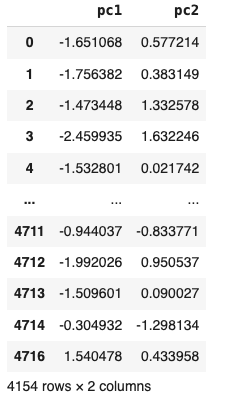
**Nhận xét:**

Vì các thuật toán phân cụm dựa vào việc tính toán khoảng cách để đánh giá sự tương đồng giữa các điểm dữ liệu, việc chuẩn hóa dữ liệu là rất quan trọng để đảm bảo kết quả phân cụm chính xác và có ý nghĩa. Việc chuẩn hóa sẽ làm cho dữ liệu trở nên đồng nhất, giới hạn các biến số có phạm vi lớn hơn, như Monetary trong trường hợp này, ảnh hưởng quá mức đến các tính toán khoảng cách. Việc này đảm bảo rằng tất cả các biến số đều tác động một cách đồng đều vào quá trình phân cụm, dẫn đến kết quả cân bằng và tốt hơn.

### 3.1.4. PCA

| pca = PCA(n\_components = 2)  pca.fit(rfm\_df\_std)  pca\_df = pd.DataFrame(pca.transform(rfm\_df\_std), columns = ['pc1', 'pc2'])  pca\_df.set\_index(rfm\_df.index, inplace = True)  pca\_df |
| --- |

Kết quả thu được:

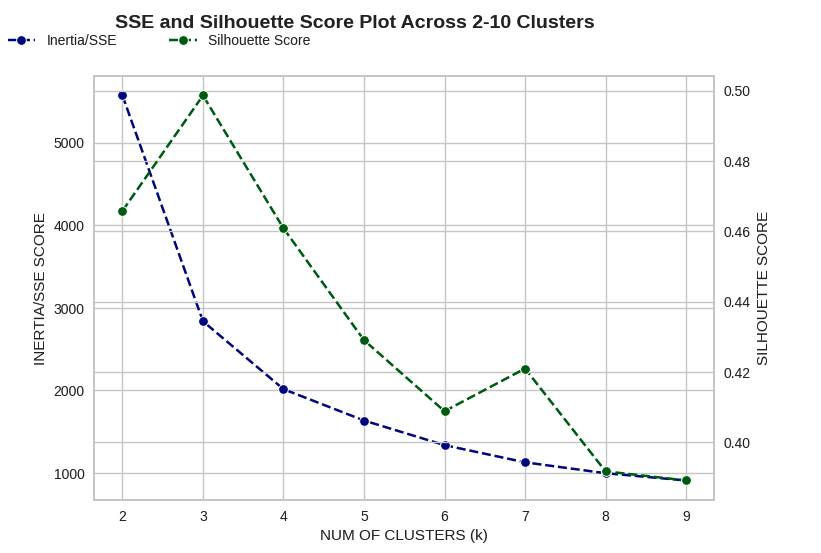


**Nhận xét:**  
Sử dụng PCA để giảm chiều dữ liệu, giảm sự tác động của đa cộng tuyến (các biến của RFM có tính đa cộng tuyến cao)

### 3.1.5. Phân cụm với K-Mean

| sse, silhouette = [], []  # Tính toán SSE và silhouette scores cho k từ 2 đến 10  for k in range(2, 10):  kmeans = KMeans(n\_clusters=k, random\_state=10).fit(pca\_df)  sse.append(kmeans.inertia\_)  silhouette.append(silhouette\_score(pca\_df, kmeans.labels\_))  fig, ax1 = plt.subplots()  # Vẽ đồ thị cho SSE Scores  sns.lineplot(x=range(2, 10), y=sse, marker='o', linestyle='--', color='darkblue', label='Inertia/SSE', ax=ax1)  # Vẽ đồ thị cho Silhouette Scores trên trục y thứ hai  ax2 = ax1.twinx()  sns.lineplot(x=range(2, 10), y=silhouette, marker='o', linestyle='--', color='darkgreen', label='Silhouette Score', ax=ax2)  ax1.set\_xlabel('NUM OF CLUSTERS (k)')  ax1.set\_ylabel('INERTIA/SSE SCORE')  ax2.set\_ylabel('SILHOUETTE SCORE')  ax2.set\_title('SSE and Silhouette Score Plot Across 2-10 Clusters', size=14, pad = 35,  x = 0.42, weight='bold')  ax1.legend(loc=2, frameon=False, bbox\_to\_anchor = [-0.16, 1.13])  ax2.legend(loc=2, frameon=False, bbox\_to\_anchor = [0.1, 1.13])  plt.show() |
| --- |

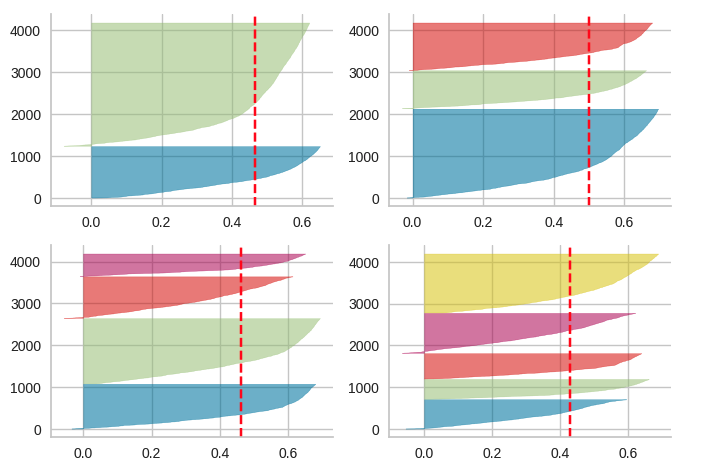
Kết quả thu được:



**Nhận xét:**Chúng ta có thể thấy rằng giá trị tối ưu của k là 3, giá trị này có điểm hình bóng cao nhất và nằm ở phần 'khuỷu tay' của biểu đồ SSE. Chúng ta có thể kiểm tra thêm bằng biểu đồ hình bóng.

| fig, ax = plt.subplots(2, 2)  for k in range(2, 6):  km = KMeans(n\_clusters = k, random\_state = 10)  q, mod = divmod(k, 2)  visualizer = SilhouetteVisualizer(km, colors = 'yellowbrick', ax = ax[q-1][mod])  visualizer.fit(pca\_df)  sns.despine(right = True)  plt.show() |
| --- |

Kết quả thu được:



**Nhận xét:**Tất cả các giá trị k này cho thấy tỷ lệ cụm không cân bằng, nhưng k = 3 có vẻ ít mất cân bằng hơn các giá trị khác. Hơn nữa, k = 3 cho thấy điểm bóng âm ít nhất trên mỗi cụm.

