

Xây dựng và đánh giá hiệu quả các mô hình học máy vào bài toán dự báo giá cổ phiếu của 3 ngân hàng hàng đầu Việt Nam

Building and evaluating machine learning models in predicting stock prices of the top 3 banks in Vietnam

1st Trịnh Gia Huy

IS403.N21

Đại học Công nghệ thông tin

Hồ Chí Minh, Việt Nam

20520556@gm.uit.edu.vn

2nd Phạm Lê Dịu Ái

IS403.N21

Đại học Công nghệ thông tin

Hồ Chí Minh, Việt Nam

20520368@gm.uit.edu.vn

3rd Nguyễn Thị Thảo Hồng

IS403.N21

Đại học Công nghệ thông tin

Hồ Chí Minh, Việt Nam

20520192@gm.uit.edu.vn

4th Lương Nguyễn Thành Nhân

IS403.N21

Đại học Công nghệ thông tin

Hồ Chí Minh, Việt Nam

20520667@gm.uit.edu.vn

5th Lâm Võ Khánh My

IS403.N21

Đại học Công nghệ thông tin

Hồ Chí Minh, Việt Nam

20520912@gm.uit.edu.vn

Tóm tắt nội dung—Để nắm bắt các biến động đại dịch covid-19 ảnh hưởng lên thị trường chứng khoán ngân hàng, nhóm sử dụng mười một mô hình là ARIMA, ARIMAX, SARIMAX, Random Forest, Linear Regression, RNN, LSTM, BNN, GRU, XGBoost, CNN-LSTM để dự báo giá cổ phiếu của các ngân hàng là VCB, BID, STB. Đánh giá ba độ đo RMSE, MAPE và MDA, tìm ra hai mô hình tốt nhất để tiếp tục dự báo giá 30 ngày tiếp theo. Kết quả thực nghiệm cho thấy 2 mô hình RNN, GRU cho ra độ đo tốt nhất trên RMSE lần lượt là 1085.73 và 1420.37 với cổ phiếu VCB, 873.23 và 926.10 với cổ phiếu STB, 1052.78 và 1043.37 với cổ phiếu BID với tỉ lệ chia dữ liệu là 6-3-1.

Index Terms—học máy, dự báo, chuỗi thời gian, Việt Nam, ngân hàng, cổ phiếu, chứng khoán, ARIMA, ARIMAX, SARIMAX, Linear Regression, Random Forest, RNN, LSTM, BNN, GRU, CNN-LSTM, XGBoost, BID, STB, VCB.

I. GIỚI THIỆU

Trong khoảng thời gian gần đây, ngành ngân hàng Việt Nam đã trải qua một sự phát triển mạnh mẽ, đóng góp quan trọng vào sự phát triển kinh tế của đất nước. Dữ liệu cổ phiếu ngân hàng cung cấp thông tin quan trọng về hiệu suất tài chính và biến động của thị trường, giúp hiểu rõ hơn về sự phát triển và tiềm năng của ngành ngân hàng Việt Nam. Năm 2020, thị trường chứng khoán Việt Nam đã gặp nhiều thách thức khi mà đại dịch COVID-19 xuất hiện và tác động lên thị trường Việt Nam khi mà chỉ số thị trường chứng khoán tại Việt Nam là VN-Index ghi nhận sự sụt giảm từ 991 điểm xuống 643 điểm trong Q1 2020[1] kéo theo sự phát triển của các ngân hàng. Vì thế với mục tiêu phân tích xu hướng, biến động dữ liệu cổ phiếu và dự báo sự tăng giảm giá cả trong tương lai, nhóm nghiên cứu đã chọn dự báo giá cổ phiếu của ba ngân hàng Việt Nam là Vietcombank, Sacombank, BIDV bằng cách sử dụng các mô hình học máy và học sâu để phân tích bài toán dự báo giá cổ phiếu.

Trong bài nghiên cứu này, nhóm sẽ trình bày đánh giá thực nghiệm về 11 mô hình dự báo chuỗi thời gian - time series để dự báo giá cổ phiếu. Cụ thể, 11 mô hình được dùng là: ARIMA, ARIMAX, SARIMAX, Linear Regression, Random Forest, RNN, LSTM, BNN, GRU, CNN-LSTM, XGBoost. Việc đánh giá hiệu quả các mô hình sẽ dựa trên 3 độ đo là RMSE, MDA, MAPE để tìm ra 2 mô hình tốt nhất để dự báo giá của 30 ngày tiếp theo của bộ dữ liệu từ đó rút ra những nhận xét về các ưu - nhược điểm của từng mô hình đối với bài toán chuỗi thời gian nói chung và bài toán dự báo giá cổ phiếu nói riêng.

