TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 19, Số 12 (2022): 2064-2075
Website: https://journal.hcmue.edu.vn
ISSN: 2734-9918

HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 19, No. 12 (2022): 2064-2075
https://doi.org/10.54607/hcmue.js.19.12.3334(2022)

Bài báo nghiên cứu

ÁP DỤNG MỘT VÀI PHƯƠNG PHÁP MÁY HỌC VÀO BÀI TOÁN DỰ BÁO THEO CHUỖI THỜI GIAN

Nguyễn Thị Hồng Thảo[1*], Đào Minh Châu[1], Vũ Thanh Nguyên[1], Phù Phước Huy[2]

[1]Trường Đại học Công nghiệp Thực phẩm Thành phố Hồ Chí Minh, Việt Nam

[2]Viện Công nghệ Thông tin – Viện Khoa học và Công nghệ Quân sự, Việt Nam

[*]Tác giả liên hệ: Nguyễn Thị Hồng Thảo – Email: thaonth@hufi.edu.vn

Ngày nhận bài: 13-9-2021; ngày nhận bài sửa: 14-01-2022; ngày duyệt đăng: 13-3-2022

TÓM TẮT

Vàng đóng vai trò cực kì quan trọng trong nền kinh tế, đặc biệt là đánh giá chỉ số lạm phát của nền kinh tế. Do đó, trong phạm vi nghiên cứu của bài báo này, chúng tôi áp dụng một vài phương pháp máy học vào bài toán dự báo dữ liệu chuỗi thời gian bằng mô hình dự báo ARIMA và SARIMA để dự báo giá vàng trong tương lai. Chúng tôi đã tiến hành một loạt thử nghiệm và đánh giá mô hình cũng như phân tích các yếu tố biến động của chuỗi thời gian để tìm ra kết quả tối ưu nhất để tăng hiệu suất dự báo.

Từ khóa: dự báo giá vàng; máy học; mô hình ARIMA; mô hình SARIMA

1. Giới thiệu

Trong những năm gần đây, việc nghiên cứu áp dụng máy học đã được ứng dụng rộng rãi như dự đoán kết quả bầu cử, phân tích dự báo đầu tư, lên kế hoạch phát triển kinh tế, đánh giá kết quả phát triển trong một số lĩnh vực… Công nghệ Máy học là một trong những phương pháp chính trong khai phá dữ liệu. Nó được sử dụng trong tiến trình khám phá tri thức. Các phương pháp máy học hoạt động trên các dữ liệu có đặc tả thông tin. Máy học mang đến nhiều lợi ích: máy học giúp xử lí rất nhiều thông tin đến từ nhiều nguồn khác nhau và dự báo các thông tin đó; ở những nơi không có chuyên gia, máy học có thể giúp tạo ra được các quyết định từ các dữ liệu có được; các thuật toán máy học có thể giúp xử lí khi dữ liệu không đầy đủ, không chính xác...

Do Máy học mang đến nhiều lợi ích cho con người, nên chúng tôi chọn nghiên cứu vài phương pháp máy học vào bài toán dự báo theo chuỗi thời gian, sử dụng phương pháp ARIMA (Autoregressive Integrated Moving Average) và SARIMA để dự báo trong tương lai. Chúng tôi chọn dữ liệu để phân tích và dự báo là giá vàng. Vàng là một trong những mặt hàng quan trọng đánh giá chỉ số lạm phát của một quốc gia, được xem như một tài sản an

Cite this article as: Nguyen Thi Hong Thao, Dao Minh Chau, Vu Thanh Nguyen, & Phu Phuoc Huy (2022). Using some machine learning methods for time series forecasting. *Ho Chi Minh City University of Education Journal of Science, 19*(12), 2064-2075.

toàn trước những biến động về kinh tế, chính trị, xã hội hoặc khủng hoảng tiền tệ. Do đó, dự báo về hành vi giá vàng rất cần thiết. Từ đó, chúng tôi đã tiến hành một loạt thử nghiệm và đánh giá hiệu quả của 2 phương pháp ARIMA và SARIMA cũng như phân tích các yếu tố biến động của chuỗi thời gian để tìm ra kết quả tối ưu nhất kết quả dự báo.

Với những vai trò quan trọng của vàng đã được phân tích, nên đã có nhiều công trình nghiên cứu về ảnh hưởng của vàng đối với nền kinh tế, biến động giá cũng như đưa ra các dự báo về giá vàng. Hiện nay, với sự phát triển của máy học vấn đề dự báo được thuận tiện hơn và có những kết quả dự báo cơ bản đáp ứng được mục đích của từng tổ chức. Nghiên cứu ảnh hưởng tin tức kinh tế đến giá cả hàng hóa, trong đó có giá vàng trong công trình của Williams (2018). Hay công trình nghiên cứu dự báo giá vàng bằng cách sử dụng các thuật toán máy học của Sami (2017).

Các công trình đã đưa ra được một số mô hình dự báo giá vàng cũng như giá một số mặt hàng khác bằng các thuật toán máy học. Đối với công trình nghiên cứu giá vàng thì thường cho kết quả chính xác với giá trị thực nghiệm của tác giả đưa ra, nhưng khi thử với những giá trị khác thì cho kết quả có độ sai lệch. Tuy nhiên, kết quả các công trình nghiên cứu này là cơ sở để nhóm tác giả nghiên cứu đánh giá so sánh thuật toán ARIMA và SARIMA trong dự báo giá vàng.