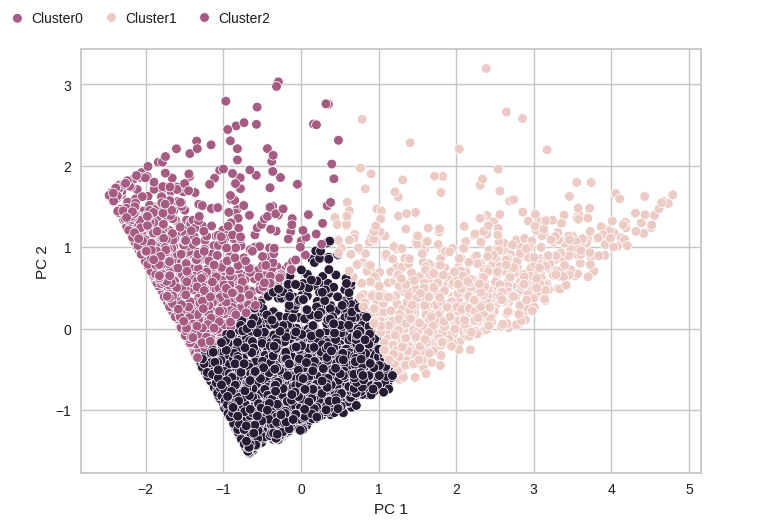
Áp dụng k bằng 3 trong quá trình phân cụm và lấy nhãn cụm.

| km = KMeans(n\_clusters = 3, random\_state = 10)  km.fit(pca\_df)  pca\_df['Cluster'] = km.labels\_  rfm\_df['Cluster'] = pca\_df['Cluster'] |
| --- |

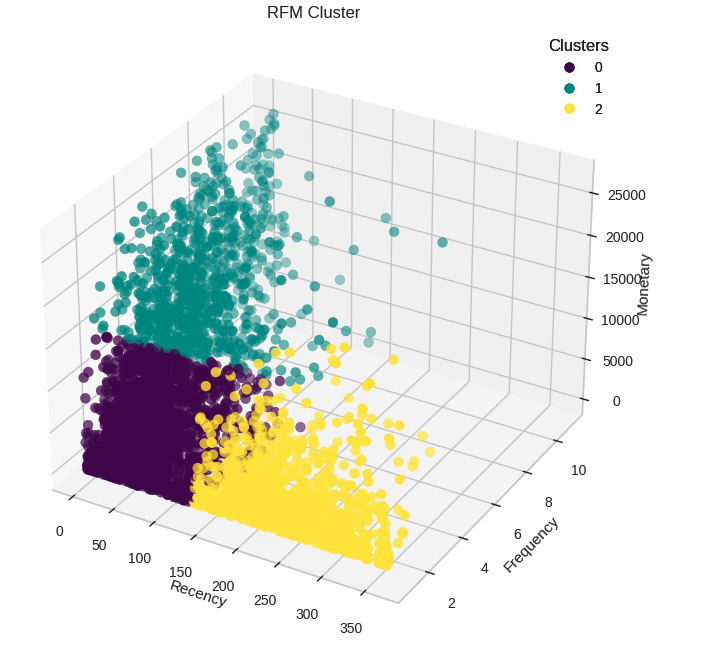
Phân tích kết quả phan cụm

| fig, ax = plt.subplots()  sns.scatterplot(x='pc1', y='pc2', hue='Cluster', data=pca\_df, ax=ax)  ax.set\_xlabel('PC 1')  ax.set\_ylabel('PC 2')  ax.legend(loc=2, ncol=3, frameon=False, bbox\_to\_anchor=[-0.145, 1.12],  labels=['Cluster' + str(i) for i in sorted(pca\_df['Cluster'].unique())],  handletextpad=0, columnspacing=1)  plt.show() |
| --- |

Kết quả thu được:



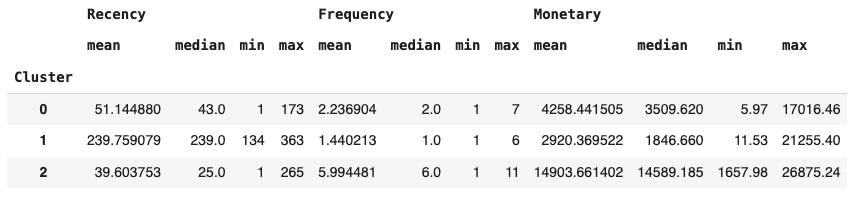
| fig = plt.figure(figsize=(10, 8))  ax = fig.add\_subplot(111, projection='3d')  scatter = ax.scatter(rfm\_df['Recency'], rfm\_df['Frequency'], rfm\_df['Monetary'],c=rfm\_df['Cluster'], cmap='viridis', s=50)  ax.set\_xlabel('Recency')  ax.set\_ylabel('Frequency')  ax.set\_zlabel('Monetary')  ax.set\_title('RFM Cluster')  legend = ax.legend(\*scatter.legend\_elements(), title="Clusters")  ax.add\_artist(legend)  plt.show() |
| --- |

Kết quả thu được:  


