RESEARCH ARTICLE

SO SÁNH CÁC MÔ HÌNH DỰ BÁO TRONG DỰ ĐOÁN GIÁ CHỨNG KHOÁN

Chu Đặng Bình An, Hoàng Đình Thăng, Trần Minh Thái

Khoa Công nghệ thông tin, Trường Đại học Ngoại ngữ -Tin học TP.HCM

21dh113175@st.huflit.edu.vn, 21dh114517@st.huflit.edu.vn, thaitm@huflit.edu.vn

TÓM TẮT— Những bài toán dự đoán bằng các mô hình là những bài toán đóng vai trò nền tảng quan trọng và được ứng dụng rộng rãi trong nhiều lĩnh vực liên quan đến đời sống con người như thời tiết, y tế hay giá cả thị trường. Những bài toán này tập trung vào việc dự đoán những kết quả của sự việc, sự kiện hay các giá trị trong tương lai dựa trên những giá trị dữ liệu lịch sử thông qua xây dựng các mô hình dự đoán. Nội dung nghiên cứu của bài báo tập trung vào việc xây dựng các mô hình dự đoán trên dữ liệu chuỗi thời gian của tập dữ liệu chứng khoán được trích dẫn từ sàn VNINDEX. Thông qua những kỹ thuật phân tích, tiền xử lý dữ liệu, lựa chọn thông số phù hợp cho từng đặc điểm của mô hình và thực hiện xây dựng, huấn luyện các mô hình để đưa ra dự đoán xu hướng giá chứng khoán. Một số phương pháp tiêu biểu được sử dụng bao gồm Autoregressive Intergrated Moving Average, Vector Autoregression, Holt-Winters và Facebook Prophet. Kết quả thực nghiệm cho thấy phương pháp Facebook Prophet là phương pháp dự đoán trên chuỗi thời gian có hiệu suất và độ chính xác cao hơn so với những phương pháp còn lại.

Từ khóa— Dự đoán chuỗi thời gian, chứng khoán, ARIMA, VAR, Holt-Winters, Facebook Prophet.

I. GIỚI THIỆU

Hiện nay, sự bùng nổ thông tin trong nhiều lĩnh vực như thị trường chứng khoán đã tạo ra lượng thông tin giao dịch mỗi giây được lưu lại là rất lớn. Thị trường chứng khoán là nơi các nhà đầu tư giao dịch chứng khoán làm tăng hay giảm khoản đầu tư ban đầu của mình. Nhiều phương pháp và kỹ thuật đã nghiên cứu dự đoán xu hướng cổ phiếu nhằm hạn chế rủi ro cho các nhà đầu tư. Thông thường, các nhà đầu tư sử dụng phân tích cơ bản và phân tích kỹ thuật để phân tích dự đoán nhằm lập chiến lược giao dịch cổ phiếu cho riêng mình. Với sự phát triển của công nghệ, nhiều mô hình thống kê, học máy, học sâu được nghiên cứu và cho ra đời để dự đoán dựa trên dữ liệu chuỗi thời gian nói chung và phân tích, dự đoán biến động giá chứng khoán nói riêng.

Trong bài báo này, chúng tôi đề xuất áp dụng các phương pháp dự báo chuỗi thời gian truyền thống như ARIMA, VAR. Tuy nhiên các mô hình dự báo truyền thống thường gặp khó khăn trong việc xử lý dữ liệu và các yếu tố như chu kỳ, mùa vụ của chuỗi thời gian không rõ ràng. Cho nên, ngoài các mô hình truyền thống chúng tôi đề xuất thêm các mô hình có thêm yếu tố mùa vụ như Holt-Winters, Facebook Prophet để cải thiện hiệu suất trong việc dự báo chuỗi thời gian nói chung và dự báo biến động trên tập dữ liệu giá chứng khoán nói riêng.

Phần còn lại của bài báo được tổ chức với bố cục bao gồm: Mục I giới thiệu bài toán. Tiếp theo Mục II mô tả bài toán và các khái niệm liên quan. Mục III trình bày các công trình nghiên cứu liên quan. Mục IV và mục V trình bày thuật toán đề xuất và các kết quả thực nghiệm. Cuối cùng Mục VI là phần kết luận và đề xuất các hướng nghiên cứu tiếp theo.

II. MÔ TẢ BÀI TOÁN

Bài toán dự đoán xu hướng giá chứng khoán là một bài toán dự báo chuỗi thời gian, trong đó mục tiêu là dự đoán xu hướng tăng, giảm hoặc đi ngang của giá chứng khoán trong tương lai dựa trên dữ liệu lịch sử và các yếu tố liên quan khác. Giá chứng khoán thường thể hiện các đặc trưng của chuỗi thời gian như xu hướng, tính mùa vụ và tính chu kỳ, đồng thời cũng chịu ảnh hưởng của nhiều yếu tố kinh tế, chính trị và xã hội. Do đó, việc dự báo giá chứng khoán là một thách thức lớn, đòi hỏi các mô hình dự báo phải có khả năng xử lý tốt các đặc trưng này và thích ứng với sự biến động của thị trường.

Trong nghiên cứu này, chúng tôi sử dụng bốn mô hình chuỗi thời gian phổ biến để dự báo xu hướng giá chứng khoán bao gồm: Autoregressive Intergrated Moving Average (ARIMA) [1, 2, 3], Vector Autoregression (VAR) [4], Holt-Winters [5] và Facebook Prophet [6, 7].

A. MÔ HÌNH ARIMA

Mô hình tự hồi quy tích hợp trung bình trượt, là một mô hình dự báo chuỗi thời gian cổ điển và phổ biến. Mô hình này sử dụng các giá trị quá khứ của chuỗi để dự đoán giá trị tương lai. ARIMA có khả năng mô hình hóa các chuỗi thời gian có xu hướng và tính mùa vụ, nhưng có thể gặp khó khăn khi xử lý dữ liệu có tính chất phi tuyến hoặc yếu tố mùa vụ phức tạp. ARIMA kết hợp ba thành phần: (i) Thành phần tự hồi quy (AR): mô tả mối quan hệ giữa các giá trị hiện tại và các giá trị quá khứ của chuỗi; (ii) Thành phần trung bình động (MA): mô tả mối quan hệ giữa các giá trị hiện tại và các sai số dự báo quá khứ và **Thành phần lấy sai phân** (I): làm cho chuỗi trở nên

dừng (stationary) bằng cách lấy sai phân của chuỗi gốc. Phương trình tổng quát của ARIMA(p,d,q) được thể hiện trong công thức (1).

$$\Delta Y_t = \Phi_1 \Delta_{t-1} + \Phi_2 \Delta Y_{t-2} + \cdots + \Phi_p \Delta Y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \quad (1)$$

Trong đó: ΔY_t là giá trị sai phân bậc d của chuỗi tại thời điểm t. Φ_1, Φ_2, Φ_p là các hệ số của phần hồi quy tự động AR với các độ trễ tương ứng; $\theta_1, \theta_2, \theta_p$ là các hệ số của phần trung bình trượt MA với các sai số độ trễ tương ứng; ε_t là nhiễu trắng tại thời điểm t, với kỳ vọng bằng 0 và phương sai không đổi.