II. CÁC NGHIÊN CỨU LIÊN QUAN

ARIMA và SARIMAX là các mô hình dự báo sử dụng phổ biến trong bài toán chuỗi thời gian, như trong việc dự báo giá cổ phiếu tại Nigeria và New York [2]. Từ kết quả, ARIMA có thể dự báo giá cổ phiếu trong trung và ngắn hạn của chỉ số S&P BSE IT và S&P BSE Sensex trên Sàn Giao dịch Chứng khoán Bomba tại Ấn Độ[3]. Trong bài nghiên cứu chiến tranh thương mại giữa Hoa Kỳ và Trung Quốc có ảnh hưởng đến giá dầu Brent hay không. Tác giả Ilma Amira Rahmayanti và các đồng sự đã dựa vào ARIMAX và có thể kết luận rằng giá dầu Brent có bị ảnh hưởng. [4] Đối với Nari Sivanandam Arunraj và các cộng sự thì SARIMAX được đề xuất cho ra kết quả tốt hơn khi mà R^2 cải thiện từ 0.386 lên 0.613 so với SARIMA khi dự báo doanh thu hằng ngày của nhà bán lẻ thực phẩm [5]. Sreelekshmy Selvin cùng đồng sự cho thấy RNN có phần trăm lỗi - error percentage lần lượt trên các công ty Infosys, TCS và Cipla là 3.90%, 7.65%, 3.83%, thấp hơn nhiều so với mô hình ARIMA[6].

Trong nghiên cứu của Abdullah Bin Omar và đồng sự, Random Forest cho thấy hiệu quả khá ổn định với các bộ dữ liệu quan sát nhỏ và thời gian dự báo ngắn hạn khi được so

sánh với mô hình Deep Neural Network tốt hơn khi dự báo dài hạn và tìm ra xu hướng trong thời gian dài[7]. Với Yuqiao Guo, tác giả đã dự báo giá cổ phiếu của Tesla, Amazon, Microsoft bằng LSTM và kết quả dự báo khá tốt [8]. Bên cạnh các phương pháp học máy đơn nhất, việc kết hợp các mô hình học máy đã được Wu và đồng sự chứng minh khi kết hợp CNN và LSTM, mô hình SACLSTM được đề xuất đã đem lại hiệu quả cao hơn so với phương pháp sử dụng riêng biệt CNN hoặc LSTM truyền thống [9]. Livieris và đồng sự cũng đề xuất 2 mô hình kết hợp CNN-LSTM cho bài toán dự báo giá vàng vào [10], mô hình đầu tiên được xây dựng với 2 lớp tích chập - convolutional layer, 1 lớp gộp - pooling và 1 lớp LSTM ít hơn mô hình thứ 2 một lớp fully-connected. Cả 2 mô hình được Livieris giới thiệu đều có hiệu quả cao hơn khi sử dụng các mô hình khác như SVR, FFNN, LSTM đơn lẻ.

Haorui Zhang sử dụng Linear Regression và LSTM để dự đoán giá cổ phiếu của Amazon [11]. Trong bài báo chỉ sử dụng giá cổ phiếu lịch sử của công ty làm nguồn dữ liệu và đánh giá các mô hình qua các độ đo MSE, MAE, RMSE, R^2 . Kết quả cho thấy trong trường hợp nguồn dữ liệu hạn chế, Linear Regression đơn giản nhất thậm chí còn tốt hơn các mô hình LSTM.

Trong nghiên cứu của Li Jidong and Zhang Ran, phương pháp XGBoost được áp dụng để dự đoán hệ số IC (Information Coefficient) trong mô hình lựa chọn cổ phiếu dựa trên nhiều yếu tố, và các hệ số IC được sử dụng để phân bổ trọng số động cho các yếu tố [12]. Trong bài báo này sử dụng ba chiến lược để so sánh xác minh sự hiệu quả của mô hình XGBoost gồm chiến lược phân bổ trọng số đồng đều, phân bổ trọng số IC, phân bổ trọng số động. Kết quả thực nghiệm chứng minh rằng mô hình XGBoost hiệu quả trong việc dự đoán hệ số IC và phân bổ trọng số động dựa trên mô hình đó có thể cải thiện hiệu suất của chiến lược lựa chọn cổ phiếu dựa trên nhiều yếu tố.

Trong bài nghiên cứu của Dang Lien Minh và các đồng sự, GRU hai luồng được áp dụng để dự đoán giá cổ phiếu bằng cách sử dụng tin tức tài chính và lịch sử giá[13]. Kết quả cho thấy GRU hai luồng hoạt động tốt hơn các mô hình tiên tiến nhất với độ chính xác tổng thể vượt trội là 66,32%. Nhận thấy các mạng thần kinh Bayes có thể đưa ra những dự đoán hợp lý với khả năng định lượng không chắc chắn, Rohitash Chandra đã sử dụng BNN để dự báo giá cổ phiếu trong giai đoạn trước và trong COVID-19[14]. Kết quả cho thấy mô hình BNN đưa ra những dự đoán hợp lý với khả năng định lượng độ không chắc chắn mạnh mẽ bất chấp sự biến động của thị trường cao trong giai đoạn đầu của đại dịch COVID-19.

III. DỮ LIỆU

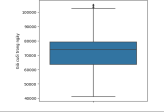
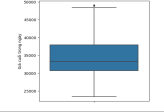
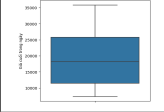
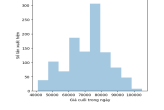
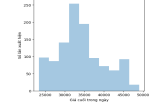
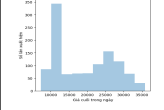
A. Tổng quan về dữ liệu

Bộ dữ liệu được sử dụng trong bài nghiên cứu dựa trên giá cổ phiếu của 3 ngân hàng hàng đầu Việt Nam là BIDV (BID)¹, Vietcombank (VCB)², Sacombank (STB)³ được thu thập từ