**Nhận xét:**

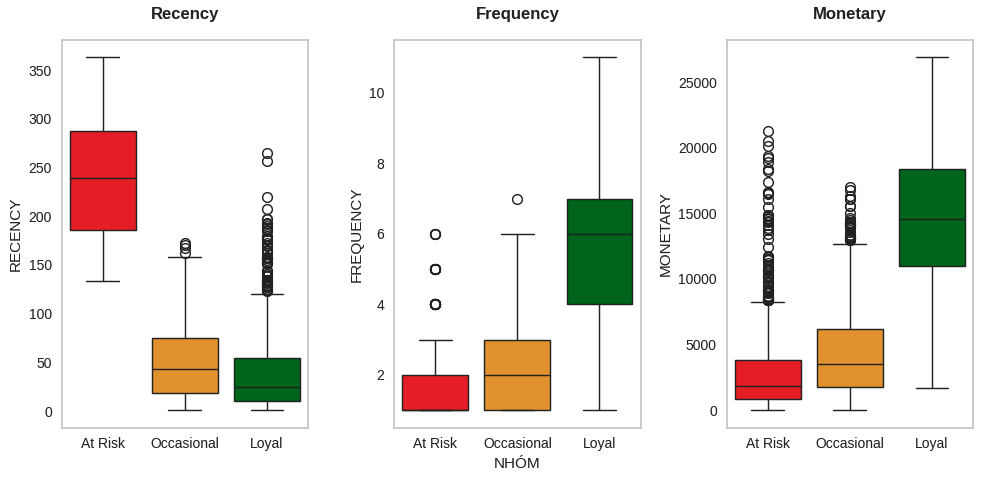
Vẫn có một số điểm chưa thật sự tốt. Mặc dù vậy hình dạng tổng thể trên biểu đồ cho ta thấy kết quả phân cụm khá ổn để có thể sử dụng được.

**Tóm tắt kết quả phân cụm và so sánh:**



Dựa trên số liệu thống kê tóm tắt của các cụm đã xác định, chúng ta có thể gắn nhãn cho từng cụm như sau:

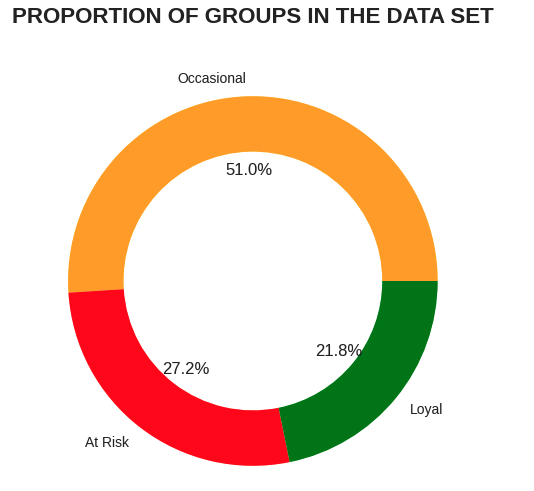
* Nhóm 0: Khách hàng không thường xuyên (R, F và M vừa phải)
* Nhóm 1: Khách hàng có nguy cơ (R cao nhất, F và M thấp nhất)
* Nhóm 2: Khách hàng trung thành (R thấp nhất, F và M cao nhất)



Chúng ta có thể nhận thấy các đặc điểm chính của từng nhóm khách hàng:

* **Nhóm khách hàng có nguy cơ:**
* Recency cao: Khách hàng không thực hiện giao dịch gần đây, cho thấy nguy cơ mất quan tâm.
* Frequency và Monetary thấp: Họ mua hàng ít và chi tiêu ít khi mua hàng.
* **Nhóm Khách hàng không thường xuyên:**
* Recency và Frequency vừa: Họ thực hiện giao dịch một cách vừa phải.
* Monetary vừa: Họ chi tiêu một số tiền trung bình cho mỗi giao dịch.
* **Khách hàng trung thành:**
* Recency thấp: Họ thực hiện giao dịch gần đây, cho thấy tính chất tích cực.
* Frequency và Monetary cao: Họ mua hàng thường xuyên và chi tiêu nhiều.

Dựa trên các đặc điểm này, cụm Khách hàng Trung Thành là mục tiêu tiềm năng để nhắm đến.



Tóm lại, phân tích dữ liệu cho thấy rằng trong tệp dữ liệu khách hàng của công ty, có ba nhóm chính: Khách hàng Không thường xuyên, Khách hàng Có nguy cơ và Khách hàng Trung thành.

* Khách hàng Không thường xuyên chiếm phần lớn, với 51.02% .
* Tiếp theo là nhóm Khách hàng Có nguy cơ, với 27.16% .
* Nhóm ít nhất là Khách hàng Trung thành, chỉ có 21.82%.

Nhìn chung, số lượng khách hàng thực sự trung thành với cửa hàng và sản phẩm ít hơn so với nhóm khách hàng không thường xuyên và có nguy cơ, điều này đáng lo ngại.

**Để giải quyết vấn đề này, công ty có thể thực hiện các chiến lược sau:**

1. **Tăng cường tương tác và giao tiếp:** Tạo ra các chiến dịch tiếp thị và quảng cáo để tăng cường tương tác với khách hàng hiện tại. Sử dụng email marketing, tin nhắn và mạng xã hội để liên lạc và tương tác thường xuyên với khách hàng, đảm bảo họ cảm thấy được quan tâm và đánh giá.
2. **Cải thiện trải nghiệm khách hàng:** Tập trung vào cải thiện trải nghiệm mua sắm trực tuyến bằng cách tối ưu hóa giao diện website, cải thiện tính dễ sử dụng và tăng tính tương tác của nền tảng. Đảm bảo quy trình mua hàng trơn tru và thuận tiện sẽ giúp tăng sự hài lòng của khách hàng.
3. **Tạo chương trình khuyến mãi và ưu đãi:** Phát triển các chương trình khuyến mãi và ưu đãi đặc biệt dành cho khách hàng trung thành để khuyến khích họ tiếp tục mua sắm và gắn bó với thương hiệu của công ty. Cung cấp điểm thưởng, phiếu quà tặng và giảm giá độc quyền là những cách hiệu quả để tạo ra sự hứng thú và khích lệ khách hàng quay lại mua hàng.
4. **Nghiên cứu và phân tích:** Tiến hành nghiên cứu để hiểu rõ hơn về nhu cầu và mong muốn của từng nhóm khách hàng. Dựa vào thông tin này, tinh chỉnh chiến lược tiếp thị và sản phẩm để đáp ứng tốt hơn nhu cầu của khách hàng và tăng cường sự hài lòng.

