

# Phân tích, dự đoán giá chứng khoán bằng mô hình thống kê

Trần Thị Mỹ Xoan - 21522815

Lê Anh Tuấn Dũng - 21521974

Lê Thị Ánh Hồng - 21520245

Nguyễn Thị Mai Liên - 21522283

Đỗ Sĩ Đạt - 21521932

Ngày 10 tháng 6 năm 2024

**Tóm tắt nội dung:** Khi thực hiện đầu tư vào các cổ phiếu chứng khoán, việc sử dụng các mô hình dự đoán giá cổ phiếu trước khi thực hiện một giao dịch đầu tư có thể giúp cho ta có các quyết định đúng đắn hơn khi ra quyết định thực hiện 1 giao dịch. Bài báo này trình bày phương pháp dự đoán giá cổ phiếu của ba công ty lớn Samsung, LG và Sony. Chúng tôi sẽ thực hiện dự đoán bằng cách sử dụng các thuật toán thống kê: ARIMA, Linear Regression, VARMA. Bộ dữ liệu được sử dụng lấy từ ngày 1/3/2019 đến ngày 1/6/2024. Sau khi thực hiện dự đoán thì ta sẽ thực hiện sử dụng các độ đo MAPE, MSE, RMSE để đánh giá các mô hình. Cuối cùng, sử dụng các mô hình để thực hiện dự đoán giá chứng khoán cho 30, 60 và 90 ngày tiếp theo.

**Từ khoá:** ARIMA, Linear Regression, VARMA. Phân tích, dự đoán giá chứng khoán.

## 1 GIỚI THIỆU

Dự đoán giá cổ phiếu là một vấn đề quan trọng trong lĩnh vực đầu tư tài chính. Việc dự đoán chính xác giá cổ phiếu có thể giúp các nhà đầu tư đưa ra quyết định đầu tư sáng suốt và giảm thiểu rủi ro. Có nhiều phương pháp khác nhau để dự đoán giá cổ phiếu, bao gồm phân tích kỹ thuật, phân tích cơ bản và học máy. Samsung, LG, SONY là những công ty hàng đầu trong ngành công nghệ, giá cổ phiếu của 3 công ty đóng vai trò là chỉ số quan trọng về động lực thị trường, khiến việc phân tích của họ trở nên cần thiết đối với các nhà đầu tư.

Báo cáo này nhằm mục đích phân tích, dự đoán giá cổ phiếu của các công ty có ảnh hưởng bằng cách sử dụng 3 thuật toán ARIMA, Linear Regression, VARMA để đưa ra dự đoán cho giá cổ phiếu. Nhờ đó các nhà đầu tư có thể lựa chọn chính xác hạng mục đầu tư vào giá cổ phiếu.

Bằng cách tận dụng những phương pháp này, chúng tôi mong muốn cung cấp hướng dẫn có giá trị cho các nhà đầu tư trong việc định hướng bối cảnh năng động của lĩnh vực công nghệ.

## 2 NGHIÊN CỨU LIÊN QUAN

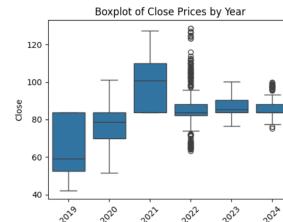
[1] Mô hình VARMA thường được sử dụng để dự báo dữ liệu chuỗi thời gian đa biến. Mô hình VARMA là sự mở rộng của mô hình ARMA trong chuỗi thời gian đơn biến (Lütkepohl, 2005; Wei, 1990) và được sử dụng với điều kiện rằng dữ liệu phải ổn định theo thời gian (Lütkepohl, 2005). Mô hình VARMA ( $p,q$ ) là sự kết hợp của mô hình VAR ( $p$ ) và mô hình vector trung bình động ( $q$ ) (VMA ( $q$ )). Bài báo này xác định 4 mô hình VARMA ( $p,q$ ) lần lượt là (1,1), (2, 1), (3, 1), (4, 1), Việc lựa chọn mô hình tốt nhất được thực hiện bằng cách sử dụng một số tiêu chí thông tin (AICC, HQC, AIC, và SBC). Các giá trị nhỏ nhất

của những tiêu chí này cho biết mô hình tốt nhất. Mô hình phù hợp nhất được chọn là VARMA(2, 1). Kết quả dự báo cho thấy sai số chuẩn tăng lên theo thời gian; sai số chuẩn trong tháng đầu tiên tương đối nhỏ so với dự đoán trung bình, nhưng tăng lên theo thời gian đến dự báo cho 12 tháng tiếp theo. Điều này cho thấy mô hình là hợp lý khi dự báo cho các khoảng thời gian ngắn, nhưng kết quả không ổn định (do sai số chuẩn cao hơn) khi dự báo cho các khoảng thời gian dài.