¹<https://www.investing.com/equities/commercial-bank-investment-develop>

²<https://www.investing.com/equities/joint-stock-commercial-bank>

³<https://www.investing.com/equities/sai-gon-thuong-tin-commercial>

	VCB	BID	STB
Số lượng mẫu	1112	1112	1112
Trung bình	71385.06	34190.49	19091.30
Trung vị	73967.5	33319.2	18425.0
Mode	63738.0, 75096.0	39000.0	11300.0
Cực tiểu	41161.0	23419.5	7300.0
25%	63584.0	30683.53	11450.0
50%	73967.5	33319.2	18425.0
75%	79225.0	37909.5	26000.0
Cực đại	105000.0	49000.0	35850.0
Độ lệch chuẩn	12609.78	5913.40	7789.38
Phương sai	159006470.64	34968326.22	60674561.30
Hệ số biến thiên	0.18	0.17	0.41
Khoảng biến thiên	63839.0	25580.5	28550.0
Skewness	-0.22	0.40	0.28
Kurtosis	-0.37	-0.49	-1.36
Boxplot			
Histogram			

Bảng I
THỐNG KÊ MÔ TẢ DỮ LIỆU

trang Investing. Dữ liệu được lấy từ ngày 01/01/2019 đến hết ngày 16/06/2023, đơn vị của dữ liệu tính theo ngày. Dữ liệu sẽ được chia thành 3 tập train, test, validate theo tỉ lệ 6-3-1, 7-2-1, 8-1-1 để đánh giá mô hình.

Từ bảng I, chúng ta có thể nhận định một số điểm đáng chú ý về giá cổ phiếu của các công ty. Cụ thể, giá cổ phiếu của VCB có xu hướng cao hơn so với BID và STB khi xét về trung bình, cực tiểu và cực đại. Biểu đồ Boxplot cho thấy VCB và BID có phân bố dữ liệu nhỏ và xuất hiện các giá trị ngoại lai ngoài phạm vi cực đại thể hiện giá cổ phiếu không thay đổi nhiều so với giá trị trung bình nhưng lại có nhiều biến động trong giá. Biểu đồ Histogram thể hiện tần suất hiện của giá cổ phiếu VCB và BID tập trung quanh giá trị trung bình còn STB có tần suất ở mức giá thấp là cao nhất. Bên cạnh đó, hệ số đối xứng Skewness chỉ ra giá cổ phiếu VCB có sự tập trung ở các mức giá thấp khác với các giá cổ phiếu BID và STB tập trung ở các mức giá cao. Ngoài ra, hệ số tập trung Kurtosis nói lên được giá cổ phiếu của cả 3 có sự phân tán dữ liệu lớn, ít tập trung xung quanh giá trị trung bình trong đó STB có sự phân tán dữ liệu lớn nhất.

B. Khai phá dữ liệu

a) *Mùa vụ*: Các bộ dữ liệu theo tuần, tháng hay quý thì không có sự lặp lại của quá trình tăng giảm nên ta có thể kết luận cả 3 bộ dữ liệu đều không thể hiện yếu tố mùa vụ.

b) *Xu hướng*: Nhìn chung, cả ba cổ phiếu VCB, BID, STB đang có xu hướng tăng trong dài hạn với đó là các đợt tăng giảm mạnh trong dữ liệu quá khứ.

c) *Chu kỳ*: Cả 3 bộ dữ liệu đều không thể hiện yếu tố chu kỳ trong dữ liệu. Quan sát các năm, trong khoảng thời gian dài hơn cũng không có sự lặp lại của quá trình tăng giảm.

d) *Nhiều*: Đối với bộ dữ liệu chuỗi thời gian của ta thì khi quan sát ta thấy cả 3 bộ dữ liệu đều không có nhiễu trắng (white noise) khi mà cả 3 bộ dữ liệu đều có trung bình khác 0.

IV. PHƯƠNG PHÁP LUẬN

A. Linear Regression

Linear Regression hay hồi quy tuyến tính là một kỹ thuật phân tích dữ liệu dùng để dự báo giá trị của một biến dựa trên giá trị của một biến khác. Cho phép thiết lập các yếu tố nào là quan trọng nhất, yếu tố nào có thể bỏ qua và cách các yếu tố đó tương tác với nhau. Công thức chung như sau:

$$y = \beta_0 + \beta_1 X + \varepsilon \quad (1)$$

Trong đó, y là giá trị dự báo của biến phụ thuộc với mọi giá trị cho trước của biến độc lập, β_0 là hệ số chặn, giá trị dự báo của y khi x là 0, β_1 là hệ số hồi quy, giá trị thay đổi khi x tăng, X là biến độc lập (biến ảnh hưởng đến giá trị dự báo y) và ε là sai số của ước lượng, hay độ biến thiên có trong ước tính về hệ số hồi quy.

B. Autoregressive integrated moving average - ARIMA

ARIMA là một mô hình chuỗi thời gian được sử dụng để dự báo giá trị của một biến dựa trên các giá trị quá khứ của biến đó. Được giới thiệu bởi Harvey [15] Mô hình gồm ba thành phần: autoregressive (AR), integrated (I), và moving average (MA). Thành phần autoregressive mô tả mối quan hệ giữa giá trị hiện tại và các giá trị quá khứ của biến, thành phần moving average mô tả mối quan hệ giữa giá trị hiện tại và các sai số của mô hình, và thành phần integrated mô tả sự khác biệt giữa các giá trị hiện tại và các giá trị quá khứ.