2. Cơ sở lí thuyết và phương pháp nghiên cứu

2.1. Cơ sở lí thuyết

2.1.1. Phương pháp ARIMA

Phương pháp ARIMA là mô hình dự báo theo chuỗi thời gian được phát minh bởi Box-Jenkin, kết quả dự báo phụ thuộc chuỗi giá trị trong quá khứ. Phương pháp ARIMA phân tích định lượng tính tương quan giữa các dữ liệu quan sát để đưa ra mô hình dự báo thông qua các giai đoạn nhận dạng mô hình, ước lượng các tham số từ dữ liệu quan sát và kiểm tra các tham số ước lượng để tìm ra mô hình thích hợp. Phương pháp ARIMA được thực hiện cho dữ liệu dừng (hay dữ liệu đã loại bỏ yếu tố xu thế). ARIMA được kết hợp bởi 3 thành thành phần chính ARIMA(p,q,d):

AR (p): Autogressive (tự hồi quy)

I (q): Integrated (chuỗi dừng)

MA (d): Moving Arverage (trung bình động).

Phương pháp ARIMA được thực hiện với dữ liệu dừng hay chuỗi thời gian dừng. Dữ liệu dừng là dữ liệu dao động xung quanh một giá trị trung bình cố định trong dài hạn, dữ liệu có giá trị phương sai xác định không thay đổi theo thời gian, dữ liệu có một giản đồ tự tương quan với các hệ số tự tương quan sẽ giảm dần khi độ trễ tăng lên. Nếu một chuỗi dừng, thì giá trị trung bình, phương sai sẽ giống nhau ở tại mọi thời điểm.

Tự hồi quy, quá trình phụ thuộc vào tổng trọng số của các giá trị quá khứ và số hạng nhiễu ngẫu nhiên:

$$Y_t = \varphi_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_n Y_{t-n} \quad (1)$$

Y_t là quan sát dừng hiện tại, Y_{t-1} ,Y_{t-2}, là quan sát dừng quá khứ; φ_0 là hằng số; φ_1,φ_2 … là các tham số phân tích hồi quy.

Mô hình tự tương quan bậc p, viết tắt là AR(p).

Trung bình động bậc q, viết tắt là MA(q): quá trình được mô tả bằng tổng trọng số của các ngẫu nhiên hiện hành có độ trễ:

$$Y_t = \mu + u_t + \theta_1 u_{t-1}$$

Trong đó, μ là giá trị trung bình của quá trình, u_t là số hạng nhiễu ngẫu nhiên, θ_1 là hệ số ước lượng, và u_{t-1} là sai số ở giai đoạn *t-1* .

Mô hình *MA* là Y_t phụ thuộc vào giá trị của sai số hiện tại và sai số quá khứ (tại thời điểm *t* và *t-1). *Y_t phụ thuộc vào giá trị sai số trước đó, Y_t không phụ thuộc giá trị trễ trong mô hình *AR*.

Kiểm định tính dừng: để kiểm định tính dừng chúng tôi sử dụng giản đồ tự tương quan

Giản đồ tự tương quan là một đồ thị biểu diễn mối quan hệ giữa hệ số tự tương quan bậc *k* với độ trễ *k* tương ứng. Hệ số tự tương quan bậc *k* (kí hiệu là r_k) được xác định theo công thức sau:

$$\rho_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})\,(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$$

Phương trình trên được gọi là *hàm tự tương quan*, kí hiệu là *ACF*.

Đối với dữ liệu mẫu, ước lượng được hệ số tự tương quan mẫu theo công thức sau:

$$r_{\mathrm{k}} = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})\,(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$$

trong đó:

$\bar{Y}$ là giá trị trung bình mẫu của chuỗi Y_t

k là độ trễ

n là số quan sát của mẫu.

Vì r_k là một hàm tự tương quan nên nó có tính chất của hàm tự tương quan ρ_k, r_k là giá trị ước lượng của ρ_k nên có những điểm khác biệt. Kết quả mô hình *MA(q)*:

$$Var(r_k) = \frac{1}{n}[1 + 2\sum_{j=1}^{q}\rho_j^2] \quad với\ k > p$$

Đối với một chuỗi thời gian, có thể thay ρ bởi *r*, lấy căn bậc hai, và nhận độ lệch chuẩn ước lượng của *r* hay sai số chuẩn của r_k cho những độ trễ lớn. Trong trường hợp tổng quát, giá trị mong đợi của hàm tự tương quan mẫu giống như giá trị thật của hàm tương quan. Vì vậy, trong thực tế, r_k có những phần không giống với ρ_k.

Hàm tự tương quan riêng phần –PACF: Mô hình *MA(q)* có hàm tự tương quan bằng không cho những độ trễ lớn hơn *q*, hàm tự tương quan lấy mẫu là một cách tốt để xác định bậc của quá trình *MA*. Tuy nhiên, hàm tự tương quan của *AR(p)* thì không bằng không sau

một số độ trễ nhất định – chúng giảm dần thay vì bằng không. Vì vậy, một hàm khác cần thiết để xác định bậc của mô hình hồi quy. Hàm này được định nghĩa là sự tương quan giữa Y_t và Y_{t-k} sau khi loại bỏ sự ảnh hưởng của ảnh hưởng của các biến bên trong Y_{t-1}, Y_{t-2}, Y_{t-3},…,Y_{t-k+1}. Hệ số này gọi là hệ số tự tương quan riêng phần ở độ trễ k và được kí hiệu bằng φ_{kk} .