Kết hợp các chiến lược này sẽ giúp công ty tăng cường cam kết và tương tác với khách hàng, từ đó tạo ra một cộng đồng khách hàng trung thành và ổn định hơn đối với thương hiệu của công ty.

## Mô hình khai phá luật kết hợp sử dụng thuật toán Apriori và FP-Growth

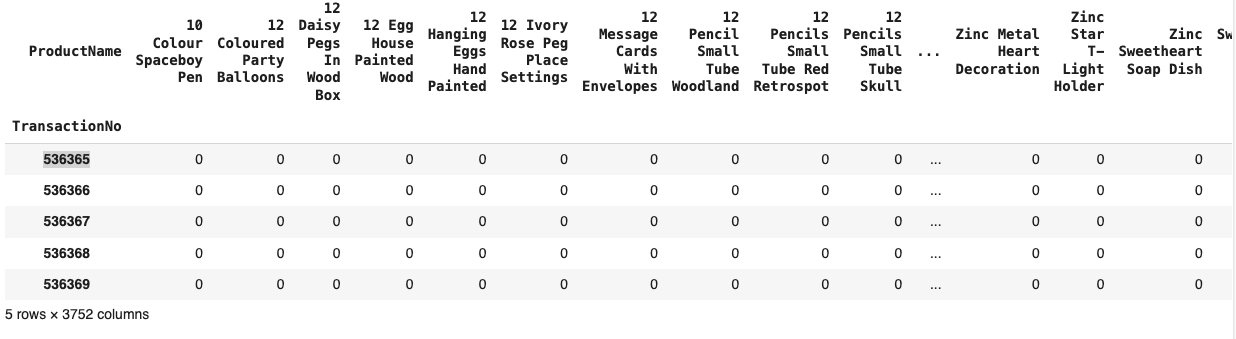
### 3.2.1. Sử dụng Thuật toán Apriori

Ý tưởng bài toán:

Như vấn đề đã đề cập ở trên việc công ty đang gặp tình trạng số lượng khách hàng thực sự trung thành với cửa hàng và sản phẩm ít hơn so với nhóm khách hàng không thường xuyên và có nguy cơ, điều này đáng lo ngại.

Các chiến lược cũng đã được đưa ra để giải quyết vấn đề, bắt đầu thực hiện công ty sẽ muốn tìm các tập sản phẩm phổ biến trong các giao dịch để đề xuất sản phẩm, giảm giá khi mua cùng nhau cải thiện sự hài lòng của khách hàng khi mua hàng, và kích thích khách hàng mua hàng.

| basket = cleaned\_df.pivot\_table(index='TransactionNo',  columns='ProductName',  aggfunc='size',  fill\_value=0)  basket = basket.applymap(lambda x: 1 if x > 0 else 0)  basket.head(5) |
| --- |

Kết quả thu được:  


**Nhận xét:**

Dữ liệu trên đang thể hiện số lượng của từng sản phẩm được mua trong mỗi giao dịch (TransactionNo). Mỗi hàng tương ứng với một giao dịch và mỗi cột tương ứng với một sản phẩm. Cụ thể:

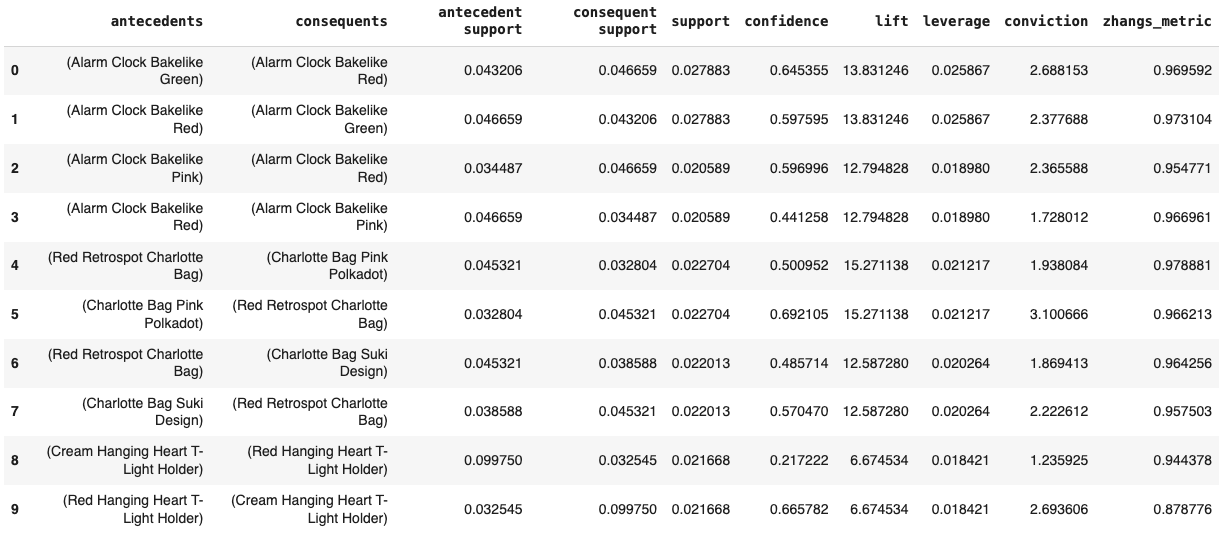
Mỗi giá trị trong bảng cho biết số lượng của mỗi sản phẩm được mua trong một giao dịch cụ thể. Ví dụ, giá trị ở hàng đầu tiên, cột "10 Colour Spaceboy Pen" và "12 Coloured Party Balloons" đều là 0, ngụ ý rằng trong giao dịch đó không có sản phẩm nào được mua. Giá trị ở hàng thứ hai, cột "12 Daisy Pegs In Wood Box" và "12 Egg House Painted Wood" cũng đều là 0, tương tự như vậy.