Mô hình ARIMA(p,d,q) tổng quát có dạng:

$$Y_t = \phi_0 + \phi_1 Y_{(t-1)} + \phi_2 Y_{(t-2)} + \dots + \phi_p Y_{(t-p)} + u_t + \theta_1 u_{(t-1)} + \dots + \theta_q u_{(t-q)} + e_t$$

Trong đó, Y_i là chuỗi dừng bậc d của chuỗi ban đầu (chuỗi khảo sát), ϕ là tham số tự hồi quy, θ là tham số trung bình đi động, u_t là nhiễu trắng, e_t là sai số dự báo, p là bậc tự hồi quy và q là bậc trung bình trượt.

C. Autoregressive Integrated Moving Average with Exogenous Variables - ARIMAX

Mô hình ARIMAX là mô hình mở rộng của ARIMA. Mô hình bao gồm các biến độc lập khác, được thêm vào cuối và được gọi là biến exogenous (biến ngoại sinh). Điều này liên quan đến việc thêm một biến bên ngoài riêng biệt khác để giúp đo lường biến endogenous (biến phụ thuộc) của mô hình. [16] ARIMAX có công thức như sau:

$$(1 - \phi_1 B - \dots - \phi_p B^p) y_t = c + \beta X_t + (1 - \theta_1 B - \dots - \theta_p B^p) \varepsilon_t$$

Trong đó, y_t là biến phụ thuộc của chuỗi thời gian tại thời điểm t , ε_t là sai số tại thời điểm t , ϕ là hệ số của thành phần tự hồi quy, θ là hệ số trung bình trượt, B là toán tử sai phân.

D. Seasonal Autoregressive Integrated Moving Average with Exogenous Variables - SARIMAX

SARIMAX được mở rộng từ mô hình SARIMA, được tăng cường với khả năng tích hợp biến ngoại sinh để cải thiện việc dự báo của mô hình [17]. Mô hình ARIMA truyền thống có nhược điểm đó là không giải thích được dữ liệu có tính mùa vụ - là một chuỗi thời gian có sự lặp lại trong dữ liệu. Vì thế, mô hình SARIMA được sử dụng để giải thích các dữ liệu có tính mùa vụ nhờ bổ sung các thành phần mùa vụ vào mô hình ARIMA thông qua các tham số mới P, D, Q, s .

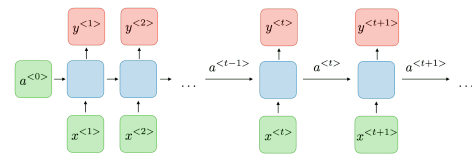
Từ bài báo của Mohamed M. Fathi và các đồng sự thì mô hình tổng quát SARIMAX(p, d, q)(P, D, Q, s) có công thức [18]:

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D Y_t = c + X_t \beta + \theta_q(B)\Theta_Q(B^s)\varepsilon_t \quad (2)$$

Trong đó công thức trên, Y_t là biến phụ thuộc của chuỗi thời gian tại thời điểm t . ε_t là sai số tại thời điểm t . c là hằng số. B là toán tử sai phân ($BY_t = Y_t - 1$). p, q, d lần lượt là các thành phần tự hồi quy, trung bình trượt và sai phân không mùa vụ. P, D, Q lần lượt là các thành phần tự hồi quy, trung bình trượt và sai phân trong mùa vụ. ϕ và Φ lần lượt là các hệ số của thành phần tự hồi quy và trung bình trượt không mùa vụ. θ và Θ lần lượt là các hệ số của thành phần tự hồi quy và trung bình trượt trong mùa vụ. X_t là biến ngoại sinh giải thích tại thời điểm t . β là hệ số biến ngoại sinh của biến ngoại sinh X_t . s là độ dài chu kỳ mùa vụ trong chuỗi thời gian.

E. Recurrent Neural Network - RNN

Cấu trúc mạng lặp lại sử dụng đầu ra - output như là một bộ nhớ linh động, là tiền thân của mô hình RNN ngày nay, được Jordan đề xuất vào năm 1986 [19]. Mô hình RNN dựa trên ý tưởng giống như mô hình mạng nơ-ron truyền thẳng. Sự khác biệt là các đầu ra của mạng nơ-ron truyền thẳng tại bất cứ thời điểm t nào, đều là hàm từ đầu vào - input tại thời điểm hiện tại cùng với trọng số - weight. Trong khi đó các đầu ra của RNN tại thời điểm t thì không chỉ dựa vào đầu vào hiện tại và weight mà còn dựa vào đầu vào trước đó, điều này khắc phục nhược điểm của mô hình mạng nơ-ron truyền thẳng là bị giới hạn bởi nó không giải thích được sự phụ thuộc thời gian - temporal dependencies. Xem minh họa cấu trúc RNN tại hình 1.