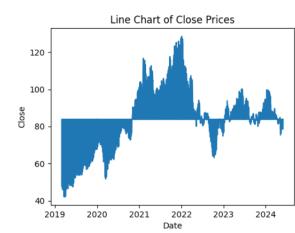
[2] Mô hình thuật toán Linear Regression được sử dụng để dự đoán giá trị của một biến phụ thuộc (biến đầu ra) dựa trên giá trị của một hoặc nhiều biến độc lập (biến đầu vào), được sử dụng để xác định và phân tích các xu hướng trong dữ liệu thời gian. Trong bài báo này, hiệu suất của hồi quy tuyến tính (LR) và hồi quy vectơ hỗ trợ (SVR) để dự đoán giá cổ phiếu của Amazon vào tháng 10 năm 2019. Phát hiện của họ chỉ ra rằng LR đạt được độ chính xác cao hơn (98,76%) so với SVR (94,32%), cho thấy khả năng phù hợp của nó trong ngắn hạn - giá kỳ hạn trước khi xác định cổ phiếu có ít biến động hơn.

Bảng 1: SONY, SAMSUNG, LG's Descriptive Statistics

	SONY	SAMSUNG	LG
Count	1920	1920	1920
Mean	83.842011	83.842011	101952.316347
Std	15.828603	15.828603	26596.447698
Min	42.030000	42.030000	41850
25%	79.987500	79.987500	87200
50%	83.842011	83.842011	101952.316347
75%	90.442500	90.442500	114000
Max	128.59	128.590000	185000



Hình 1: SONY stock price's boxplot



Hình 2: SONY stock price's histogram

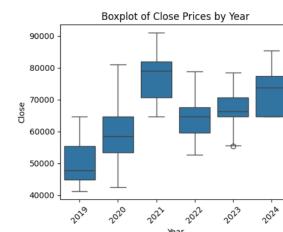
### 3 PHƯƠNG PHÁP

#### 3.1 DATASET

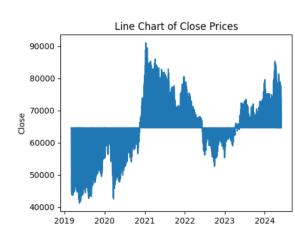
Trong tâm của bài viết này là dự đoán giá cổ phiếu. Vì vậy, cả ba tập dữ liệu chúng tôi sử dụng trong nghiên cứu đều là giá cổ phiếu của các công ty lớn. Các công ty này là SONY, SAMSUNG và LG. Dữ liệu của SAMSUNG, LG, SONY được thu thập từ ngày 1 tháng 3 năm 2019 đến ngày 1 tháng 6 năm 2024.

Mô tả từng cột trong dữ liệu :

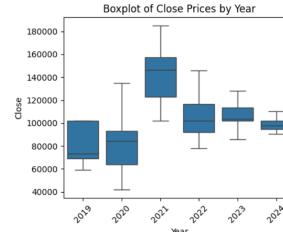
- + Date : Ngày giao dịch.
- + Open: Giá mở cửa của cổ phiếu vào ngày giao dịch.
- + High: Giá cao nhất của cổ phiếu trong ngày giao dịch.
- + Low: Giá thấp nhất của cổ phiếu trong ngày giao dịch.
- + Close: Giá đóng cửa của cổ phiếu vào ngày giao dịch.



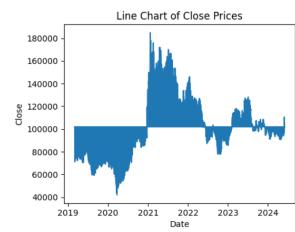
Hình 3: SAMSUNG stock price's boxplot



Hình 4: SAMSUNG stock price's histogram



Hình 5: LG stock price's boxplot



Hình 6: LG stock price's histogram

## 3.2 THUẬT TOÁN VARMA

Mô hình VARMA thường được sử dụng để dự báo dữ liệu chuỗi thời gian đa biến. Mô hình VARMA là một mở rộng của mô hình ARMA trong chuỗi thời gian đơn biến (Lutkepohl, 2005; Wei, 1990) và được sử dụng với điều kiện dữ liệu phải là dừng theo thời gian (Lutkepohl, 2005). Mô hình VARMA ( $p, q$ ) là sự kết hợp của mô hình VAR ( $p$ ) và mô hình trung bình trượt vector ( $q$ ) (VMA ( $q$ )).