Cách tính giá trị của hàm tự tương quan riêng phần:

$$\varphi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1} \varphi_{k-1,j}\rho_{k-j}}{1 - \sum_{j=1}^{k-1} \varphi_{k-1,j}\rho_j}$$

trong đó $\varphi_{k,j} = \varphi_{k-1,j} - \varphi_{kk}\varphi_{k-1,k-j} \quad với\ j = 1,2,\dots,k-1$

Với các công thức đã được đưa ra, chúng tôi thực hiện quy trình thuật toán ARIMA:

Bước 1. Tính ACF và PACF của dữ liệu gốc để kiểm tra xem chuỗi gốc có dừng hay không. Nếu dừng, chuyển sang Bước 3

Bước 2. Lấy log rồi lấy sai phân bậc một của dữ liệu gốc hoặc lấy sai phân trực tiếp nếu dữ liệu ít biến động, sau đó tính ACF và PACF của dữ liệu chuyển đổi

Bước 3. Phân tích giản đồ tự tương quan để xác định các mô hình có thể có

Bước 4. Ước lượng các mô hình dự kiến

Bước 5. Đối với mỗi mô hình được ước lượng:

Kiểm tra hệ số của độ trễ cao nhất có ý nghĩa thống kê hay không. Nếu không giảm bớt độ trễ của p và q.

Kiểm tra ACF và PACF đối với phần dư.

So sánh các sai số dự báo.

Phân tích đồ thị phần dư.

Phân tích đồ thị giá trị dự báo và giá trị thực tế.

Khi đánh giá các mô hình ARIMA, nên so sánh giữa các mô hình.

Bước 6. Nếu có thay đổi so với mô hình gốc, quay lại Bước 4.

Mô hình ARIMA được áp dụng đối với những chuỗi thời gian dừng. Hầu hết, dữ liệu chuỗi thời gian là các chuỗi có xu thế (giá trị trung bình của Y_t trong năm này có thể khác năm kia hoặc các chuỗi không dừng). Để áp dụng được mô hình ARIMA thì các chuỗi thời gian phải khử các yếu tố xu thế trong các chuỗi dữ liệu gốc thông qua quy trình lấy sai phân.

2.1.2. Phương pháp SARIMA

Phương pháp dự báo SARIMA được phát triển từ phương pháp ARIMA khi chuỗi dữ liệu có tính mùa vụ. Các độ trễ theo mùa của y_t trong mô hình ARIMA để trở thành mô hình SARIMA. Mô hình SARIMA có dạng tổng quát: (p,d,q)(P,D,Q)s, trong đó, P là AR và Q là MA, D là bậc sai phân có tính mùa, s là số kì trong một vòng chu kì (s = 12 khi chuỗi dữ liệu theo tháng). Phân tích ACF và PACF tại các trễ là bội số của độ dài mùa s giúp kết luận các giá trị P, Q phù hợp cho mô hình. Đối với thành phần mùa MA, biểu đồ ACF cho thấy một

đỉnh nhọn ở các trễ mùa, còn đối với thành phần mùa AR thì biểu đồ PACF thể hiện đỉnh nhọn này.

Tiến trình xây dựng thuật toán SARIMA thực hiện qua 6 bước sau:

Bước 1. Kiểm định tính dừng, tính mùa vụ của dữ liệu gốc Các chuỗi được kiểm định tính dừng của dữ liệu thời gian bằng phương pháp kiểm định đơn vị (unit root test) – kiểm định Dickey-Fuller (ADF).

Bước 2. Chuyển dữ liệu sang sai phân. Việc chuyển đổi được thực hiện bằng cách tính sai phân giữa các giá trị quan sát dựa vào giả định các phần khác nhau của các chuỗi thời gian. Nghiên cứu thực nghiệm khi chuyển chuỗi dữ liệu sang sai phân nên dừng lại sai phân bậc 2.

Bước 3. Ước lượng các giá trị tham số của mô hình sau khi chuyển thành chuỗi sai phân bậc d để đạt tính dừng.

Bước 4. Xây dựng mô hình SARIMA.

Bước 5. Lựa chọn mô hình và kiểm định mô hình các tham số của mô hình sẽ được ước lượng bằng MLE.

Bước 6. Dự báo sau khi kiểm định sai số, nếu mô hình là phù hợp, sẽ được sử dụng vào việc dự báo. Nghiên cứu tính chỉ tiêu % sai số của dự báo so với giá trị quan sát được và được tính bằng sự chênh lệch giữa giá trị quan sát so với giá trị dự báo.

2.2. Phương pháp nghiên cứu

Phương pháp định lượng: Dự báo chuỗi thời gian dựa trên phân tích giá vàng và thời gian thay đổi theo từng ngày.

Thu thập và xử lí số liệu số: Chuỗi dữ liệu sử dụng 748 dòng quan sát từ 02/01/2019 đến 09/06/2021. Mỗi dòng quan sát bao gồm giá vàng và thời gian liên tục mỗi ngày lưu trên phần mềm Excel

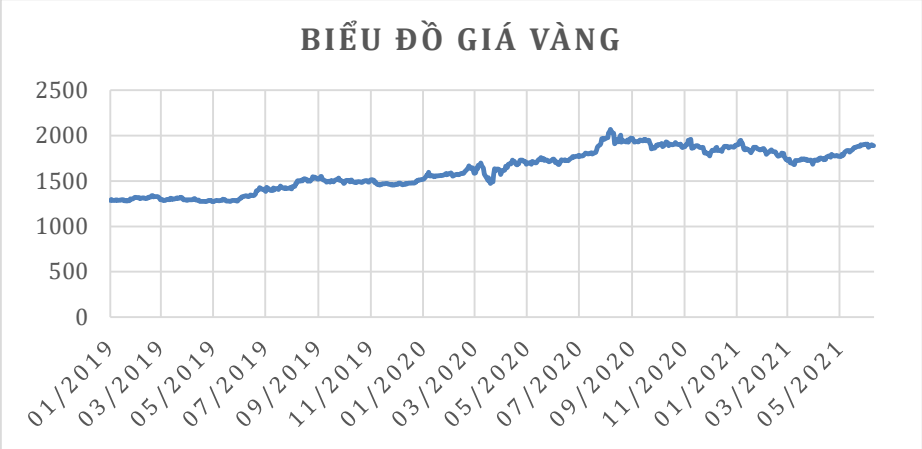
Nguồn số liệu: Lấy từ nguồn dữ liệu thực tế trên https://www.tradingview.com