Hình 1. Mô hình hóa mô hình RNN

Mô hình RNN được Shuai Li và các đồng sự được mô tả theo công thức sau [20]:

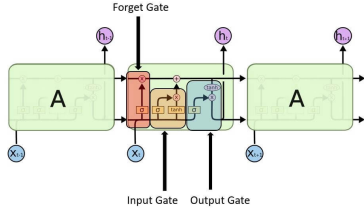
$$\mathbf{h}_t = \sigma(\mathbf{W}\mathbf{X}_t + \mathbf{U}\mathbf{h}_{t-1} + \mathbf{b}) \quad (3)$$

Trong đó $\mathbf{X}_t \in \mathbb{R}^M$ và $\mathbf{h}_t \in \mathbb{R}^N$ lần lượt là các giá trị đầu vào và trạng thái ẩn tại time step. $\mathbf{W} \in \mathbb{R}^{N \times M}$, $\mathbf{U} \in \mathbb{R}^{N \times N}$

và $\mathbf{b} \in \mathbb{R}^N$ lần lượt là các weight của giá trị đầu vào và giá trị lặp lại và các bias của mạng nơ-ron. σ là hàm kích hoạt được sử dụng trong mạng, và N là số lượng nơ-ron được sử dụng trong tầng RNN hiện tại.

F. Long Short-Term Memory - LSTM

So với mạng nơ-ron truyền thống, mạng nơ-ron tái phát (RNN) có bộ nhớ trong, chúng ghi nhớ tất cả thông tin đã được lưu trữ trong quá khứ và sử dụng nó để ra quyết định trong các bước tiếp theo. Mặc dù RNN hoạt động tốt khi xử lý dữ liệu chuỗi ngắn, nhưng chúng gặp hai vấn đề chính là gradient vanishing và gradient exploding.

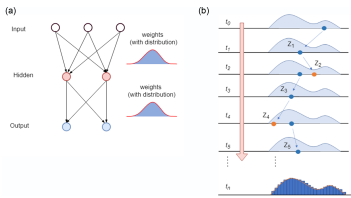


Hình 2. Kiến trúc mô hình LSTM

Để giải quyết các vấn đề của RNN, có thể sử dụng Long Short-Term Memory (LSTM). LSTM hoạt động như một RNN cụ thể với cải tiến trong việc xác định sự phụ thuộc dài hạn trong dữ liệu chuỗi. Khác với RNN, LSTM thực hiện một cách tính toán phức tạp hơn trong việc tính toán trạng thái ẩn bằng cách thay thế các nơ-ron trong lớp ẩn truyền thống bằng các ô nhớ. Chúng có khả năng điều chỉnh thông tin thông qua ba cấu trúc cổng (Input Gate, Output Gate, Forget Gate) và các cổng này cập nhật thông tin theo cách lựa chọn bằng cách học những loại thông tin trong chuỗi quan trọng để giữ lại hoặc xác định thông tin không hữu ích và loại bỏ chúng. [8]

G. Bayesian Neural Network - BNN

Mạng nơ-ron Bayesian cung cấp một phương pháp thực thi xác suất cho một mạng nơ-ron tiêu chuẩn với khác biệt chính là trọng số và độ lệch được biểu diễn thông qua phân phối xác suất hậu nghiệm như được thể hiện trong Hình 3. Tương tự như các mạng nơ-ron tiêu chuẩn, mạng nơ-ron Bayesian cũng có khả năng xấp xỉ hàm liên tục vô hạn. Quá trình suy luận

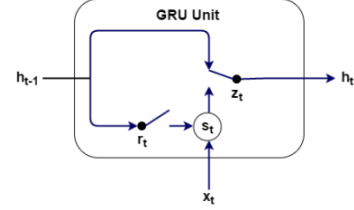


Hình 3. Kiến trúc mô hình BNN

bắt đầu bằng cách thiết lập phân phối tiên nghiệm (prior) trên trọng số và độ lệch. Sau đó, hệ thống lấy mẫu (chẳng hạn như MCMC) sử dụng một hàm hợp lý tính toán khẩu độ huấn luyện và chấp nhận hoặc từ chối một mẫu được đề xuất.

H. Gated Recurrent Unit - GRU

Gated Recurrent Unit (GRU) được đề xuất vào năm 2014 nhằm giải quyết vấn đề gradient biến mất bằng cách thay đổi cách tính toán các đơn vị ẩn. Một đơn vị ẩn ht của GRUs được tính toán dựa trên xt và ht-1, nhưng với cơ chế cổng khác biệt. Hình 4 thể hiện cấu trúc đơn giản của một đơn vị GRU.



Hình 4. Cấu trúc của GRU

Các phương trình sau được sử dụng để tính toán một đơn vị ẩn GRU[21]:

$$z_t = \sigma(x_t U^z + h_{t-1} U^z + b^z) \quad (4)$$

$$r_t = \sigma(x_t U^r + h_{t-1} W^r + b^r) \quad (5)$$

$$s_t = \tanh(x_t U^s + (h_{t-1} \odot r_t) W^s + b^s) \quad (6)$$

$$h_t = (1 - z_t) \odot s_t + z_t \odot h_{t-1} \quad (7)$$

Hàm σ được gọi là hard sigmoid và ký hiệu \odot mang ý nghĩa phép nhân từng phần.

Các cổng GRU r_t và z_t được gọi là cổng đặt lại và cập nhật tương ứng. Cổng đặt lại quyết định cách kết hợp đầu vào mới x_t với bộ nhớ trước đó h_{t-1} để tính toán s_t . s_t có thể xem như một trạng thái ẩn "ứng cử viên". Việc tính toán của h_t sử dụng cổng cập nhật z_t để xác định bao nhiêu bộ nhớ trước đó cần được giữ lại. Một mô-đun GRU với 2 lớp ẩn để nắm bắt các tương tác đặc trưng cấp cao giữa các bước thời gian khác nhau. Các đơn vị trong lớp ẩn thứ hai được tính toán tương tự như trong lớp ẩn đầu tiên.

I. XGBoost

XGBoost hoạt động bằng phương pháp Gradient Boosting - là sự kết hợp việc sử dụng một tập hợp các mô hình yếu (weak learner) để tạo ra một mô hình mới dự báo mạnh hơn. Ý tưởng cơ bản của mô hình gradient boosting là xây dựng một loạt các mô hình dự báo tuần tự, trong đó mỗi mô hình mới cố gắng cải thiện sai số dự báo của mô hình trước đó. Để làm điều này, mô hình tối ưu hóa gradient boosting sử dụng phương pháp gradient descent. Mô hình hóa mô hình dự báo được cải thiện để xấp xỉ giá trị thực tế. [22]

Mô hình minh họa cấu trúc mô hình XGBoost như Hình 5.

Về cơ bản, XGBoost được tính toán tổng quát như sau:

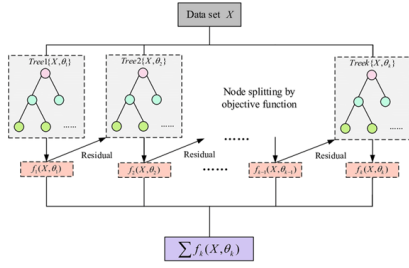
$l(x_1, x_2)$ - cost function

$f_i(x)$: thứ tự tại cây đang thực thi

$\hat{y}_i^0 = 0$: khởi tạo giá trị dự báo ban đầu.

Giá trị đóng góp của cây thứ nhất trong việc dự báo giá trị trong việc dự báo giá trị trong mỗi mẫu dữ liệu (x_i)

$$\hat{y}_i^1 = \hat{y}_i^0 + f_1(x_i)$$



Hình 5. Mô hình hóa mô hình hóa mô hình XGBoost

$$\hat{y}_i^2 = \hat{f}_1(x_i) + f_2(x_i) = \hat{y}_i^1 + f_2(x_i)$$

Quá trình lặp đi lặp lại cho đến khi đạt được giá trị dự báo cuối cùng.

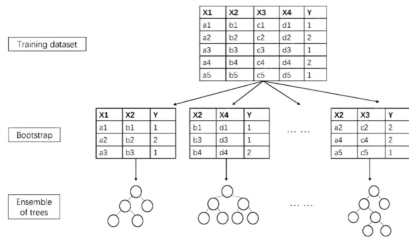
$$\hat{y}_i^t = \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)$$

$$obj^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^n \Omega(f_i)$$

$$= \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + constant$$

J. Rừng ngẫu nhiên - Random Forest

Nhằm khắc phục tình trạng overfitting của mô hình cây quyết định, mô hình Random forest được giới thiệu lần đầu bởi Breiman (2001)[23], tác giả đã đề xuất tạo một tập hợp các cây quyết định được xây dựng dựa trên một tập con ngẫu nhiên của các mẫu dữ liệu với các thuộc tính quyết định cũng được lựa chọn ngẫu nhiên. Dữ liệu của mỗi cây được lấy ngẫu nhiên và có thể trùng lặp(Bootstrapping) để đem lại các kết quả dự báo từ bộ dữ liệu huấn luyện độc lập nhau. Từ đó, kết quả dự báo là giá trị trung bình của tập hợp cây quyết định. Xem minh họa cấu trúc rừng ngẫu nhiên ở Hình 6.



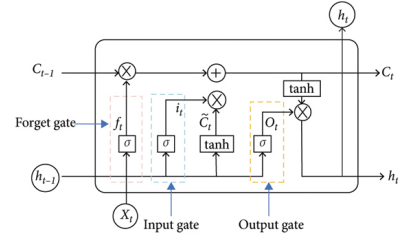
Hình 6. Mô hình hóa mô hình Random Forest

Trong bài nghiên cứu, nhóm sử dụng hàm Random Forest thư viện Scikit-learn với các cây được tạo nên từ mô hình cây quyết định Classification and Regression Trees - CART[24].

K. CNN-LSTM

Mạng nơon tích chập (Convolutional Neural Network) là một loại mạng nơon đặc biệt để xử lý dữ liệu có cấu trúc dạng lưới-ma trận. CNN gồm hai phần chính là Convolutional layer (lớp tích chập) và pooling layer (lớp gộp)[25]. Dữ liệu đầu vào dạng chuỗi thời gian sẽ được xem như ma trận một chiều, được xử lý bởi các lớp tích chập và các hàm kích hoạt. Dữ liệu đầu ra sau đó được các lớp gộp rút trích các đặc trưng của dữ liệu. Việc kết hợp một lượng lớn các lớp giúp cho CNN có ưu thế trong việc rút trích các đặc trưng của quá khứ, tiếp tục đưa vào các lớp của mô hình LSTM nhằm huấn luyện mô

hình. Tổng quát một cấu trúc CNN-LSTM, được tổng quát hóa như Hình 7[26].



Hình 7. Mô hình CNN-LSTM

V. THỰC NGHIỆM

A. Độ đo

Với n ngày được mô hình dự báo, ta có các Y_A là giá trị thực ngày i, Y_B là giá trị thực ngày i-1, f là giá trị dự báo ngày i, có thể xây dựng ba độ đo như sau:

a) *Root Mean Square Error - RMSE*: độ lớn trung bình sai số tuyệt đối giữa giá trị thực tế và giá trị được dự báo.

$$RMSE = \sqrt{\frac{\sum (Y_A - f)^2}{n}}$$

b) *Mean Absolute Percentage Error - MAPE*: tỉ lệ phần trăm trung bình giữa sai số tuyệt đối và giá trị thực tế.

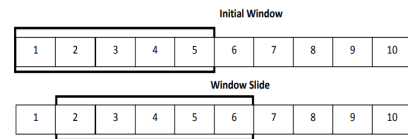
$$MAPE = \frac{1}{n} \sum \left| \frac{Y_A - f}{Y_A} \right|$$

c) *Mean Directional Accuracy - MDA*: tỉ lệ phần trăm các dự đoán xu hướng đúng so với giá trị thực tế.

$$MDA = \frac{1}{n} \sum (1_{(Y_A - f)(Y_A - Y_B) > 0})$$

B. Cửa sổ trượt - Sliding window

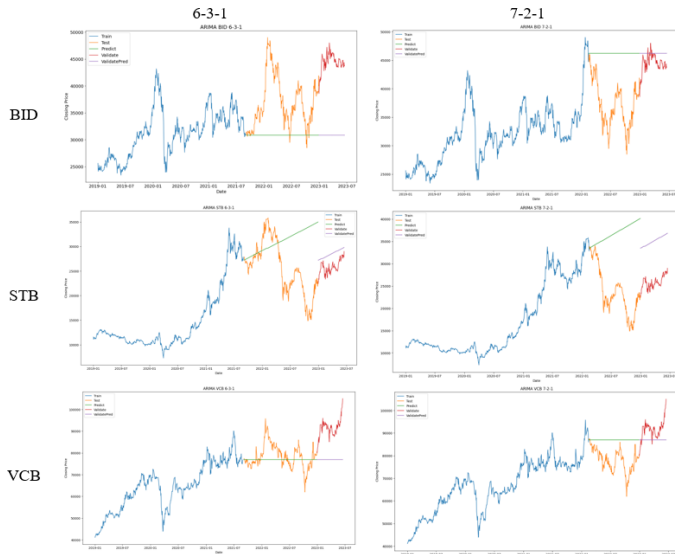
Cửa sổ trượt là kỹ thuật xử lý dữ liệu nhằm biến đổi dữ liệu dạng chuỗi thời gian sang dạng ma trận để phù hợp với giá trị đầu vào của các mô hình học máy, học sâu, như hình 8[27].



Hình 8. Tiến trình cửa sổ trượt

C. ARIMA

Qua Hình 9, ARIMA cho kết quả dự đoán đúng về xu hướng của giá cổ phiếu ở cả 3 bộ dữ liệu và 3 tỉ lệ dữ liệu theo thời gian dài hạn. Tuy nhiên, khi xem xét trên chỉ số RMSE, giá trị chênh lệch giữa giá thực tế và dự báo là quá lớn và khó thể sử dụng. Việc mô hình quá đơn giản trong khi dữ liệu cổ phiếu có nhiều biến động đã làm lộ nhược điểm của các mô hình thống kê như ARIMA vốn phụ thuộc vào các chu kỳ, biến động thường xuyên xảy ra nên mô hình gặp hạn chế khi gặp các biến động bất thường.



Hình 9. Kết quả dự báo mô hình ARIMA

D. ARIMAX

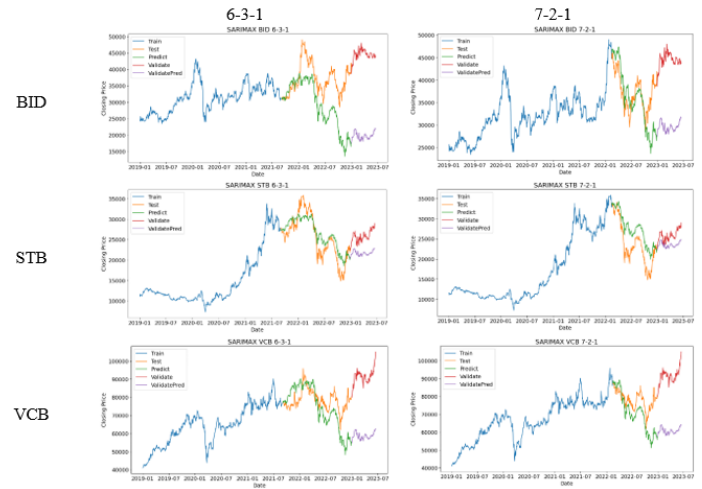
Quan sát kết quả dự báo ở hình 10 và chi tiết ở bảng II, mô hình ARIMAX sử dụng dữ liệu VN-Index làm biến ngoại sinh và cho ra kết quả tốt nhất khi được chia ở tỉ lệ 7-2-1 với RMSE lần lượt là 10713.03, 4207.05 và 5300.1 trên các cổ phiếu VCB, STB, BID. Mô hình dự báo tốt các xu hướng(tăng/giảm) của giá cổ phiếu tương lai, MDA cao tuy nhiên giá trị dự báo đa số đều nhỏ hơn giá trị thực tế ở các trường hợp làm cho chỉ số RMSE và MAPE khi so sánh với các mô hình khác là thấp hơn đáng kể. Đáng lưu ý, các cổ phiếu đều có xu hướng giảm trong tương lai khi được dự báo trong khi thực tế giá các cổ phiếu đều có xu hướng tăng(ngoại trừ STB).