- Công thức:

$$\mathbf{y}_t = \mathbf{c} + \sum_{i=1}^p \Phi_i \mathbf{y}_{t-i} + \sum_{j=1}^q \Theta_j \epsilon_{t-j} + \epsilon_t \quad (1)$$

Trong đó:

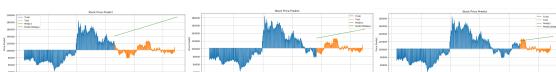
- $\mathbf{y}_t$  là vector của các biến tại thời điểm  $t$ .
- $\mathbf{c}$  là một vector hằng số.
- $\Phi_i$  là ma trận hệ số của chuỗi thời gian (độ trễ, lag)  $i$  của mô hình AR.
- $\Theta_j$  là ma trận hệ số của chuỗi thời gian (độ trễ, lag)  $j$  của mô hình MA.
- $\epsilon_t$  là vector của các nhiễu ngẫu nhiên tại thời điểm  $t$ .

## 4 THỰC NGHIỆM

### 4.1 MÔ HÌNH DỰ ĐOÁN

#### 4.1.1 ARIMA

#### 4.1.2 LINEAR REGRESSION



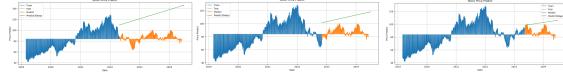
Hình 7: LG 30 DAYS 6:4, 7:3, 8:2



Hình 8: LG 60 DAYS 6:4, 7:3, 8:2



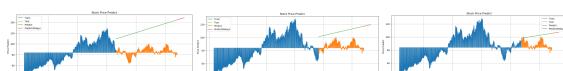
Hình 9: LG 90 DAYS 6:4, 7:3, 8:2



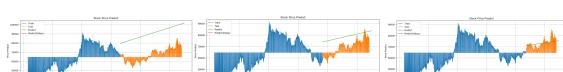
Hình 10: SONY 30 DAYS 6:4, 7:3, 8:2



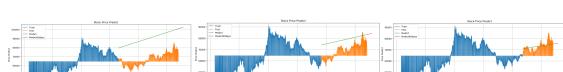
Hình 11: SONY 60 DAYS 6:4, 7:3, 8:2



Hình 12: SONY 90 DAYS 6:4, 7:3, 8:2



Hình 13: SAMSUNG 30 DAYS 6:4, 7:3, 8:2

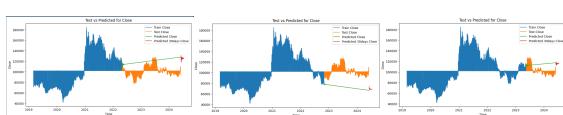


Hình 14: SAMSUNG 60 DAYS 6:4, 7:3, 8:2

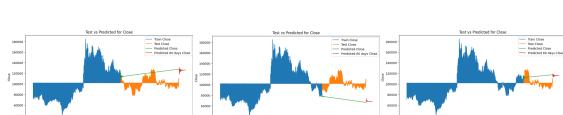


Hình 15: SAMSUNG 90 DAYS 6:4, 7:3, 8:2

#### 4.1.3 VARMA



Hình 16: LG 30 DAYS 6:4, 7:3, 8:2



Hình 17: LG 60 DAYS 6:4, 7:3, 8:2



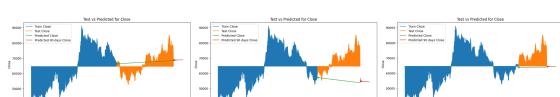
Hình 18: LG 90 DAYS 6:4, 7:3, 8:2



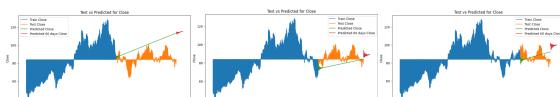
Hình 23: SAMSUNG 60 DAYS 6:4, 7:3, 8:2



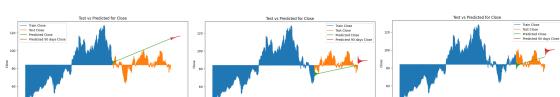
Hình 19: SONY 30 DAYS 6:4, 7:3, 8:2



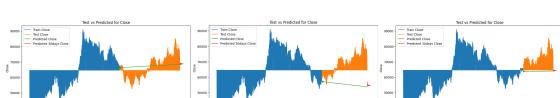
Hình 24: SAMSUNG 90 DAYS 6:4, 7:3, 8:2



Hình 20: SONY 60 DAYS 6:4, 7:3, 8:2



Hình 21: SONY 90 DAYS 6:4, 7:3, 8:2



Hình 22: SAMSUNG 30 DAYS 6:4, 7:3, 8:2

## 5 TÀI LIỆU THAM KHẢO

[1]: Warsono, Edwin Russel, Wamiliana\*, Widiarti, Mustofa Usman, "Modeling and Forecasting by the Vector Autoregressive Moving Average Model for Export of Coal and Oil Data (Case Study from Indonesia over the Years 2002-2017)". International Journal of Energy Economics and Policy, 2019

[2] D. Bhuriya, G. Kaushal, A. Sharma, and U. Singh, "Stock market predication using a linear regression," in 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA),