**PHÂN TÍCH CHUỖI THỜI GIAN DỰ BÁO LƯỢNG MUA HÀNG CỦA KHÁCH HÀNG VỚI MÔ HÌNH ARIMA**

Một trong những quan tâm tiên quyết của các cửa hàng là tối đa hóa lợi nhuận thu được. Đặc biệt là những cửa hàng thương mại điện tử, nơi mà số đơn đặt hàng thành công đóng góp rất quan trọng vào lợi nhuận của hàng. Ở đây, nhóm sẽ dựa trên dữ liệu lịch sử đặt hàng của một cửa hàng, sau đó phân tích và dự đoán trong thời gian tiếp theo, liệu số đơn đặt hàng là bao nhiêu, có xu hướng tăng lên hay giảm đi. Qua đó, giúp cho cửa hàng có những quyết định kịp thời về các chương trình khuyến mại, về giá cả mặt hàng khi biết những biến động về sức mua của người tiêu dùng.

Để thực hiện phân tích bằng mô hình ARIMA, ta sẽ thực hiện những bước sau :

* Visualize chuỗi thời gian để có những phán đoán sơ bộ về chuỗi
* Kiểm tra tính ổn định của chuỗi
* Lựa chọn các tham số cho mô hình ARIMA
* Train và đánh giá kết quả mô hình

# Tập dữ liệu

Tập dữ liệu E-Commerce của UCI Machine Learning Repository public. Dữ liệu bao gồm các lịch sử giao dịch từ 01/12/2010 đến 09/12/2011 từ một số cửa hàng bán lẻ trực tuyến online đến có trụ sở từ vương quốc Anh. Công ty chủ yếu bán quà tặng độc đáo. Nhiều khách hàng của công ty là khách hàng sỉ.

Dữ liệu bao gồm **692778** bản ghi giao dịch mua của khách hàng. Bao gồm 8 trường dữ liệu

|  |  |  |
| --- | --- | --- |
| 1 | InvoiceNo | Mã đơn hàng |
| 2 | StockCode | Mã sản phẩm |
| 3 | Description | Mô tả tên sản phẩm |
| 4 | Quantity | Số lượng sản phẩm |
| 5 | InvoiceDate | Ngày mua |
| 6 | UnitPrice | Giá của một sản phẩm |
| 7 | CustomerID | Mã khách hàng |
| 8 | Country | Quốc gia của cửa hàng |

Một dữ liệu mẫu:

Graphical user interface

Description automatically generated

Ở đây, ta tập trung dự đoán số lượng hóa đơn có thể bán của cửa hàng trong mỗi ngày, nên ta sẽ cần xử lý trước dữ liệu.

# Xử lý dữ liệu

## Làm sạch dữ liệu

* Kiểm tra xem có thuộc tính nào bị missing-value hay không?

Table

Description automatically generated

Nhận thấy, chỉ có 2 thuộc tính có missing-value là Description và CustomerID. Với những hóa đơn bị thiếu CustomerID, ta sẽ thay bằng giá trị -999.

Ở đây, ta sẽ quan tâm nhiều hơn đến trường CustomerID, xem xét xem liệu có thể loại bỏ hẳn những hóa đơn không rõ người mua hay không.

Graphical user interface, text, application

Description automatically generated

Dễ thấy:

* 24,94 % dữ liệu giao dịch không có CustomerID
* Hơn 16% tổng doanh thu là từ những khách hàng không có ID này

Vì vậy chúng ta không thể loại bỏ những hóa đơn bị thiếu CustomerID này.

* Lọc các hóa đơn hợp lệ:

Ngoài ra, trong các hóa đơn mua hàng có một số hóa đơn liên quan đến đơn hàng bị hủy hay đến khoản nợ khó đòi và những hóa đơn không đóng góp gì vào doanh thu của cửa hàng.

Các hóa đơn có InvoiceNo bắt đầu bằng chữ '**C**' là các đơn hàng đã bị hủy. Vì chúng không được tính vào tổng doanh thu, ta có thể xóa các trường giao dịch này khỏi dữ liệu.

Tương tự, một số giao dịch có InvoiceNo bắt đầu bằng '**A**' có liên quan đến việc điều chỉnh nợ khó đòi.

Graphical user interface, text, application, email

Description automatically generated

A picture containing graphical user interface

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Sau khi lọc tất cả 3 loại hóa đơn trên ra khỏi bộ dữ liệu, ta thu được các bản ghi thực sự có giá trị:

Graphical user interface

Description automatically generated with low confidence

* Với dữ liệu hiện tại, ta có một số thông tin sau :

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Nhận xét :

a) Những tháng nghỉ lễ (10, 11, 12) có lượng mua thực sự cao hơn. Số lượng mua hàng có xu hướng tăng trong kỳ nghỉ lễ. Có thể là nhu cầu sản phẩm quà tặng cao hơn trong quý cuối năm. Ngoài ra, có thể có khả năng khách hàng bán lẻ có xu hướng tích trữ nguồn cung cấp quà tặng trong thời gian này vì giá sản phẩm có thể tương đối giảm.

b) Không có mối tương quan rõ ràng giữa các ngày trong tháng và số lượng mua hàng.

c) Hầu như không có mua hàng vào các ngày thứ bảy hàng tuần (vì không hiện trong biểu đồ). Sức mua tăng dần từ chủ nhật tuần trước đến thứ 2, thứ 3 tuần sau và đạt đỉnh tại thứ 5.

d) Việc mua hàng chỉ được thực hiện trong khoảng thời gian từ 7 giờ sáng đến 8 giờ tối, giờ mua hàng có thể trùng với giờ bán hàng của các cửa hàng bán lẻ.

## Chuyển đổi dữ liệu

Do những bản ghi thu được ở trên được phân chia ở mức sản phẩm (nhiều bản ghi thuộc cùng một hóa đơn) nên với mục tiêu phân tích, ta tiến hành gộp các hóa đơn theo ngày; đồng thời, bổ sung giá trị cho những ngày không bán được đơn nào (giá trị bằng 0). Ta thu được kết quả sau :

Table

Description automatically generated

Như vậy, ta đã thu được chuỗi thời gian là tổng số hóa đơn bán được theo ngày, trong khoảng thời gian từ 1/12/2010 đến 9/12/2011. Chuỗi này gồm 374 bản ghi, tương ứng với mỗi ngày.

# Visualize data

Chuỗi thời gian thu được có dạng như sau :

Chart

Description automatically generated

Ta có thể thấy những nhận xét ở mục 2.2 là khá đúng, nhất là khi quan sát từng tháng (ví dụ: tháng 1/2011, 2/2011, 3/2011, 6/2011)

Chart, scatter chart

Description automatically generated

Để rõ ràng hơn, ta Phân tích chuỗi thành **3 thành phần** :

* Xu hướng (**trend**): Sự liên tục của các giá trị tăng hoặc giảm trong chuỗi thời gian đã cho
* Thời vụ (**seasonality**): Chu kỳ lặp lại trong một khoảng thời gian cụ thể (ngày, tuần, tháng, v.v.) trong chuỗi thời gian đã cho
* Phần còn lại (**residual**): Đây là các giá trị bất thường thể hiện sự bất thường ngẫu nhiên của các giá trị trong chuỗi thời gian đã cho.

Background pattern

Description automatically generated with low confidence

Nhận thấy, chuỗi thời gian đã cho:

* Có xu hướng giảm mạnh vào đầu năm mới, tăng mạnh về cuối năm.
* Có chu kỳ lặp lại theo tuần.

# Kiểm tra tính ổn định của chuỗi

## Chuỗi ổn định

Một điều quan trọng khi làm việc với chuỗi thời gian đó là ta phải đảm bảo chuỗi thời gian dùng để phân tích là chuỗi ổn định (stationary time series). Bởi vì chỉ với chuỗi ổn định, ta mới có thể tin tưởng vào kết quả phân tích hay dự đoán của mô hình.

Chuỗi ổn định là chuỗi trong đó giá trị trung bình và phương sai đều không đổi theo thời gian. Nói cách khác thuộc tính của nó không phụ thuộc vào thời gian mà chuỗi được quan sát. Một chuỗi thời gian ổn định, cần thỏa mãn các tiêu chí sau :

* Giá trị trung bình không đổi
* Phương sai không đổi
* Cấu trúc tự tương quan không đổi (tự tương quan là mối tương quan giữa giá trị hiện tại và giá trị trong quá khứ)
* Không có thành phần chu kỳ

## Kiểm tra tính ổn định của chuỗi

Ta kiểm tra tính ổn định bằng giả thuyết **Augmented Dickey-Fuller Test (Kiểm tra ADF). Đây là một bài kiểm tra phổ biến để xác định chuỗi có ổn định hay không. Cụ thể, nó sẽ kiểm tra 2 giả thuyết**

* **Giả thuyết rỗng \_ H0 : Chuỗi không ổn định**
* **Giả thuyết thay thế \_ H1 : Chuỗi ổn định**

**Kết quả kiểm tra là một giá trị p-value. Nếu :**

* **p-value ≤ 0.05, ta có thể bác bỏ H0 và kết luận chuỗi đã cho ổn định.**
* **p-value > 0.05, thì không thể bác bỏ H0 và kết luận chuỗi đã cho không ổn định.**

Sau khi kiểm tra chuỗi đã cho, ta thu được kết quả:

Graphical user interface, application

Description automatically generated

Do chuỗi không ổn định, nên ta cần thực hiện một số bước Differences để chuyển chuỗi về ổn định.

## Chuyển chuỗi đã cho về chuỗi ổn định

Có rất nhiều cách để chuyển một chuỗi đã cho về dạng ổn định : chuyển một cách tuyến tính (first difference, second difference) hoặc chuyển bằng cách phi tuyến ( lấy căn bậc hai hay logarith giá trị của chuỗi).

Ta thấy hình dạng chuỗi đã cho có dạng tuyến tính, nên ta thực hiện chuyển bằng cách tuyến tính.

Việc thực hiện các lần Differences cần được tối thiểu hóa và thực hiện có thứ tự.

Ví dụ ở chuỗi đã cho, ta nhận thấy chuỗi có tính chu kỳ nên ta áp dụng “Seasonal Difference” trước, sau đó nếu chuỗi vẫn chưa ổn định thì mới áp dụng tiếp “First Difference” rồi “Second Difference”

Sau khi áp dụng “Seasonal Difference”, ta kiểm tra tính ổn định của chuỗi và có thể xem xét ảnh hưởng của phép Difference đến chuỗi thế nào (bằng cách phân tích lại chuỗi thành 3 thành phần như đã làm bên trên).

A picture containing shape

Description automatically generated

A picture containing text, antenna

Description automatically generated

Có thể thấy chuỗi đã ổn định (qua kiểm tra ADF) và mô phỏng chuỗi cũng không còn thể hiện xu hướng rõ ràng hay tính chu kỳ.

Như vậy, chuỗi đã sẵn sàng để phân tích.

# Xây dựng model ARIMA

Bởi vì chuỗi có tính chu kỳ nên ta sẽ sử dụng model Seasonal ARIMA (SARIMA). Để xây dựng mô hình SARIMA phù hợp, ta cần chọn bộ tham số (p,d,q)x(P,D,Q,s). Trong đó,

* (**p,d,q**) : bậc của model AR, số lần thực hiện Differences , bậc của model MA
* (**P,D,Q,s**) : có ý nghĩa tương tự như trên nhưng theo chu kỳ và s đại diện cho tính chu kỳ (*periodicity*)

## Mô hình SARIMA sơ bộ

Ta có thể xác định bộ tham số cho mô hình SARIMA sơ bộ bằng biểu đồ của các Hàm tự tương quan (ACF) và Hàm tự tương quan từng phân (PACF).

Trong đó, biểu đồ ACF giúp xác định giá trị q và Q; còn PACF giúp xác định p và P; cặp giá trị (d, D) được xác định ở bước chuyển chuỗi về dạng ổn định ở bước trước. Ta chỉ quan tâm đến những vị trí có giá trị vượt ngoài khoảng xanh (vị trí có ý nghĩa về mặt thống kê).

Ta có thể gọi các hàm có sẵn trong thư viện statsmodels để vẽ hai biểu đồ này. Chuỗi thời gian đã cho có 2 biểu đồ tương ứng là :

Chart, histogram

Description automatically generatedA picture containing chart

Description automatically generated

Nhìn vào biểu đồ ACF/PACF thấy :

* **PACF** có mức tăng đột biến âm đáng kể tại các vị trí **lag(độ trễ) 7, 14, 21**,... (đây là suy giảm cấp mũ \_ *expotential decay*) ; và bên cạnh đó ở biểu đồ **ACF** có mức tăng đột biến âm đáng kể tại **lag 7**, qua đó cho thấy giá trị **Q = 1** (thể hiện tính thời vụ theo tuần) và giá trị **P = 0**
* **ACF** có mức tăng đột biến dương đáng kể tại **lag 1, 2 và 3** nên **q = 3**
* **PACF** có mức đột biến dương đáng kể tại **lag 1** nên **p = 1**

Như vậy, với tham số khởi đầu của mô hình SARIMA, ta chọn bộ tham số tiềm năng là **SARIMA(1, 0, 3)x(0, 1, 1, 7)**

Tiếp theo, ta sẽ triển khai mô hình SARIMA sơ bộ sử dụng các thông số xác định ở trên và đánh giá chất lượng mô hình qua thành phần residual.

Kết quả đánh giá thể hiện ở 2 biểu đồ ACF/PACF mới như sau:

Chart

Description automatically generated with low confidence

A picture containing histogram

Description automatically generated

Trong đồ thị trên,

* Không có đột biến đáng kể nào ngoại trừ ở độ trễ lag 1 trong các đồ thị ACF và PACF của phần dư của mô hình.
* Các thông số này là điểm khởi đầu có thể chấp nhận được đối với mô hình và có thể được cải thiện hơn nữa bằng cách sử dụng tìm kiếm lưới (Grid Search)

## Chọn bộ tham số phù hợp nhất

Sau khi có những tham số khởi tạo ban đầu, ta có thể sử dụng thư viện pmdarima để xác định bộ tham số tốt nhất cho dữ liệu

A picture containing scatter chart

Description automatically generated

Vì đây là mô hình SARIMA nên ta cần chọn m (có ý nghĩa như s), seasonal = True và 1 ≤ D ≤2 (luôn phải áp dụng “Seasonal Difference”). Hàm sẽ tự động chọn bộ tham số tốt nhất sao cho có giá trị AIC (*Akaike information criteri*on) nhỏ nhất.

Kết quả thu được như sau :

Table

Description automatically generated

Sau khi thực hiện tìm kiếm, bộ tham số tốt nhất cho mô hình là **SARIMA(3,0,0)x(1,1,1,7)** vì có chỉ số AIC thấp nhất. Bây giờ ta tiến hành train và đánh giá mô hình.

## Train và đánh giá mô hình

* Walk-forward train/test

Với phân tích chuỗi thời gian, ta không thể chia tập Train/Test bằng cách split thông thường hay dùng cross validation thông thường. Cách tiếp cận tốt hơn - và trực quan hơn nhiều - là mô phỏng các mô hình theo trình tự "đi bộ về phía trước", định kỳ đào tạo lại mô hình để kết hợp tất cả dữ liệu có sẵn tại thời điểm đó (Như hình minh họa bên dưới)

Graphical user interface

Description automatically generated

Các dữ liệu mới liên tục được được cập nhật vào làm dữ liệu huấn luyện mỗi khi mô hình cần được huấn luyện lại.

Ở đây, nhóm chia tập train/test theo tỷ lệ 3-1 và thực hiện walk-forward validation trên từng ngày trong tập test một.

* Đánh giá model với residuals plot

Kết quả thu được của mô hình của nhóm như sau :

Graphical user interface

Description automatically generated

Các biểu đồ về phần residuals của chuỗi khá đẹp. Trong đó,

**Biểu đồ “Histogram plus estimated density”**: Đường KDE màu cam theo sát với đường N(0,1) màu xanh lá. Đây là một dấu hiệu tốt cho thấy phần dư được phân phối bình thường.

**Biểu đồ “Standardized residual for y”**: Phần còn lại (residual) theo thời gian không hiển thị bất kỳ tính thời vụ rõ ràng nào và có vẻ là white noise.

**Biểu đồ “Normal Q-Q”**: Cho thấy phân phối có thứ tự của phần dư (chấm màu xanh) tuân theo xu hướng tuyến tính của các mẫu được lấy từ phân phối chuẩn với N (0, 1). Đây là dấu hiệu cho thấy phần dư được phân phối bình thường.

**Biểu đồ “Correlogram”**: Cho thấy phần dư của chuỗi thời gian có mối tương quan thấp với các lag của chính nó. Đây là một yếu tố tốt.

Về mặt sơ bộ thì model của nhóm có vẻ tốt. Tiếp theo, sẽ đánh giá chi tiết về các độ đo accuracy của model.

* Đánh giá qua Accuracy Metrics

Ở đây, nhóm áp dụng độ đo là: Root Mean Squared Error (RMSE). Nó cho biết mức độ phù hợp của mô hình với dữ liệu – mức độ gần của các điểm dữ liệu quan sát được với các giá trị dự đoán của mô hình.



Kết quả thu được cho thấy mô hình chưa được hoàn hảo lắm. Có thể cần nhiều dữ liệu hơn để thu được nhiều thông tin về chuỗi thời gian hơn.

## Dự đoán

Kết quả dự đoán được mô tả ở hình dưới : Đường màu xanh là giá trị thực, đường màu đỏ là kết quả dự đoán của mô hình. Nhận thấy, các xu hướng tăng giảm của dự đoán trùng khớp với giá trị thực, tuy nhiên độ chính xác chưa thực sự hoàn hảo về số lượng hóa đơn được dự đoán. Ví dụ, trong những ngày đầu tháng 12, số hóa đơn mua hàng có giảm mạnh, tuy kết quả dự đoán cho thấy số hóa đơn có xu hướng giảm nhưng vẫn dừng ở mức cao (nhiều hóa đơn).

Histogram

Description automatically generated with low confidence

# Cải tiến mô hình

* Có thể cung cấp thêm dữ liệu theo nhiều năm để có thể có cái nhìn chính xác hơn về doanh thu của cửa hàng.