| frequent\_itemsets = apriori(basket, min\_support=0.02, use\_colnames=True)  rules = association\_rules(frequent\_itemsets, metric="lift", min\_threshold=1)  frequent\_itemsets.sort\_values('support', ascending=False) |
| --- |

Kết quả thu được:  


**Nhận xét:**  
Kết quả từ quá trình khai phá các tập hợp phổ biến với thuật toán Apriori cho thấy các sản phẩm và tổ hợp sản phẩm có tần suất xuất hiện cao nhất trong các giao dịch.

**Kết quả của quá trình khai phá kết hợp với thuật toán Apriori:**

****

**Ý nghĩa của việc khai phá với bài toán đặt ra:**

* **Thiết Kế Các Gói Khuyến Mãi Hấp Dẫn:**

Thông tin: Với các cặp sản phẩm có tỉ lệ mua cùng nhau cao (high confidence), cửa hàng có thể thiết kế các gói khuyến mãi hoặc giảm giá cho các nhóm sản phẩm này.

Ứng dụng: Ví dụ, nếu "Alarm Clock Bakelike Green" và "Alarm Clock Bakelike Red" thường được mua cùng nhau, cửa hàng có thể tạo ra các gói khuyến mãi giảm giá khi mua cả hai sản phẩm.

* **Tăng Doanh Số Bằng Chiến Lược Bán Kèm:**

Thông tin: Kết quả phân tích cung cấp các thông tin về lift (mức độ tăng khả năng mua sản phẩm kết hợp) và confidence (mức độ tin cậy của việc mua sản phẩm cùng nhau).

Ứng dụng: Cửa hàng có thể áp dụng chiến lược bán kèm cho các sản phẩm có lift cao. Ví dụ, khi khách hàng mua "Charlotte Bag Pink Polkadot", hệ thống có thể gợi ý mua thêm "Red Retrospot Charlotte Bag" với một mức giảm giá nhỏ, dựa trên việc hai sản phẩm này thường được mua cùng nhau.

* **Cải Thiện Chiến Lược Marketing:**

Thông tin: Kết quả từ thuật toán Apriori giúp xác định các sản phẩm nào có mối liên hệ mạnh mẽ với nhau.

Ứng dụng: Cửa hàng có thể sử dụng thông tin này để thực hiện các chiến dịch quảng cáo tập trung vào các nhóm sản phẩm liên quan. Điều này giúp tăng hiệu quả của các chiến dịch quảng cáo và tối ưu hóa chi phí marketing.

### 3.2.2. Sử dụng thuật toán FP-Growth

Tìm các mẫu sản phẩm đặc trưng cho từng quốc gia. Điều này có thể giúp hiểu rõ hơn về sở thích mua sắm của khách hàng từ các quốc gia khác nhau.

Từ đó có thể đưa ra các chính sách và khuyến mãi phù hợp với các từng quốc gia:

- Hiểu rõ hơn về sở thích mua sắm của khách hàng từ các quốc gia khác nhau.

- Phát triển các chiến lược tiếp thị và bán hàng được nhắm mục tiêu đến từng quốc gia.

- Khuyến mãi và đề xuất sản phẩm phù hợp với thị hiếu mua sắm của từng quốc gia.

**Từ Ma trận basket ở phần trước, thêm vào thông tin quốc gia**

|  |
| --- |

### 

**Tìm luật kết hợp cho các quốc gia**

|  |
| --- |

**Xây dựng hàm để tìm sản phẩm đặc trưng cho từng quốc gia**

|  |
| --- |

**Sử dụng hàm đã định nghĩa để tìm sản phẩm đặc trưng cho từng quốc gia**

|  |
| --- |

|  |
| --- |

**Hiểu hành vi mua hàng:** Bằng cách phân tích các mẫu mua hàng, ta có thể nhận biết các sản phẩm thường được mua cùng nhau bởi khách hàng từ cùng một quốc gia. Điều này giúp hiểu rõ hơn về sở thích, nhu cầu và hành vi mua sắm của khách hàng trong từng quốc gia cụ thể.

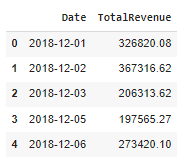
**Tối ưu hóa kinh doanh:** Các sản phẩm được xác định là đặc trưng cho mỗi quốc gia có thể được sử dụng để tối ưu hóa chiến lược kinh doanh, từ việc phát triển sản phẩm đến quảng cáo và tiếp thị.

## Dự báo doanh thu sử dụng LSTM và ARIMA

Để có thể xác định việc chạy các chiến dịch có hiệu quả hay không thì chúng ta cần 1 thước đo, 1 điểm để so sánh. Do đó việc dự đoán doanh thu trong tương lai có thể giúp doanh nghiệp xem xét các yếu tố để có thể so sánh nó với thực tế, đánh giá hiệu quả của các chiến dịch trong tương lai.

### 3.3.1. Tùy chỉnh lại bộ dữ liệu

| df\_grouped = cleaned\_df.groupby('Date')['Revenue'].sum().reset\_index()  df\_grouped.columns = ['Date', 'TotalRevenue']  df\_grouped.head() |
| --- |



|  |
| --- |

# 

### 3.3.2. LSTM

Sau Khi Dùng GridSearchCV thì xác định:

| best\_params = {'activation': 'relu', 'batch\_size': 32, 'dropout': 0.0,'epochs': 80, 'optimizer': 'adam', 'units': 65} |
| --- |

Chuẩn hóa dữ liệu về khoảng [0, 1]

|  |
| --- |

Chia dữ liệu thành các chuỗi thời gian

|  |
| --- |

Chia dữ liệu thành tập huấn luyện và tập kiểm tra

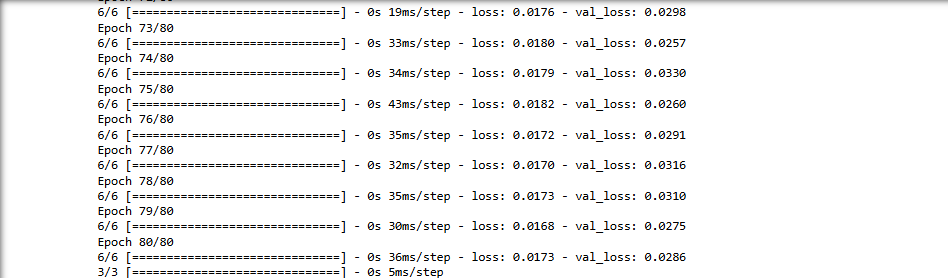
|  |
| --- |

Xây dựng mô hình LSTM với tham số tốt nhất

|  |
| --- |

Huấn luyện mô hình

|  |
| --- |



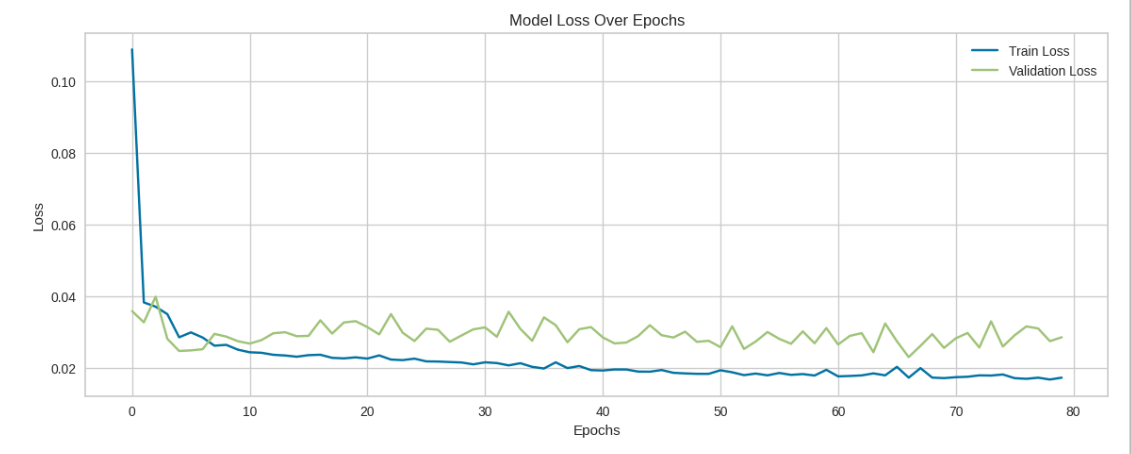
Đánh giá mô hình trên tập kiểm tra

|  |
| --- |

|  |
| --- |

Biểu đồ thể hiện kết quả dự đoán

|  |
| --- |



Đưa các dự đoán và giá trị thực tế trở lại tỷ lệ gốc của chúng

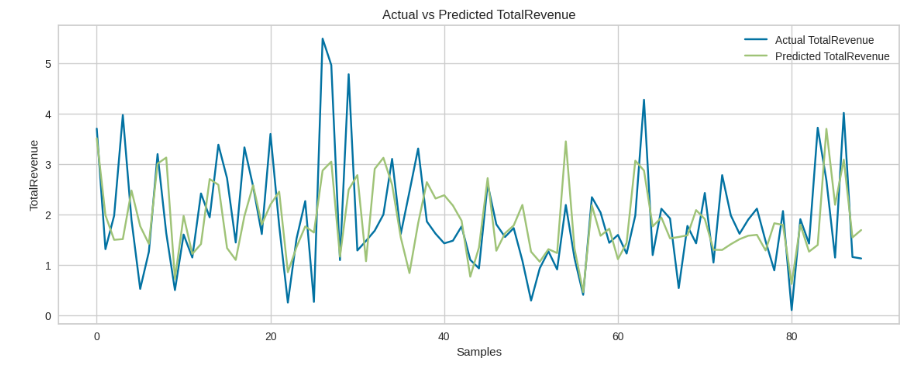
|  |
| --- |

Tạo và hiển thị Data Frame chứa giá trị dự đoán và giá trị thực tế:

|  |
| --- |

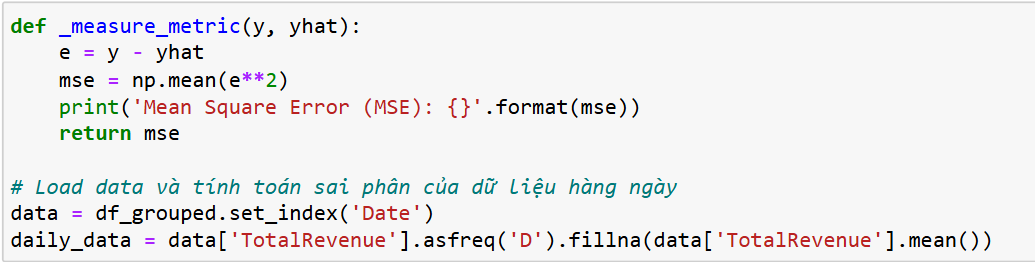
Hiển thị kết quả so sánh bằng biểu đồ

|  |
| --- |



### 3.3.3. ARIMA

Load lại data và định nghĩa hàm để đo lường các chỉ số lỗi

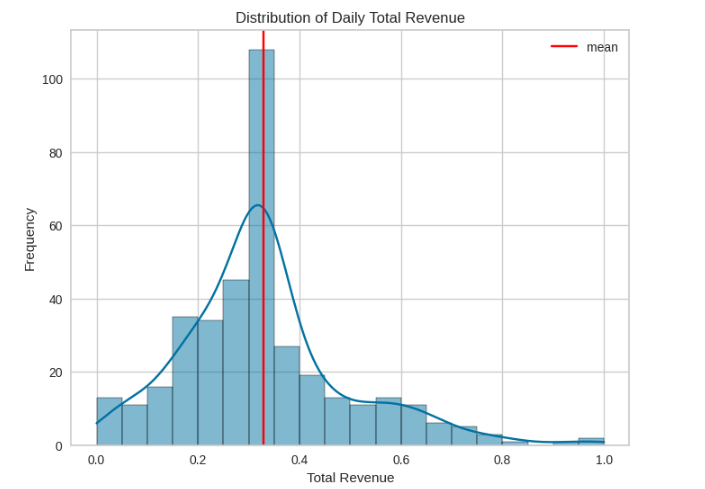


Trong mô hình ARIMA ((p, d, q)):

* (p) là bậc của phần tự hồi quy (AR)
* (d) là bậc của sai phân (differencing)
* (q) là bậc của phần trung bình trượt (MA)

Tiến hành phân tích để chọn p, d, q cho mô hình:

|  |
| --- |



ADF Statistic for Original Data: -2.362117451842031

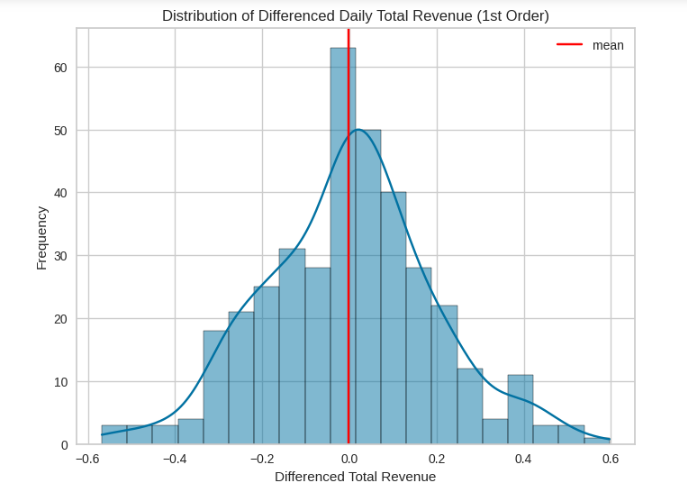
p-value for Original Data: 0.15271509659661692

Nhận xét với dữ liệu gốc:

* ADF Statistic: -2.2619191469273927
* p-value: 0.18454399136033556 Giá trị p lớn hơn 0.05, điều này có nghĩa là không thể bác bỏ giả thuyết gốc (chuỗi thời gian có gốc đơn và không dừng). Nói cách khác, dữ liệu gốc là không dừng.

**Tiếp tục cho sai phân bậc 1**

|  |
| --- |

****

DF Statistic for Differenced Data (1st Order): -8.343260289238136

p-value for Differenced Data (1st Order): 3.1344588996153606e-13

Dữ liệu sau khi lấy sai phân bậc 1:

* ADF Statistic: -8.681035609714435
* p-value: 4.2855275734739095e-14

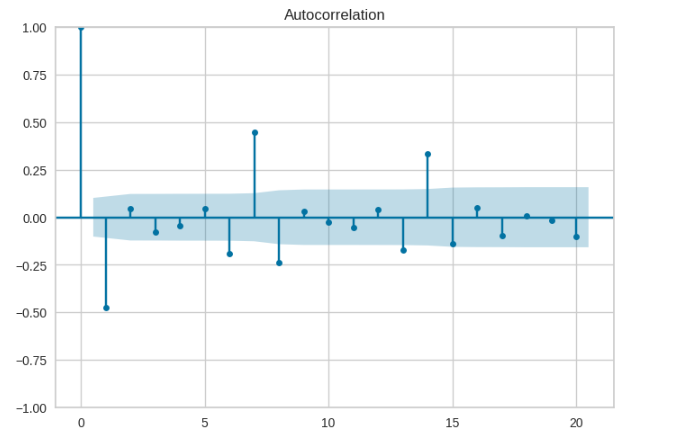
Giá trị p nhỏ hơn 0.05 rất nhiều, điều này có nghĩa là có thể bác bỏ giả thuyết gốc. Nói cách khác, dữ liệu sau khi lấy sai phân bậc 1 là dừng.

Vì chuỗi thời gian đã trở nên dừng sau khi lấy sai phân bậc 1, không cần thiết phải tiếp tục lấy sai phân bậc 2 và vẽ kết quả kiểm tra ADF. Sai phân bậc 1 đã đủ để làm chuỗi thời gian dừng.

**-> Nên chọn d = 1**

**Tiếp tục phân tích:**

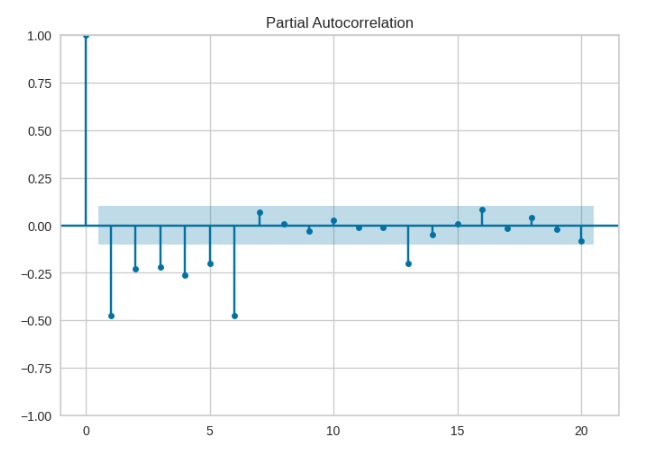
|  |
| --- |

****

Dựa trên biểu đồ ACF trong hình :

Tham số p (AR): Xác định bằng biểu đồ PACF. Chọn p là độ mà tại đó PACF có giá trị đáng kể và nằm ngoài vùng tin cậy.

- > 0, 1, 6, 7, 8, 14

****

Tham số q (MA): Xác định bằng biểu đồ ACF. Chọn q là số độ mà tại đó ACF có giá trị đáng kể và nằm ngoài vùng tin cậy.

-> 0, 1, 2, 3,4 ,5, 6, 13

Vậy tham số xác định ở đây là:

* p = [0, 1, 2, 3, 4, 5, 6,1 3]
* d = 1
* q = [0, 1, 6, 7, 8 , 14]

**Chuẩn bị dữ liệu cho mô hình ARIMA**

|  |
| --- |

**Xây dựng và đánh giá mô hình ARIMA**

|  |
| --- |

|  |
| --- |

|  |
| --- |

|  |
| --- |

Dự đoán và đánh giá mô hình

|  |
| --- |

|  |
| --- |

Vẽ biểu đồ kết quả dự đoán và so sánh với giá trị thực tế

|  |
| --- |

|  |
| --- |

### 3.3.4. So sánh hiệu suất mô hình

|  |
| --- |

|  |
| --- |

Dựa trên biểu đồ về MSE của hai mô hình LSTM và ARIMA:

* MSE của LSTM: MSE thấp hơn cho thấy mô hình LSTM có khả năng dự báo tốt hơn trong trường hợp này.
* MSE của ARIMA: MSE cao hơn so với LSTM, điều này có thể chỉ ra rằng mô hình ARIMA không bám sát dữ liệu thực tế như LSTM.

Như vậy doanh nghiệp có thể sử dụng mô hình dự đoán được xây dựng từ LSTM để dự đoán doanh thu nhằm có chiến lược tốt hơn để thu hút khách hàng cũng như lấy giá trị dự báo đấy để đánh giá các hoạt động trong tương lai.

# KẾT LUẬN

Thương mại điện tử ngày càng đóng vai trò quan trọng trong nền kinh tế số, mang đến nhiều cơ hội phát triển cho doanh nghiệp. Việc khai thác dữ liệu giao dịch bán hàng bằng các thuật toán tiên tiến như K-Means, Apriori, FP-Growth, ARIMA và LSTM là một hướng đi hiệu quả để dự đoán doanh thu trong tương lai. Doanh nghiệp có thể dựa vào dự báo này để đánh giá hiệu quả hoạt động, tối ưu hóa chiến lược kinh doanh và đưa ra quyết định phù hợp, từ đó nâng cao khả năng cạnh tranh trên thị trường.

Dự đoán doanh thu không chỉ giúp doanh nghiệp đánh giá hiệu quả hoạt động mà còn là công cụ hữu ích để:

* **Lập kế hoạch tài chính:** Dự đoán doanh thu giúp doanh nghiệp dự trù các khoản chi phí, đầu tư và đảm bảo nguồn vốn hoạt động hiệu quả.
* **Quản lý kho hàng:** Dự đoán nhu cầu thị trường giúp doanh nghiệp sắp xếp lượng hàng hóa hợp lý, tránh tình trạng tồn kho hoặc thiếu hụt.
* **Triển khai chiến dịch marketing:** Dự đoán doanh thu giúp doanh nghiệp điều chỉnh chiến dịch marketing phù hợp với mục tiêu, tối ưu hóa hiệu quả chi tiêu.
* **Đánh giá hiệu quả hoạt động:** Dự đoán doanh thu giúp doanh nghiệp so sánh kết quả kinh doanh thực tế với dự báo, đánh giá hiệu quả của các chiến dịch và đưa ra điều chỉnh kịp thời.

Như vậy, việc dự đoán doanh thu trong thương mại điện tử là một công cụ quan trọng giúp doanh nghiệp vận hành hiệu quả, nâng cao khả năng cạnh tranh và phát triển bền vững.

# TÀI LIỆU THAM KHẢO

[1] Tài liệu các file PDF môn khai phá dữ liệu của GVHD Thầy - Ths. Trần Trọng Bình.

[2] Deep AI KhanhBlog, *13.1. Các bước của thuật toán k-Means Clustering*. Available at: <https://phamdinhkhanh.github.io/deepai-book/ch_ml/KMeans.html> , [Accessed 30 May 2024].

[3] Le Hoang, *Khai phá dữ liệu và lớp bài toán khai thác các tập phổ biến (P2)*. Available at: <https://viblo.asia/p/khai-pha-du-lieu-va-lop-bai-toan-khai-thac-cac-tap-pho-bien-p2-m68Z0W06KkG> [Accessed 30 May 2024].

[4] Khoa học dữ liệu - Khanh's blog, *Bài 19 - Mô hình ARIMA trong time series*. Available at: <https://phamdinhkhanh.github.io/2019/12/12/ARIMAmodel.html> [Accessed 28 May 2024].

[5] Baeldung, *Choosing the best q and p from ACF and PACF plots in ARMA-type modeling*. Available at: <https://www.baeldung.com/cs/acf-pacf-plots-arma-modeling> [Accessed 28 May 2024].

[6] Aayush Bajaj, *Anomaly Detection in Time Serie*s. Available at: <https://neptune.ai/blog/anomaly-detection-in-time-series#:~:text=The%20procedure%20for%20detecting%20anomalies,That's%20it> [Accessed 28 May 2024].

[7] scikit-learn, *MinMaxScalers*. Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>. [Accessed 26 May 2024].

[8] Nguyen Minh Duc, Thuật toán Apriori khai phá luật kết hợp trong Data Mining

<https://viblo.asia/p/thuat-toan-apriori-khai-pha-luat-ket-hop-trong-data-mining-3P0lPEv85ox> [Accessed 28 May 2024]

[9] Sruthy, Frequent Pattern (FP) Growth Algorithm In Data Mining

<https://www.softwaretestinghelp.com/fp-growth-algorithm-data-mining/> [Accessed 28 May 2024]