B. MÔ HÌNH VAR

Mô hình vector tự hồi quy, là một mô hình dự báo chuỗi thời gian đa biến, cho phép mô tả mối quan hệ giữa nhiều biến chuỗi thời gian khác nhau. VAR biểu diễn mỗi biến là một hàm tuyến tính của các giá trị quá khứ của chính nó và các biến khác trong mô hình. VAR có thể nắm bắt được sự tương tác giữa các biến, nhưng đòi hỏi dữ liệu phải dừng và có thể gặp khó khăn khi xử lý dữ liệu có số chiều cao. Phương trình tổng quát của VAR(p) theo công thức (2).

$$Y_t = c + A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots + A_p Y_{t-p} + \varepsilon_t \quad (2)$$

Trong đó: Y_t là vector cột các biến nội sinh tại thời điểm *t*; *c* là vector cột các hằng số; A_1, A_2, A_p là các ma trận hệ số ảnh hưởng tương ứng với các biến tại thời điểm *t-1, t-2*; ε_t là vector cột của các sai số ngẫu nhiên hay là nhiễu trắng.

C. MÔ HÌNH HOLT-WINTERS

Phương pháp làm trơn số mũ, là một mô hình dự báo chuỗi thời gian có khả năng xử lý cả xu hướng và tính mùa vụ. Holt-Winters sử dụng ba thành phần: (i) **Mức độ (α)**: ước tính giá trị trung bình của chuỗi tại mỗi thời điểm. (ii) **Xu hướng (β)**: ước tính sự thay đổi của mức độ theo thời gian và (iii) **Yếu tố mùa vụ (γ)**: ước tính sự lặp lại theo chu kỳ của chuỗi. Phương pháp Holt-Winters được phân loại thành 2 mô hình nhỏ Multiplicative Holt-Winters (MHW) theo công thức (3) và Additive Holt-Winters (AHW) theo công thức (4).

AHW: Sử dụng khi sự biến đổi mùa vụ không phụ thuộc vào mức độ của chuỗi thời gian.

$$\begin{aligned} \hat{y}_t + h|t &= \ell_t + hb_t + s_t + h - m(k+1) \\ \ell_t &= \alpha(y_t - s_t - m) + (1-\alpha)(\ell_t - 1 + b_t - 1) \\ b_t &= \beta * (\ell_t - \ell_t - 1) + (1 - \beta *) b_t - 1 \\ s_t &= \gamma(y_t - \ell_t - 1 - b_t - 1) + (1-\gamma) s_t - m \end{aligned} \quad (3)$$

MHW: Sử dụng khi sự biến đổi mùa vụ phụ thuộc vào mức độ của chuỗi thời gian.

$$\begin{aligned} \hat{y}_t + h|t &= (\ell_t + hb_t) s_t + h - m(k+1) \\ \ell_t &= \alpha * y_t/(s_t\text{-}m) + (1-\alpha)(\ell_t - 1 + b_t - 1) \\ b_t &= \beta * (\ell_t - \ell_t - 1) + (1 - \beta *) b_t - 1 \\ s_t &= \gamma * y_t/(\ell_t - 1 - b_t - 1) + (1-\gamma) s_t - m \end{aligned} \quad (4)$$

Trong đó: $\hat{y}t$ là giá trị dự đoán vào thời điểm t; $\ell_t$là mức độ cập nhật tại thời điểm t; $b_t$là xu hướng cập nhật tại thời điểm t (khoảng thời gian diễn ra xu hướng); s_t là yếu tố mùa vụ cập nhật tại thời điểm t (khoảng thời gian diễn ra mùa vụ); *m* là chu kỳ mùa vụ, thường m =12 tháng (lượng dữ liệu được xác định); *k* là phần nguyên của (h−1)/m để đảm bảo rằng các ước tính chỉ số mùa vụ được sử dụng cho dự báo đến từ năm cuối cùng của dữ liệu.

D. MÔ HÌNH FACEBOOK PROPHET

Mô hình Prophet được phát triển bởi Facebook và đã trở nên phổ biến trong dự báo chuỗi thời gian nhờ vào tính linh hoạt, dễ triển khai và khả năng xử lý tốt các yếu tố xu hướng, mùa vụ và các sự kiện đặc biệt. Prophet được thiết kế đặc biệt để dễ sử dụng cho những người không có kiến thức sâu về thống kê nhưng vẫn có thể tạo ra các dự báo đáng tin cậy.

Prophet chia chuỗi thời gian thành ba thành phần chính: xu hướng, tính mùa vụ, và các ngày lễ hoặc sự kiện đặc biệt, cùng với nhiễu ngẫu nhiên. Prophet cung cấp hai lựa chọn: mô hình cộng (Additive Seasonality) theo công thức (5) và mô hình nhân (Multiplicative Seasonality) theo công thức (6), tùy thuộc vào tính chất biến động của chuỗi thời gian.

Trong mô hình cộng, các thành phần được cộng lại với nhau, thích hợp khi dữ liệu có biên độ dao động ổn định, khi xu hướng tăng nhưng biên độ của các dao động mùa vụ không thay đổi nhiều.

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \qquad (5)$$

Với mô hình nhân, các thành phần này sẽ nhân với nhau, phù hợp khi dữ liệu có biên độ dao động tăng theo xu hướng, tạo ra sự thay đổi phi tuyến tính.

$$y(t) = g(t) * s(t) * h(t) * \epsilon_t \qquad (6)$$

Trong đó, *g(t)* là hàm xu hướng, đại diện cho sự thay đổi không có tính chu kỳ của chuỗi thời gian. Đây có thể là một xu hướng tuyến tính hoặc phi tuyến tính; *s(t)* là hàm mùa vụ, đại diện cho sự thay đổi có tính chu kỳ của chuỗi thời gian, ví dụ như sự thay đổi theo tuần, tháng, hoặc năm; *h(t)* là hàm đại diện cho các ảnh hưởng của ngày lễ, những ngày đặc biệt do người dùng cung cấp; $\epsilon_t$là sai số ngẫu nhiên, đại diện cho những dao động không thể dự đoán trước được trong chuỗi thời gian.

Prophet phù hợp cho dự báo các chuỗi thời gian có tính biến động mạnh và xử lý linh hoạt các xu hướng dài hạn và các yếu tố mùa vụ, đồng thời cho phép tùy chỉnh dễ dàng các yếu tố đặc biệt, một đặc điểm quan trọng khi dự báo giá chứng khoán vốn phụ thuộc vào các sự kiện bất ngờ. Prophet cũng hỗ trợ tốt trong các tình huống có dữ liệu thiếu hoặc nhiễu ngẫu nhiên.

Trong thực tế, khi làm việc với bộ dữ liệu sẽ gặp những giai đoạn có xu hướng ổn định trong khoảng thời gian này, biến động lớn trong khoảng thời gian kia thì việc chọn mô hình thích hợp có thể gặp khó khăn, việc xác định một mô hình có thể không đủ. Vì vậy, áp dụng cả mô hình cộng và mô hình nhân để so sánh và đánh giá hiệu quả việc dự báo trung hạn trong tương lai.

III. CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN

Có nhiều công trình nghiên cứu các phương pháp, ứng dụng phân tích và xử lý dữ liệu với các mô hình học máy trong các bài toán làm việc với chuỗi thời gian, đặc biệt là trong lĩnh vực kinh tế, cụ thể ở đây là dự báo xu hướng giá chứng khoán. Những nghiên cứu này đề cập đến đến các phương pháp, các cách tiền xử lý dữ liệu và các mô hình học máy, học sâu ứng dụng vào việc dự báo xu hướng giá chứng khoán.

Nghiên cứu của Mehar Vijh và các cộng sự [8] tập trung vào việc dự đoán giá đóng cửa của cổ phiếu bằng cách sử dụng các kỹ thuật học máy, cụ thể là Mạng nơ-ron nhân tạo (ANN) và Random Forest (RF). Các tác giả đã sử dụng dữ liệu lịch sử về giá cổ phiếu (giá mở, cao, thấp, đóng) để tạo ra các biến mới, sau đó sử dụng các biến này làm đầu vào cho mô hình. Kết quả thực nghiệm cho thấy rằng mô hình ANN đưa ra kết quả dự đoán tốt hơn RF dựa trên chỉ số đánh giá MAPE và RMSE.

Một công trình nghiên cứu khác của V Kranthi Sai Reddy [9] đề xuất sử dụng phương pháp Support Vector Machine (SVM) với Radial Basis Function (RBF) kernel, để dự đoán thị trường chứng khoán. Các tính năng được sử dụng bao gồm biến động giá cổ phiếu, động lượng giá, biến động chỉ số và động lượng chỉ số. Nghiên cứu sử dụng môi trường Weka và YALE Data Mining để thực hiện các thí nghiệm. Kết quả cho thấy SVM có thể hoạt động trên tập dữ liệu lớn được thu thập từ các thị trường tài chính toàn cầu khác nhau và không gặp vấn đề về quá khớp.

Shunrong Shen, Haomiao Jiang và Tongda Zhang [10] đề xuất một thuật toán dự đoán mới khai thác mối tương quan thời gian giữa các thị trường chứng khoán toàn cầu và các sản phẩm tài chính khác nhau để dự đoán xu hướng giá chứng khoán với phương pháp SVM (Support Vector Machine). Kết quả thực nghiệm cho thấy thuật toán đạt được độ chính xác dự đoán là 74,4% trên NASDAQ, 76% trên S&P 500 và 77,6% trên DJIA.

Himanshu Gupta và Aditya Jaiswal [11] đã tiến hành nghiên cứu về dự báo giá chứng khoán sử dụng nhiều mô hình học sâu khác nhau bao gồm LSTM, RNN và CNN. Trong nghiên cứu này, dữ liệu từ chỉ số S&P 500 được sử dụng để thử nghiệm các mô hình học sâu này. Kết quả cho thấy mô hình LSTM có độ chính xác cao nhất khi so sánh với các mô hình khác nhờ khả năng ghi nhớ các phụ thuộc dài hạn trong chuỗi thời gian.

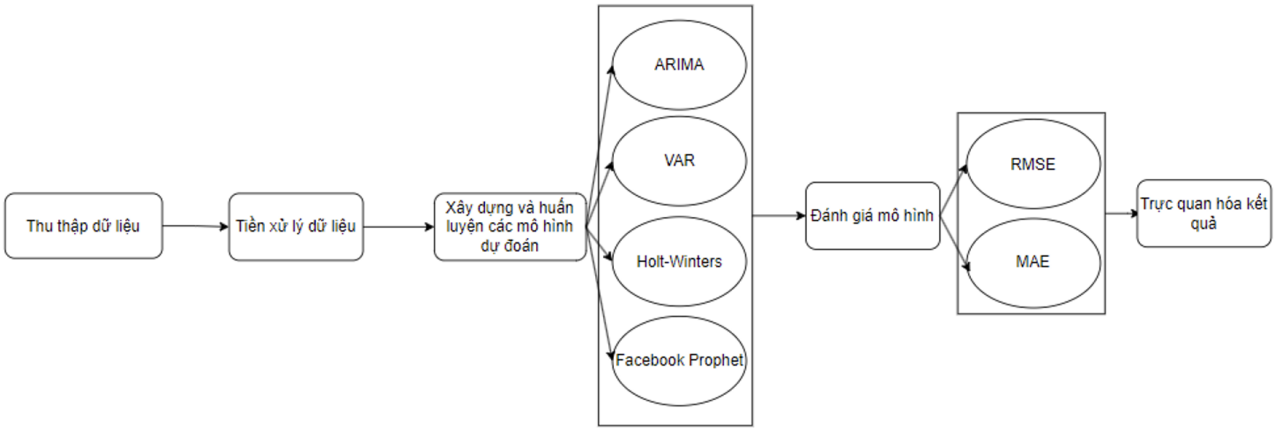
Một nghiên cứu khác của Allison Koenecke [12] tập trung vào việc áp dụng các mạng nơ-ron học sâu vào dự báo chuỗi thời gian tài chính của Microsoft. Tác giả sử dụng LSTM kết hợp với học tăng dần (curriculum learning) để cải thiện độ chính xác trong dự báo. Kết quả thực nghiệm đã cho thấy phương pháp học sâu này mang lại hiệu quả dự báo vượt trội hơn so với các kỹ thuật truyền thống.

Cuối cùng, Alexiei Dingli và Karl Sant Fournier [13] đã thử nghiệm sử dụng mạng nơ-ron tích chập (CNN) để dự báo hướng biến động giá chứng khoán trong ngắn hạn. Mặc dù kết quả của mô hình CNN chưa vượt qua các mô hình truyền thống như Logistic Regression và Support Vector Machines, nhưng các tác giả khẳng định rằng với các kỹ thuật tinh chỉnh thêm, CNN có tiềm năng vượt trội trong các bài toán dự báo chuỗi thời gian tài chính.

Nhìn chung, các công trình nghiên cứu đều đề cập đến các bước tiền xử lý dữ liệu chuỗi thời gian, đồng thời đề xuất các mô hình dự đoán với những cải tiến dựa trên nhiều nguồn dữ liệu chứng khoán khác nhau. Những kết quả từ các bài nghiên cứu không chỉ cải thiện độ chính xác của các mô hình mà còn mở ra nhiều hướng nghiên cứu khác nhau trong việc xử lý dữ liệu, tối ưu hóa mô hình trong lĩnh vực này.

IV. MÔ HÌNH ĐỀ XUẤT

Hình 1 thể hiện sơ đồ xử lý của quá trình dự đoán xu hướng giá chứng khoán. Quá trình xử lý gồm 6 giai đoạn: (i) Thu thập dữ liệu, (ii phân tích và tiền xử lý dữ liệu, (iii) xây dựng và huấn luyện các mô hình, (iv) đánh giá mô hình dự đoán và (v) trực quan hóa kết quả.



Hình 1. *Sơ đồ tổng quan các bước thực hiện.*

A. THU THẬP DỮ LIỆU

Tập dữ liệu chứng khoán là tập dữ liệu được thu thập bằng thư viện VNSTOCK [14] trên sàn chứng khoán HOSE với 30 mã chứng khoán khác nhau đến từ nhiều công ty thuộc mọi lĩnh vực từ kinh tế, thương mại, công nghệ thông tin, dầu khí, v.v... Tập dữ liệu này gồm 39.280 dòng dữ liệu với 7 thuộc tính được thu thập từ ngày 02/01/2019 đến ngày 26/04/2024. Dữ liệu có 7 thuộc tính được mô tả trong Bảng 1.

Bảng 1. *Mô tả về tập dữ liệu*

STT	Đặc trưng	Kiểu dữ liệu	Ý nghĩa	Ví dụ
1	time	datetime	Ngày diễn ra phiên giao dịch	02/01/2019
2	open	Numeric	Giá mở đầu cho phiên giao dịch hay còn gọi là giá mở cửa	94.900
3	high	Numeric	Giá cao nhất đạt được trong phiên giao dịch cùng ngày	96.200
4	low	Numeric	Giá thấp nhất đạt được trong phiên giao dịch cùng ngày	94.800
5	close	Numeric	Giá kết thúc phiên giao dịch cùng ngày hay còn gọi là giá đóng cửa	96.000
6	volume	Numeric	Khối lượng giao dịch trong phiên giao dịch cùng ngày	2.017.000
7	ticker	Nominal	Mã chứng khoán của các công ty trên sàn giao dịch	FPT

Dữ liệu chứng khoán được chia thành 2 tập dữ liệu train và test. Với dữ liệu tập train là từ ngày 02/01/2019 đến ngày 29/12/2023 với 36.940 dòng dữ liệu. Với tập test là từ ngày 02/01/2024 đến ngày 26/04/2024 với 2.340 dòng dữ liệu.

B. TIỀN XỬ LÝ DỮ LIỆU

Thực hiện chuyển đổi dạng dữ liệu cho tập dữ liệu thành dạng datetime với đặc trưng time được thiết lập thành index cho tập dữ liệu.

C. MÔ HÌNH ARIMA

Tiến hành kiểm tra tính dừng cho chuỗi dữ liệu với tập kiểm định thống kê ADF, nếu chuỗi không dừng thì thực hiện tính sai phân để chuỗi dừng sau đó thực hiện vẽ biểu đồ ACF và PACF để xác định các độ trễ bậc p, q cho mô hình.

```
ADF Test Statistic: -1.075136
p-value: 0.724885
Critical Values:
{'1%': -3.4356133204636095, '5%': -2.8638642784217305, '10%': -2.5680074748794373}
Chuỗi không dừng
```

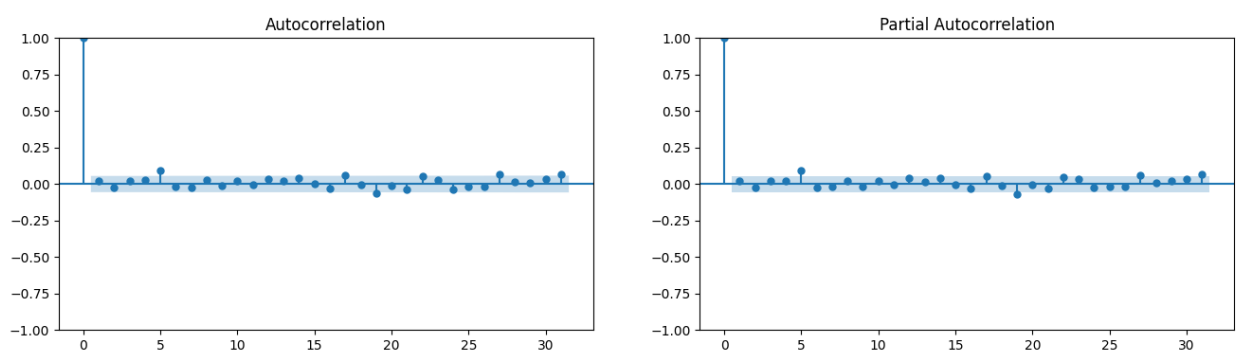
Hình 2. *Kết quả kiểm tra tính dừng trên một mã SSI.*

Kết quả trong Hình 2 cho thấy rằng chuỗi dữ liệu mã SSI là chuỗi không dừng, chúng ta phải tiến hành tính sai phân để biến đổi thành chuỗi dừng.

```
ADF Test Statistic: -14.024543
p-value: 0.000000
Critical Values:
{'1%': -3.4356133204636095, '5%': -2.8638642784217305, '10%': -2.5680074748794373}
Chuỗi dừng
```

Hình 3. *Kết quả kiểm tra tính dừng của mã SSI sau khi thực hiện sai phân*

Khi chuỗi dừng rồi chúng ta tiến hành vẽ biểu đồ ACF và PACF (Hình 4) để xác định các tham số cho mô hình.



Hình 4. *Biểu đồ ACF và PACF của mã SSI.*

Sau khi biến đổi cho chuỗi dừng, thực hiện vẽ hai biểu đồ ACF và PACF để xác định bậc p, q cho mô hình. Trong đó, để xác định bậc p, q chúng ta dựa vào đỉnh hay bậc có giá trị độ trễ cao nhất để xác định mô hình. Ta thấy được, ở 2 biểu đồ trong Hình 4, thì ngay ở điểm 0 của cả hai biểu đồ đều có độ trễ vượt qua mức 5%, tức là giá trị ở đỉnh 0 của cả hai biểu đồ có sự chênh lệch cao so với các đỉnh còn lại. Cho nên, ta có thể rút ra rằng bậc q và bậc p của mô hình cho mã SSI đều là ARIMA (0,1,0).

D. MÔ HÌNH VAR

Thực hiện gom nhóm dữ liệu giữa 2 cột *time, ticker* rồi xử lý các mục bị trùng lặp bằng giá trị trung bình *mean*. Sau đó sử dụng hàm *pivot* của Pandas để biến đổi dữ liệu với các *ticker* là các cột, thiết lập *date* trở thành index và mỗi hàng tương ứng theo thời gian thì sẽ là giá đóng cửa của các mã chứng khoán. Sau khi thực hiện *pivot* dữ liệu, các mã chứng khoán trong cột ticker trước đó sẽ trở thành các cột dữ liệu chứa giá trị của giá đóng cửa theo từng mã. Tại đây, từng cột dữ liệu của các mã chứng khoán sẽ đại diện cho các biến riêng biệt đưa vào mô hình đa biến VAR (Hình 5).

E. MÔ HÌNH HOLT-WINTERS

Thực hiện chuyển đổi dạng dữ liệu cột 'time' thành 'Datetime', ta tiếp tục đổi tên cột 'time' thành 'ds', cột 'close' thành 'y' và tiến hành chỉ chọn cột 'y' được gắn theo chỉ mục theo thời gian để đơn giản hóa dữ liệu đầu vào cho mô hình Holt-Winters cho cả mô hình cộng và nhân.

F. MÔ HÌNH FACEBOOK PROPHET

Quy trình thực hiện tiền xử lý dữ liệu tương tự với mô hình Holt-Winters.

```
(ticker          ACB      BCM      BID      BVH      CTG      FPT      GAS  \
 time
 2019-01-02   9230.0  22290.0  22360.0  78110.0  12690.0  17800.0  60180.0
 2019-01-03   8890.0  20430.0  21350.0  78290.0  12080.0  17630.0  58720.0
 2019-01-04   8950.0  21080.0  21050.0  77760.0  12250.0  17760.0  58590.0
 2019-01-07   9070.0  20980.0  21550.0  77760.0  12120.0  18060.0  59830.0
 2019-01-08   9010.0  21080.0  21350.0  77410.0  11950.0  18140.0  61220.0
 ...             ...      ...      ...      ...      ...      ...      ...
 2023-12-25  23350.0  61900.0  43200.0  39300.0  26900.0  96000.0  76500.0
 2023-12-26  23250.0  62300.0  43000.0  39500.0  26800.0  97200.0  76400.0
 2023-12-27  23300.0  62600.0  43000.0  39550.0  26850.0  96900.0  76300.0
 2023-12-28  23750.0  62700.0  42700.0  39600.0  27100.0  96600.0  76000.0
 2023-12-29  23900.0  62900.0  43400.0  39500.0  27100.0  96100.0  75500.0

 ticker          GVR      HDB      HPG  ...      TCB      TPB      VCB  \
 time                                   ...
 2019-01-02   8400.0   9040.0  10800.0  ...  25600.0   7850.0  34560.0
 2019-01-03   8310.0   9030.0  10340.0  ...  25000.0   7670.0  34560.0
 2019-01-04   8310.0   9300.0  10410.0  ...  24950.0   7640.0  35080.0
 2019-01-07   7960.0   9260.0  10290.0  ...  25100.0   7720.0  35200.0
 2019-01-08   8130.0   9130.0  10110.0  ...  25100.0   7740.0  35460.0
 ...             ...      ...      ...  ...      ...      ...      ...
 2023-12-25  20200.0  19150.0  27450.0  ...  30950.0  17050.0  81800.0
 2023-12-26  20300.0  19200.0  27800.0  ...  30800.0  17000.0  82800.0
 2023-12-27  20250.0  19300.0  27750.0  ...  30850.0  17350.0  82700.0
 2023-12-28  20300.0  19350.0  27950.0  ...  31500.0  17300.0  82800.0
 2023-12-29  21200.0  20300.0  27950.0  ...  31800.0  17400.0  80300.0
```

Hình 5. *Kết quả biến đổi dữ liệu cho mô hình VAR*

G. XÂY DỰNG MÔ HÌNH

1. MÔ HÌNH ARIMA

Bảng 2. *Mô hình ARIMA*

METHOD: ARIMA_Forecast(data, p, d, q)
INPUT:
data: Chuỗi thời gian đầu vào
p: Bậc của thành phần tự hồi quy (AR)
d: Bậc của quá trình lấy sai phân (I)
q: Bậc của thành phần trung bình động (MA)
OUTPUT:
forecast: Chuỗi các giá trị dự báo
BEGIN:
1. IF NOT is_stationary(data): //kiểm tra tính dừng
2. data = difference(data, d)
3. p, q = identify_pq_parameters(data) // thực hiện xác định thông số p,d
4. model = ARIMA(data, order=(p, d, q))
5. model_fit = model.fit()
6. forecast = model_fit.forecast(steps=horizon)
7. RETURN forecast
END

2. MÔ HÌNH VAR

Bảng 3. *Mô hình VAR*

METHOD: VAR_Forecast(data, p)
INPUT:
data: Tập hợp các chuỗi thời gian đầu vào
p: Số lượng độ trễ (lags)
OUTPUT:

```
   forecast: Ma trận các giá trị dự báo
BEGIN:
1. FOR EACH series IN data:
2.    IF NOT is_stationary(series):
3.       series = difference(series)
4. p = identify_lag_order(data)
5. model = VAR(data)
6. model_fit = model.fit(p)
7. forecast = model_fit.forecast(model_fit.y, steps=horizon)
8. RETURN forecast
END
```

3. MÔ HÌNH HOLT-WINTERS

Bảng 4.*Mô hình Holt-Winters*

```
METHOD: HoltWinters_Forecast(data, seasonal_type, seasonal_period)
INPUT:
   data: Chuỗi thời gian đầu vào
   seasonal_type: Loại mô hình ('additive' hoặc 'multiplicative')
   seasonal_period: Chu kỳ mùa vụ
OUTPUT:
BEGIN
   forecast: Chuỗi các giá trị dự báo
1. IF seasonal_type == 'additive':
2.    model = ExponentialSmoothing(data, trend='add', seasonal='add', seasonal_periods=seasonal_period)
3. ELSE IF seasonal_type == 'multiplicative':
4.    model = ExponentialSmoothing(data, trend='mul', seasonal='mul', seasonal_periods=seasonal_period)
5. model_fit = model.fit()
6. forecast = model_fit.forecast(steps=horizon)
7. RETURN forecast
END
```

4. MÔ HÌNH FACEBOOK PROPHET

Bảng 5.*Mô hình Facebook Prophet*

```
METHOD: Prophet_Forecast(data, horizon)
INPUT:
   data: Chuỗi thời gian đầu vào
   horizon: Số bước dự báo
OUTPUT:
BEGIN:
   forecast: Dataframe chứa các giá trị dự báo
1. df = convert_to_prophet_format(data)
2. model = Prophet()
3. # Điều chỉnh các tham số nếu cần
4. model.fit(df)
5. future = model.make_future_dataframe(periods=horizon)
6. forecast = model.predict(future)
7. RETURN forecast
END
```

V. KẾT QUẢ THỰC NGHIỆM

Sau khi thực hiện thu thập dữ liệu, tiền xử lý dữ liệu. Chúng ta tiến hành xây dựng bốn mô hình thống kê để ứng dựng vào việc dự đoán xu hướng giá chứng khoán: ARIMA, VAR, Holt-Winters, Facebook Prophet. Với mô hình ARIMA sẽ sử dụng thư viện statsmodels với từ khóa *ARIMA* để gọi ra mô hình; mô hình VAR sẽ sử dụng statsmodels với từ khóa *VAR*; mô hình Holt-Winters sẽ sử dụng thư viện statsmodels với từ khóa *ExponentialSmoothing* và mô hình Facebook Prophet sẽ sử dụng thư viện prophet để xây dựng mô hình.

Sử dụng hai độ đo Root Mean Square Error (RMSE) và Mean Absolute Error (MAE) từ thư viện scikit-learn để đánh giá hiệu suất của các mô hình.

(i) MAE [15] là 1 metric đánh giá mô hình bằng cách tính trung bình giá trị tuyệt đối sai số giữa giá trị thực tế và giá trị dự đoán được tính theo công thức 6. Trong đó: x_i là các giá trị thực tế và y_i là các giá trị dự đoán

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \qquad (6)$$

(ii) RMSE [15] là căn bậc hai của MSE. RMSE đo lường mức độ phù hợp của các giá trị dự đoán từ mô hình so với các giá trị thực tế được quan sát trong tập dữ liệu được tính theo công thức 7.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (7)$$

Các thông số ước lượng bằng cách xác định độ trễ bằng ACF và PACF để xây dựng cho mô hình ARIMA cho từng mã như trong Bảng 6.

Bảng 6. *Thông số ước lượng mô hình ARIMA cho từng mã*

Mã	Thông số ước lượng	Mã	Thông số ước lượng
SSI	ARIMA(0, 1, 0)	POW	ARIMA(2, 1, 2)
ACB	ARIMA(0, 1, 0)	SAB	ARIMA(0, 1, 0)
BCM	ARIMA(0, 1, 0)	SHB	ARIMA(1, 1, 1)
BID	ARIMA(0, 1, 0)	SSB	ARIMA(0, 1, 0)
BVH	ARIMA(0, 1, 0)	STB	ARIMA(0, 1, 0)
CTG	ARIMA(0, 1, 0)	TCB	ARIMA(0, 1, 0)
FPT	ARIMA(0, 1, 0)	TPB	ARIMA(0, 1, 0)
GAS	ARIMA(0, 1, 0)	VCB	ARIMA(4, 1, 4)
GVR	ARIMA(0, 1, 0)	VHM	ARIMA(0, 1, 0)
HDB	ARIMA(0, 1, 0)	VIB	ARIMA(0, 1, 0)
HPG	ARIMA(0, 1, 0)	VIC	ARIMA(0, 1, 0)
MBB	ARIMA(0, 1, 0)	VJC	ARIMA(6, 1, 6)
MSN	ARIMA(3, 1, 3)	VNM	ARIMA(0, 1, 0)
MWG	ARIMA(0, 1, 0)	VPB	ARIMA(0, 1, 0)
PLX	ARIMA(5, 1, 5)	VRE	ARIMA(0, 1, 0)

Mô hình VAR được ước lượng thông số mô hình bằng cách xác định độ trễ bằng AIC với kết quả trả về thông số tốt nhất để xây dựng mô hình là VAR (1).

Với mô hình Holt-Winters trải thì các thông số alphas, betas, gammas sẽ được tổ hợp ba thông số này từ 0.1 đến 0.9 với bước nhảy là 0.1. Sau đó thực hiện vòng lặp, lặp qua từng tổ hợp để chọn ra thông số tốt nhất cho mô hình. Sau đó chọn chế độ cho mô hình Holt-Winters là Additive hay Multiplicative.

Mô hình Facebook Prophet là mô hình thống kê có tính chất mùa vụ nên các thông số sẽ phụ thuộc vào xu hướng, chu kỳ theo ngày, tháng, năm. Ở đây, khi xây dựng mô hình này thì thông số *changepoint_prior_scale* hay độ nhạy điểm thay đổi xu hướng thường được đặt ở mặc định là 0.05. Tiếp đến là thông số *seasonality_prior_scale* là thông số làm mịn các thành phần mùa vụ thì sẽ lấy ở mức phản ánh vừa đủ các dao động là 10. Thông số chu kì diễn ra trong 1 tháng tức là 30 ngày. Chúng ta sẽ chọn chế độ cho mô hình là Additive hay Multiplicative và bật chế độ mùa vụ theo tháng và theo tuần trong mô hình Facebook Prophet để tiến hành dự đoán.

Bảng 7. *Kết quả trung bình thang đo đánh giá MAE, RMSE*

Model	MAE	RMSE
ARIMA	42557,947	42486,06
VAR	4400,065	3817,61
Additive Holt-Winters	3533,63	4077,04
Multiplicative Holt-Winters	3557,57	4091,04
Additive Seasonality Facebook Prophet	3358,7	4003,01
Multiplicative Seasonality Facebook Prophet	3887,96	4605,6

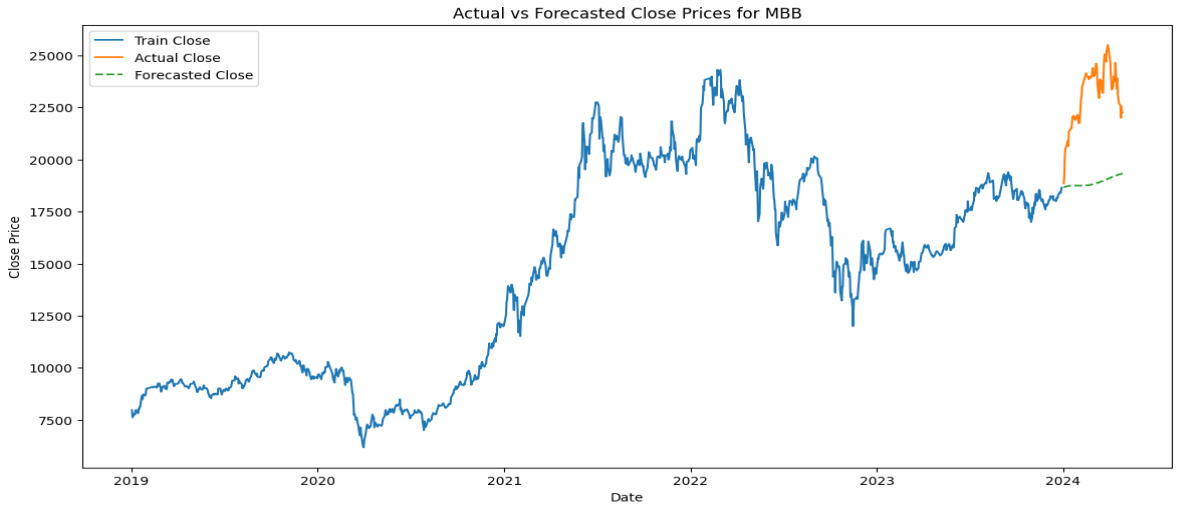
Dựa vào Bảng 7 ta thấy mô hình cộng Facebook Prophet có kết quả dự báo tốt nhất với MAE thấp nhất và RMSE cũng thấp, chứng tỏ độ chính xác cao trong việc dự báo chứng khoán. Mô hình VAR và Holt-Winters (cả cộng và nhân) cũng cho kết quả khá tốt với MAE và RMSE thấp. Trong khi đó, mô hình ARIMA có kết quả dự báo kém nhất với MAE và RMSE rất cao, cho thấy mô hình này không phù hợp với dữ liệu chứng khoán phức tạp và có yếu tố mùa vụ và xu hướng phi tuyến tính.

Kết quả đánh giá thang đo ở trên là tổng hợp của 30 mã chứng khoán. Việc mô hình ARIMA cho kết quả không cao là do dữ liệu có tính chất mùa vụ mạnh, mô hình không có các biến xử lý tình mùa vụ phù hợp và dữ liệu được trải dài trong gần 5 năm nên cũng có thể ảnh hưởng đến hiệu suất mô hình.

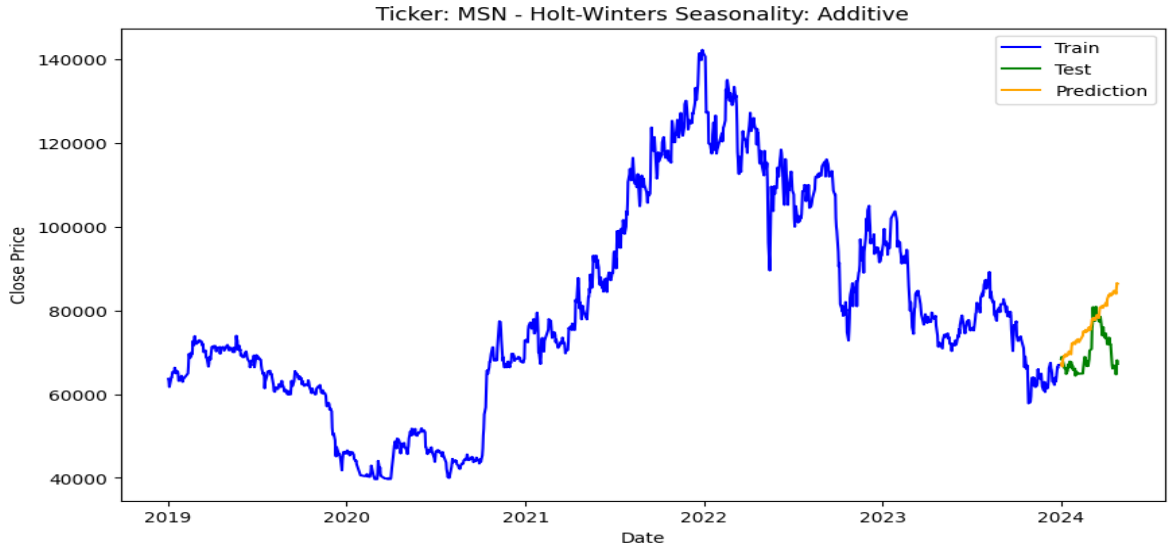
Mô hình VAR vì chỉ có thành phần tự hồi quy cho nên kết quả mô hình có độ chính xác cải thiện hơn so với mô hình ARIMA. Riêng hai mô hình Holt-Winters và Facebook Prophet đều có thành phần chu kỳ, mùa vụ trong việc xây dựng mô hình nên kết quả dự đoán cho ra gần với kết quả thực tế mà chúng ta thu thập được. Và dựa vào thang đo đánh giá và các biểu đồ trực quan kết quả đã cho thấy mô hình mới nhất Facebook Prophet là mô hình thể hiện được xu hướng lên xuống rõ ràng hơn trên toàn bộ tập dữ liệu chứng khoán.

Một số kết quả trực quan dự đoán của các mô hình được thể hiện trong Hình 6, Hình 7, Hình 8 và Hình 9.

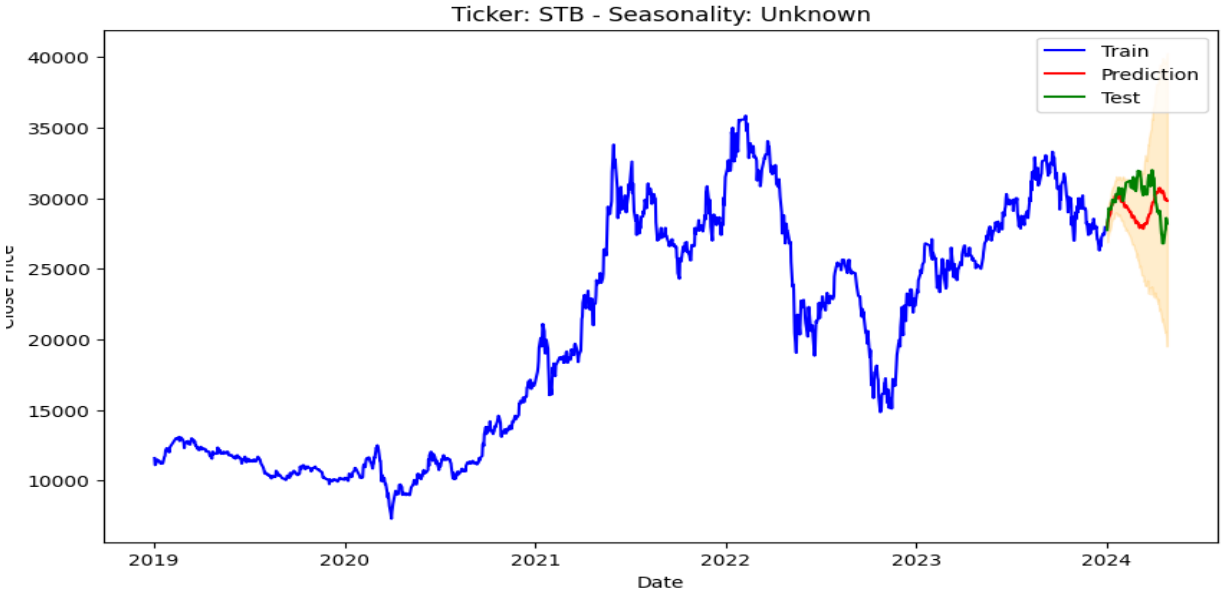
Hình 6. *Kết quả dự đoán của mô hình ARIMA*



Hình 7. *Kết quả dự đoán của mô hình VAR*



Hình 8. *Kết quả dự đoán của mô hình Holt-Winters*



Hình 9. *Kết quả dự đoán trên mô hình Facebook Prophet*

Dựa trên kết quả của các biểu đồ trực quan, đồng thời dựa trên kết quả của thang đo đánh giá trên các mô hình ta có thể nhận xét như sau:

Mô hình ARIMA (Hình 6): Kết quả dự đoán của mô hình ARIMA chỉ là một đường thẳng lên hoặc xuống. Dựa vào thang đo thì cũng thấy được rằng sự chênh lệch giữa giá trị dự đoán và giá trị thực tế rất lớn. Điều này cho thấy ARIMA không phù hợp để xây dựng mô hình dự đoán trên tập dữ liệu này.

Mô hình VAR (Hình 7): Đường dự đoán của mô hình VAR có đặc trưng là những có hình dạng cánh cung, với những sự biến thiên thay đổi rất nhỏ. Đường dự đoán thường là những đường vòng cung đi từ dưới lên hoặc đi từ trên xuống. Kết quả dự đoán trên mô hình VAR đối với nhiều mã có sự chênh lệch rất lớn giữa các giá trị thực tế và các giá trị dự đoán. Nó được thể hiện rõ ở các điểm đánh giá thang đo MAE, RMSE của mô hình.

Mô hình Holt-Winters (Hình 8): Đường dự đoán trên mô hình cộng và mô hình nhân của Holt-Winters là những đường có hình dạng sóng lên xuống ở các điểm. Đặc điểm của đường dự đoán này thể hiện theo 1 chiều hướng nhất định bao gồm chiều hướng đi lên, đi xuống hoặc đi ngang. Sự chênh lệch của kết quả dự đoán giữa hai mô hình có sự khác biệt rõ. Đối với mô hình cộng, bám sát với giá trị thực tế, kết quả thang đo đánh giá thấp hơn so với mô hình nhân. Tuy nhiên đối với một số mã chứng khoán thì kết quả dự đoán trên mô hình nhân lại đưa ra kết quả tốt hơn ví dụ như mã FPT.

Mô hình Facebook Prophet (Hình 9): Đối với đường dự đoán của mô hình Prophet, đường dự đoán thể hiện hình dạng lên xuống ở các điểm giá trị, gần sát với hình dạng biến thiên của một dữ liệu chuỗi thời gian. Kết quả của hai mô hình cộng và nhân đều đưa ra kết quả trên toàn tập dữ liệu rất sát với dữ liệu thực tế trên tập kiểm thử. Kết quả thang đo cho thấy, sự chênh lệch giữa các điểm giá trị thực tế và giá trị dự đoán không quá cao so với các mô hình còn lại được đề cập trong bài báo cáo.

Qua kết quả cho thấy mô hình Facebook Prophet cho ra kết quả tốt hơn so với các mô hình còn lại là vì mô hình này có áp dụng đặc điểm mùa vụ của chuỗi thời gian vào việc dự đoán. Cho nên kết quả của mô hình Facebook Prophet sẽ bám sát với các giá trị dữ liệu thực tế.

VI. KẾT LUẬN

Với việc xây dựng và ứng dụng các mô hình dự đoán dữ liệu chuỗi thời gian trên tập dữ liệu chứng khoán mang đến một số ý nghĩa, lợi ích trong lĩnh vực kinh tế - tài chính. Việc xây dựng các mô hình dự đoán xu hướng giá chứng khoán có thể giúp cho những người mới tham gia vào lĩnh vực này có thể đưa ra các lựa chọn đầu tư phù hợp hoặc đối với các chuyên gia nghiên cứu về lĩnh vực kinh tế thì các mô hình dự đoán xu hướng giá chứng khoán có thể giúp xác định các yếu tố tác động đến lĩnh vực này. Với sự giúp đỡ của các mô hình dự đoán, các chuyên gia có thể phát hiện, đưa ra cảnh báo những yếu tố có ảnh hưởng tới nền kinh tế.

Với các kết quả thực nghiệm trên các mô hình ARIMA, VAR, Holt-Winters, Facebook Prophet được huấn luyện trên tập dữ liệu chứng khoán cho thấy việc lựa chọn những mô hình phù hợp với dữ liệu từng mã chứng khoán giúp tăng độ chính xác cho việc dự đoán xác với thực tế. Kết quả của bài báo cho thấy Facebook Prophet là mô hình tối ưu nhất với việc đưa ra các dự đoán.

VII. TÀI LIỆU THAM KHẢO

[1] Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M. (1978). Time Series Analysis: Forecasting and Control, The Statistician, Vol. 27, pp. 265-265.

[2] Pham Đình Khánh (2019), Mô hình ARIMA trong time series, https://phamdinhkhanh.github.io/2019/12/12/ARIMAmodel.html, Accessed date: Jun 30, 2024.

[3] Adam Hayes (2024). Autoregressive Integrated Moving Average (ARIMA) Prediction Model, https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp, Accessed date: Jul 02, 2024.

[4] N. Hashimzade, M. A. Thornton (2013). Handbook of Research Methods and Applications in Empirical Macroeconomics, Edward Elgar Pub.

[5] Nurhamidah, Nusyirwan, Ahmad Faisol (2020). Forecasting Seasonal Time Series Data using The Holt-Winters Exponential Smoothing Method of Additive Models, Jurnal Matematika Integratif, Vol. 16, No. 2, pp. 151-157.

[6] B. Letham, S. J. Taylor (2017). Forecasting at Scale. DOI: https://doi.org/10.7287/peerj.preprints.3190v2

[7] Prophet (2017). Meta Open Source, https://facebook.github.io/prophet/, Accessed date: Aug 17, 2024.

[8] M. Vijh, D. Chandola, V. A. Tikkiwal, A. Kumar (2019). Stock Closing Price Prediction using Machine Learning Techniques, Procedia Computer Science, Vol. 167, pp. 599-606.

[9] Abhinit Davane1, Shailesh Bhagat, Mehul Bhoi, Prajaktee Rane (2021). Stock Market Prediction Using Machine Learning International, Research Journal of Engineering and Technology (IRJET). Vol. 8, No. 5, pp. 2289-2292.
[10] S. Shen, H. Jiang, T. Zhang (2012). Stock Market Forecasting Using Machine Learning Algorithms, Stanford University.
[11] Himanshu Gupta, Aditya Jaiswal (2024). A Study on Stock Forecasting Using Deep Learning and Statistical Models, ArXiv, pp. 1-9. DOI: https://doi.org/10.48550/arXiv.2402.06689.
[12] A. Koenecke (2020). Applying Deep Neural Networks to Financial Time Series Forecasting, Mining Data for Financial Applications, pp. 1-21.
[13] Alexiei Dingli, Karl Sant Fournier (2017). Financial Time Series Forecasting – A Deep Learning Approach, International Journal of Machine Learning and Computing, Vol. 7, No. 5, pp. 118-122.
[14] T. Vũ (2022). VNStocks, https://vnstocks.com/docs/category/t%C3%A0i-li%E1%BB%87u, Accessed date: Jul 16, 2024.
[15] N. K. Reddy (2024). Regression Metrics, https://www.geeksforgeeks.org/regression-metrics/, Accessed date: Aug 09, 2024.

COMPARISON OF FORECASTING MODELS FOR STOCK PRICE PREDICTION

Chu Dang Binh An, Hoang Dinh Thang, Tran Minh Thai

ABSTRACT— Predictive problems using machine learning models are fundamental and widely applied in various domains, including weather forecasting, healthcare, and financial markets. These problems focus on predicting the outcomes of events or future values based on historical data values through the construction of predictive models. This study focuses on developing predictive models for time series data using a stock dataset from the VNINDEX exchange. Through analysis techniques, data preprocessing, parameter selection appropriate for each model characteristic, and the construction and training of machine learning models to predict stock price trends, several representative methods are used, including Autoregressive Integrated Moving Average, Vector Autoregression, Holt-Winters, and Facebook Prophet. Experimental results show that the Facebook Prophet outperforms other methods in accuracy predictions and overall performance.

Keywords— Time series forecasting, stock market, ARIMA, VAR, Holt-Winters, Facebook Prophet.

Trần Minh Thái là tiến sỹ Công nghệ thông tin (CNTT). Hiện tại, TS. Thái là giảng viên và trưởng bộ môn Hệ thống Thông tin thuộc khoa Công nghệ thông tin, Trường Đại học Ngoại ngữ - Tin học TP.HCM. Lĩnh vực nghiên cứu của TS. Thái liên quan đến vấn đề khai thác dữ liệu, ẩn dữ liệu, xử lý dữ liệu lớn và nhận dạng.

Chu Đặng Bình An hiện là sinh viên chuyên ngành Khoa học dữ liệu, khoa Công nghệ thông tin tại Trường Đại học Ngoại ngữ - Tin học TP.HCM (HUFLIT).

Hướng nghiên cứu chính: Khai thác dữ liệu, học máy, trí tuệ nhân tạo.

Hoàng Đình Thăng hiện là sinh viên chuyên ngành Khoa học dữ liệu, khoa Công nghệ thông tin tại Trường Đại học Ngoại ngữ - Tin học TP.HCM (HUFLIT).

Hướng nghiên cứu chính: Phân tích và tiền xử lý dữ liệu, học máy, trí tuệ nhân tạo.