3. Kết quả thực nghiệm

3.1. Phương pháp ARIMA

Với các phương pháp đề xuất, chúng tôi đã tiến hành thử nghiệm trên chuỗi dữ liệu gồm 748 dòng quan sát từ ngày 2 tháng 01 năm 2019 đến ngày 9 tháng 6 năm 2021, được lấy từ các nguồn dữ liệu thực tế thông qua trang web Trading View (www.tradingview.com). Dự đoán chuỗi thời gian dựa trên phân tích giá vàng và thời gian thay đổi từng ngày. Mỗi dòng quan sát bao gồm giá vàng và thời gian liên tục trong ngày được lưu trữ trong phần mềm Excel.

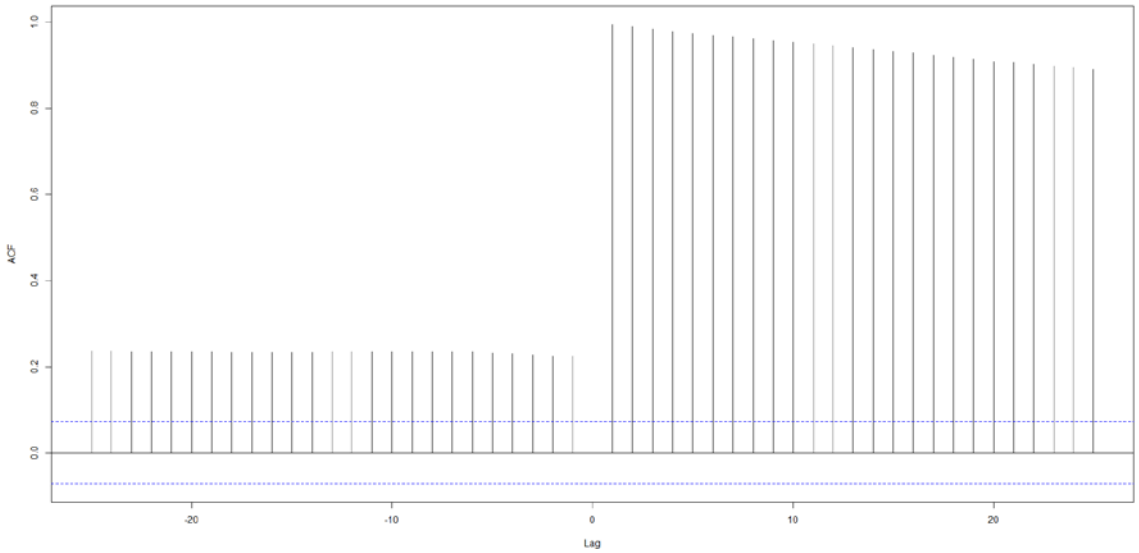
Các kết quả phân tích dự liệu dựa trên phần mềm Rstudio. Trước khi tiến hành phân tích dữ liệu, biểu diễn biểu đồ giá vàng trong thời gian quan sát.



Hình 1. *Biểu đồ giá vàng từ 02/01/2019 đến 10/06/2021*

Kết quả biểu đồ giá vàng Hình 1 từ 02/01/2019 đến 10/06/2021, ta thấy giá vàng là chuỗi dữ liệu không dừng vì có yếu tố xu hướng và yếu tố mùa vụ. Để kiểm định lại nhận định trên ta tiến hành phân tích biểu đồ giản tự tương quan ACF của tập dữ liệu giá vàng

Biểu đồ tương quan của dữ liệu vàng (02/01/2019 đến 10/06/2021)

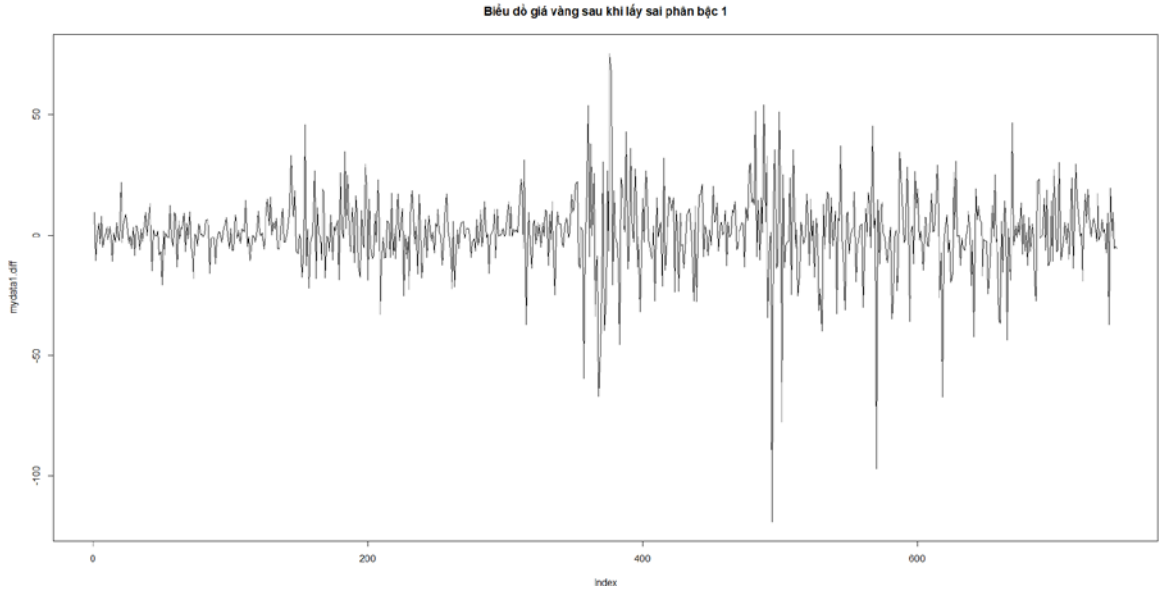


Hình 2. *Biểu đồ tương quan tính độ trễ*

Dựa vào biểu đồ tương quan tính độ trễ, chúng tôi thấy hệ số ACF của tập dữ liệu giảm rất chậm và liên tục nằm ngoài giới hạn tin cậy. Kết luận dữ liệu giá vàng là chuỗi không dừng.

Do dữ liệu giá vàng là chuỗi không dừng nên chúng tôi cần loại bỏ tính xu hướng và mùa vụ của tập dữ liệu bằng cách tiến hành lấy sai phân hoặc log trước khi vận dụng phương pháp ARIMA.

Biểu đồ giá vàng sau khi lấy sai phân bậc 1

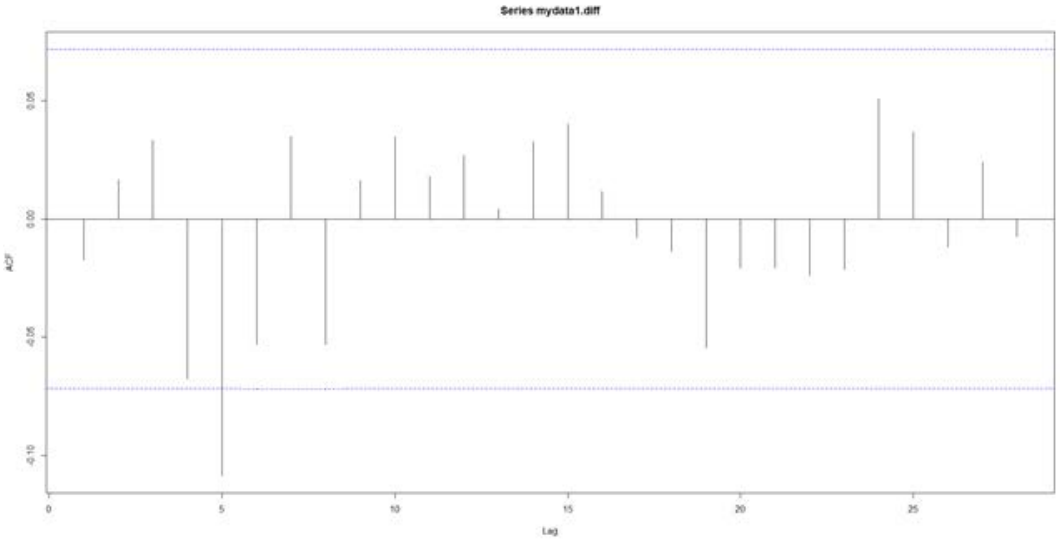


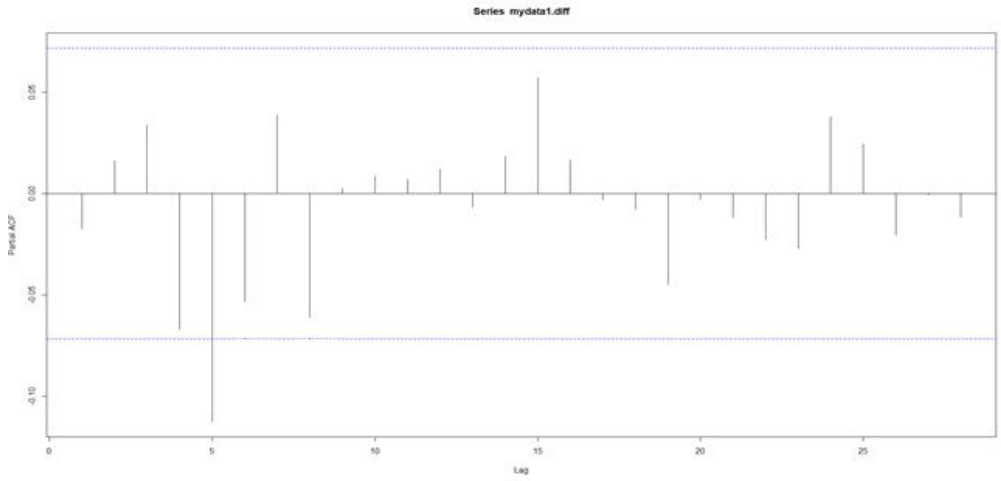
Hình 3. *Biểu đồ sai phân bậc 1*

Hình 3 biểu diễn biểu đồ giá vàng sau khi lấy sai phân bậc 1, chúng tôi thấy tất cả các giá trị dao động xung quanh giá trị [-100,50], có vài điểm nằm ngoài khoảng trên. Sau khi lấy sai phân lần 1 chúng tôi thấy dữ liệu có thể có tính dừng.

Chúng tôi xét biểu đồ tương quan ACF và tương quan riêng phần PACF của dữ liệu giá vàng sau khi lấy sai phân bậc 1

Biểu đồ tương quan và tương quan riêng phần của dữ liệu giá vàng áp dụng cho mô hình ARIMA





Hình 4. *Biểu đồ tương quan và tương quan riêng*

Biểu đồ hàm tự tương quan sai phân bậc 1 của chuỗi dữ liệu giá vàng cho thấy chuỗi dữ liệu sau khi lấy sai phân đã loại bỏ thuộc tính mùa vụ và xu hướng, như vậy đã đạt chuỗi dừng. Kết luận *q=1*.

Biểu đồ PACF có độ trễ là 1. Vậy *p=1*.

Một số mô hình ARIMA có thể ứng với các giá trị p=1 q=1,d=1

ARIMA(2,1,2)

ARIMA(1,1,0)

ARIMA(0,1,1)

ARIMA(0,1,0)

ARIMA(1,1,1)

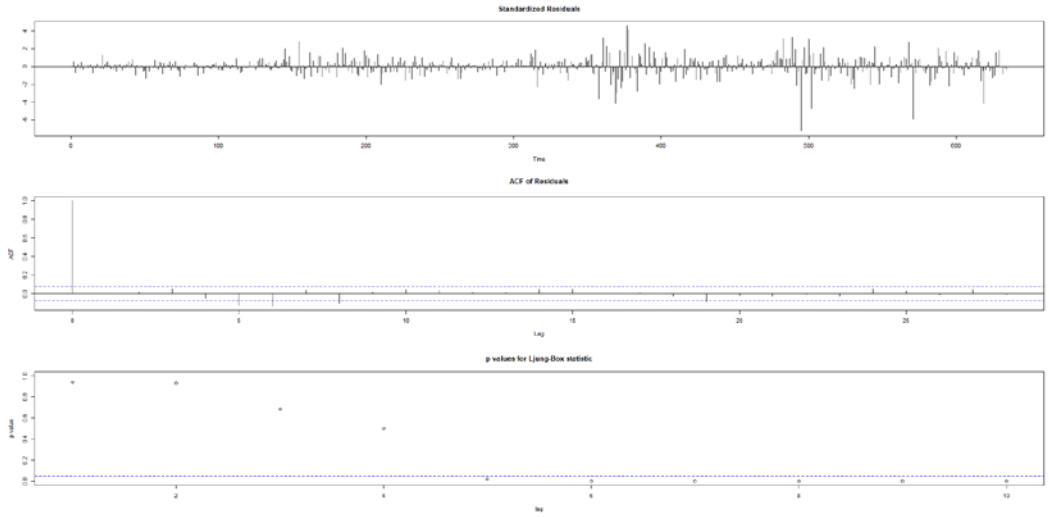
Chúng tôi tiếp tục kiểm tra và so sánh các mô hình ARIMA để tìm mô hình ARIMA tối ưu nhất.

```
Fitting models using approximations to speed things up...

ARIMA(2,1,2) with drift         : 5337.634
ARIMA(0,1,0) with drift         : 5335.344
ARIMA(1,1,0) with drift         : 5338.015
ARIMA(0,1,1) with drift         : 5337.283
ARIMA(0,1,0)                    : 5335.202
ARIMA(1,1,1) with drift         : 5339.947

Now re-fitting the best model(s) without approximations...

ARIMA(0,1,0)                    : 5340.798

Best model: ARIMA(0,1,0)

Series: seriesdata2
ARIMA(0,1,0)

sigma^2 estimated as 269.4:  log likelihood=-2669.4
AIC=5340.79   AICc=5340.8   BIC=5345.24
```

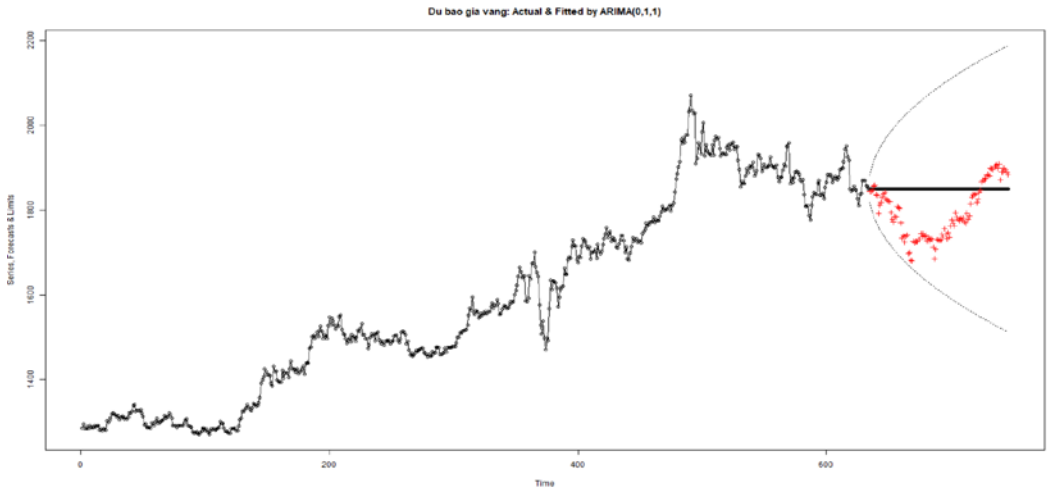
Hình 5. *So sánh AIC của các mô hình ARIMA đề xuất với dữ liệu vàng*

Theo như kết quả so sánh giá trị của chỉ tiêu AIC của các mô hình đề xuất thì mô hình thứ 5 (ARIMA(0,1,0)) là mô hình có giá trị AIC thấp nhất, điều này có nghĩa là mô hình này phù hợp với tập dữ liệu hơn so với các mô hình còn lại.

Biểu đồ giá trị phần dư, tương quan và kiểm định p-values
của thống kê Ljung-Box theo mô hình ARIMA(0,1,0)



Hình 6. *Biểu đồ phần dư, tương quan và giá trị kiểm định*
Biểu đồ so sánh giá trị thực tế và giá trị dự báo theo mô hình ARIMA(0,1,0)



Hình 7. *Biểu đồ so sánh giá trị thực tế và giá trị dự báo theo ARIMA*

Theo như kết quả từ Hình 7 thì các giá trị dự báo và giá trị thực tế có khoảng chênh lệch khá lớn.

3.2. Phương pháp SARIMA

Dựa vào phần biểu đồ tương quan ACF ở trên chúng tôi thấy biểu đồ thể hiện biến động tăng và giảm ở các độ trễ nằm trong giới hạn tin cậy, ngoại trừ độ trễ thứ 2 có giá trị của ACF gần với giới hạn tin cậy 95%. Biểu đồ hàm tự tương quan sai phân bậc 2 của chuỗi dữ liệu này cho thấy chuỗi dữ liệu sau khi lấy sai phân bậc 2 đã loại bỏ thuộc tính mùa vụ và xu hướng, như vậy đã đạt chuỗi dừng. Kết luận q+Q=1.

Phần biểu đồ tương quan riêng phần PACF, tại độ độ trễ 2 nằm ngoài phạm vi tin cậy. Chọn p+P=1.

Một số mô hình có thể ứng với các giá trị p+P=1,q+Q=1, d+D=1

Mô hình SARIMA(2,1,2)(0,1,0)$_{12}$

Mô hình SARIMA(0,1,0)(0,1,0)$_{12}$

Mô hình SARIMA(1,1,0)(0,1,0)$_{12}$

Mô hình SARIMA(0,1,1)(0,1,0)$_{12}$

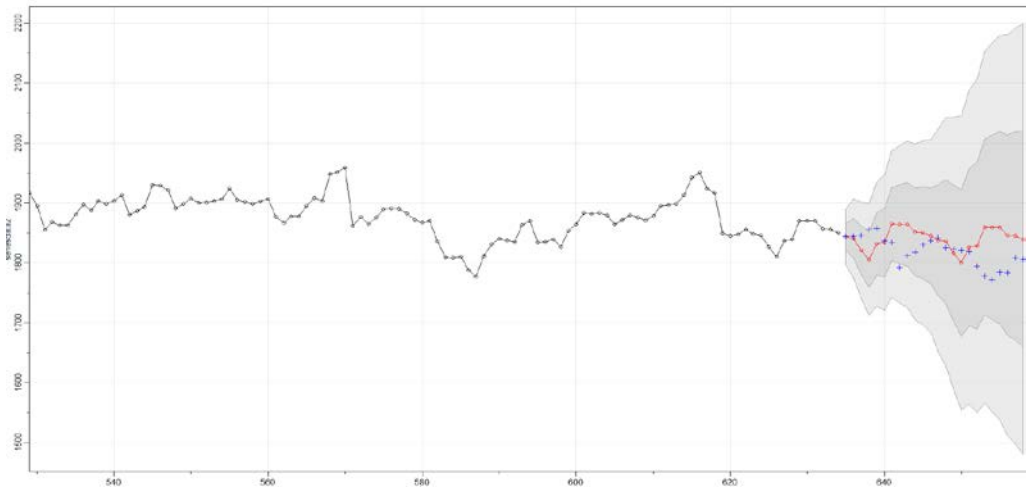
Mô hình SARIMA(1,1,1)(0,1,0)$_{12}$

```
 ARIMA(2,1,2)(0,1,0)[12]                    : Inf
 ARIMA(0,1,0)(0,1,0)[12]                    : 82.15163
 ARIMA(1,1,0)(0,1,0)[12]                    : 84.61297
 ARIMA(0,1,1)(0,1,0)[12]                    : 84.49562
 ARIMA(1,1,1)(0,1,0)[12]                    : 88.41251

 Best model: ARIMA(0,1,0)(0,1,0)[12]

Series: Train
ARIMA(0,1,0)(0,1,0)[12]

sigma^2 estimated as 84.06:  log likelihood=-39.85
AIC=81.71   AICc=82.15   BIC=82.11
```

Hình 8. *So sánh AIC của các mô hình SARIMA đề xuất với dữ liệu vàng*

Theo như kết quả so sánh giá trị của chỉ tiêu AIC của 5 mô hình đề xuất thì mô hình thứ 2 (tức là mô hình SARIMA(0,1,0)(0,1,0)$_{12}$) mô hình có giá trị AIC thấp nhất, điều này có nghĩa là mô hình này phù hợp với tập dữ liệu so với các mô hình còn lại.

Biểu đồ so sánh giá trị thực tế và giá trị dự báo theo mô hình SARIMA(0,1,0)(0,1,0)$_{12}$



Hình 9. *Kết quả dự báo theo mô hình SARIMA so với giá trị thực tế*

Theo như kết quả từ Hình 9 thì các giá trị dự báo và giá trị thực tế có khoảng chênh lệch ít.

3.3 Đánh giá và so sánh các kết quả thực nghiệm

Sau khi, chúng tôi chạy thực nghiệm dữ liệu giá vàng trên từng phương pháp ARIMA và SARIMA. Chúng tôi tổng hợp 2 phương pháp đánh giá dự báo giá vàng dữ liệu tính từ ngày 02/01/2019 đến 10/06/2021.

Đánh giá và so sánh kết quả dự báo của các phương pháp

Mô hình	MAE	MSE	RMSE	MAPE	MPE
Dữ liệu giá vàng từ ngày từ 02/01/2019 đến 10/06/2021					
ARIMA(0,1,0)	70.26160714	6911.57365357	83.13587465	0.03993752	-0.03185311
SARIMA $(0,1,0)(0,1,0)_{12}$	58.56428571	6005.85447714	77.49744820	0.03351040	-0.02915437

Trong đó, MAE: Sai số tuyệt đối trung bình (Mean Absolute Error); MSE: Sai số bình phương trung bình(Mean Squared Error); RMSE: Sai số bình phương trung bình gốc (Root Mean Squared Error); MAPE: Sai số phần trăm tuyệt đối trung bình (Mean Absolute Percentage Error); MPE: Sai số phần trăm trung bình (Mean Percentage Error); là các thước đo độ chính xác của một mô hình dự báo.

Đối với tập dữ liệu giá vàng tính từ ngày 02/01/2019 đến 10/06/2021 được sử dụng để dự báo theo các mô hình ARIMA và SARIMA thì theo như kết quả đánh giá dự báo thì mô hình SARIMA$(0,1,0)(0,1,0)_{12}$ có độ chính xác cao hơn, phù hợp để dự báo giá vàng tính từ ngày 02/01/2019 đến 10/06/2021. Từ đó cho thấy dự báo theo mô hình SARIMA là hiệu quả hơn so với mô hình ARIMA.

4. Kết luận

Trong bài báo này, chúng tôi sử dụng máy học là 2 phương pháp ARIMA và SARIMA ứng dụng vào bài toán dự báo giá vàng. Dữ liệu nghiên cứu của chúng tôi được thu thập từ thực tiễn được sử dụng để huấn luyện và thử nghiệm nhằm tăng độ chính xác cho quá trình phân tích dự báo khi sử dụng các phương pháp máy học, đầy đủ và liên tục theo chu kỳ 24 tháng. Theo kết quả thực nghiệm thì mô hình ARIMA (0,1,0) hợp lí nhất để dự báo giá vàng đối với phương pháp ARIMA. Đối với phương pháp SARIMA thì mô hình SARIMA$(0,1,0)(0,1,0)_{12}$ là mô hình cho kết quả tốt nhất so với các mô hình còn lại. Kết quả đánh giá thực nghiệm nghiên cứu dữ liệu giá vàng cho thấy phương pháp SARIMA tối ưu hơn phương pháp ARIMA. Kết quả dự báo này có thể ứng dụng vào thực tế. Các phương pháp này có thể được nghiên cứu và mở rộng để dự báo các loại hàng hóa khác trên thị trường.

❖ ***Tuyên bố về quyền lợi:*** *Các tác giả xác nhận hoàn toàn không có xung đột về quyền lợi.*

TÀI LIỆU THAM KHẢO

Chandar, S. K., & Sumathi, M., & Sivanadam, S. N. (2016). Forecasting Gold Prices Based on Extreme Learning Machine. Int *J Comput Commun Control, 11*(3), 372-380.

Changshou, L., & Liying Z., & Qingfeng W. (2003). Application of SARIMA Model in Cucumber Price Forecast. *Applied Mechanics and Materials, 373-375,* 1686-1690.

Chen, H. H., & Chen, M. Ch., & Chiu, C. Ch. (2014). The Integration of Artificial Neural Networks and Text Mining to Forecast Gold Futures Prices. *Commun Stat Simul Comput*. Retrieved June 15, 2021, from https://doi.org/10.1080/03610918.2013.786780

Futian, W., & Yinhao, C., & Zheng, W., & Muzhou H., & Jianshu L., & Zhongchu, T.(2020). Gold price forecasting research based on an improved online extreme learning machine algorithm. *Journal of Ambient Intelligence and Humanized Computing*, 11, 4101-4111.

Hussein, S. F. M., & Shah, M. B. N., & Jalal, M. R., & Abdullah, S. (2011). *Gold Price Prediction Using Radial Basis Function Neural Network*. Institute of Electrical and Electronics Engineers (IEEE). Fourth International Conference on Modeling, Simulation and Applied Optimization. Retrieved June 15, 2021, from https://10.1109/ICMSAO.2011.5775457

Vu, T. N. (2004). *Ung dung cac thuat toan may hoc trong phân tích dự báo* [*Application of machine learning algorithms in predictive analytics*]. *VNUHCM Journal of Science and Technology Development*, (1).

Peter, G. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing, 50*, 159-175.

Sami, I., & Junejo, K. N. (2017). *Predicting future gold rates using machine learning approach*. Int *J Adv Comput Sci Appl, 8*(12).

Sami, I., & Junejo, K. N. (2017). *Predicting Future Gold Rates Using Machine Learning Approach*. Int *J Adv Comput Sci Appl., 8*(12). Retrieved June 15, 2021, from https://doi.org/10.14569/IJACSA.2017.081213

Williams, S. (2018). *7 common factors that influence gold prices*. Retrieved from The Motley fool. Retrieved May 20, 2021, from https://www.fool.com/investing/2016/10/13/7-common-factors-that-influence-gold-prices.aspx

---

USING SOME MACHINE LEARNING METHODS FOR TIME SERIES FORECASTING

Nguyen Thi Hong Thao[1*], Dao Minh Chau[1], Vu Thanh Nguyen[1], Phu Phuoc Huy[2]

[1]Ho Chi Minh City University of Food Industry, Vietnam

[2]Information Technology Institution, Academy of Military Science and Technology, Vietnam

[*]Corresponding author: Nguyen Thi Hong Thao – Email: thaonth@hufi.edu.vn

Received: September 13, 2021; Revised: January 14, 2022; Accepted: March 13, 2022

ABSTRACT

Gold is considered critical in the economy, especially in assessing the inflation index of the economy. Several machine learning methods were applied in time series data using ARIMA and SARIMA models to forecast future gold prices. We conducted a series of experiments and evaluated the proposed models as well as analyzing the factor variables of time series to find the best results to increase predictive performance.

Keywords: gold price prediction; machine learning; ARIMA and SARIMA models