Hình 10. Kết quả dự báo mô hình ARIMAX

E. SARIMAX

Mô hình SARIMAX trong phần thực nghiệm sẽ sử dụng lại bộ dữ liệu VN-Index ở mô hình ARIMAX làm biến ngoại sinh và xét thêm yếu tố mùa vụ trong tập dữ liệu. Kết quả đạt được ở hình 11 cho thấy, khi xét thêm chu kỳ mùa vụ để cải tiến mô hình thì có sự cải thiện khi mà chỉ số RMSE trên VCB, tỉ lệ 6-3-1, 7-2-1 lần lượt là 10591.88, 10550.70 thấp hơn so với mô hình ARIMAX. Ngược lại, khi xét thêm yếu tố mùa vụ, khả năng dự báo dự báo xu hướng tăng giảm của mô hình kém đi so với ARIMAX trên tất cả bộ dữ liệu. Nhìn chung, sự cải thiện của mô hình SARIMAX khi xét thêm yếu tố mùa vụ là không đáng kể và sai số dự báo vẫn còn rất lớn khi xét thêm biến ngoại sinh.



Hình 11. Kết quả dự báo mô hình SARIMAX

F. Linear Regression

Thực hiện huấn luyện dữ liệu bằng mô hình Linear Regression trước tiên xác định biến Price phụ thuộc vào Index được biểu diễn qua hình 12. Mô hình dự đoán khi được chia ở tỉ lệ 6-3-1 với RMSE lần lượt với 3 độ dữ liệu BID, STB, VCB lần lượt là 7218.50, 12346.95 và 4843.72. Kết quả cho thấy giá trị dự đoán có sự chênh lệch nhiều so với kết quả thực tế, đối với sự biến động nhỏ sử dụng mô hình quá đơn giản cho ra kết quả không tốt. Với tập validation, Linear Regression cũng cho ra các nhận định tương tự.

G. Random Forest

Từ kết quả dự đoán ở hình 13, mô hình Random Forest dự đoán tốt nhất ở tỉ lệ 7-2-1 trên cả ba bộ dữ liệu, RMSE lần lượt là 1617.63, 1797.28 và 1062.12 tại bộ dữ liệu VCB, STB và BID. Mô hình có độ nhạy với các biến động tốt, tuy nhiên với các biến động lớn, mô hình lộ ra các khuyết điểm khi không thể đưa ra các quyết định phù hợp (không tăng/giảm mạnh như thực tế mà chỉ tối đa ở mức trần của dữ liệu huấn luyện quá khứ). Khi so sánh, Random Forest vẫn cho kết quả tốt hơn cả những mô hình học sâu như LSTM, CNN-LSTM



Hình 12. Kết quả dự báo mô hình Linear Regression

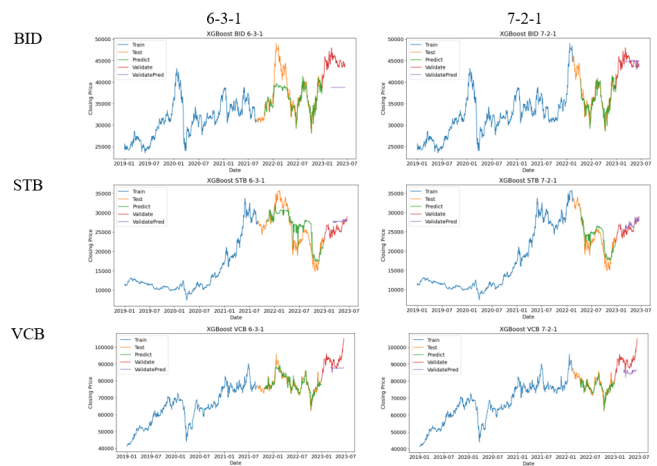
ở nhiều trường hợp khi xét trên độ đo RMSE hay MDA. Trên tập Validation, mô hình cũng có những thể hiện tương tự.



Hình 13. Kết quả dự báo mô hình Random Forest

H. XGBoost

Kết quả nhận được khi thực hiện huấn luyện tập dữ liệu qua mô hình XGBoost như hình 14. Mô hình dự đoán tốt nhất được chia theo tỉ lệ 7-2-1 đối với bộ dữ liệu BID và STB cho kết quả RMSE lần lượt là 1202.68 và 1996.40, tương đồng với độ đo MAPE cũng cho kết quả tốt đối với hai bộ dữ liệu trên. Còn đối với bộ dữ liệu VCB tỉ lệ 6-3-1 lại cho kết quả tốt hơn với RMSE là 2007.07 và MAPE là 1.91. Với khả năng xử lý dữ liệu lớn và mỗi mô hình mới cố gắng cải thiện sai số thì mô hình XGBoost huấn luyện tốt đối với bộ dữ liệu có độ biến động lớn.



Hình 14. Kết quả dự báo mô hình XGBoost

I. RNN

Quan sát hình 18, ta thấy được mô hình RNN cho được kết quả rất tốt khi mà độ đo RMSE của mô hình RNN cho ra kết quả thấp nhất so với các mô hình khác đối với dữ liệu VCB, BID ở các tỉ lệ 6-3-1, 7-2-1. MAPE của mô hình cũng cho được các kết quả tốt khi mức độ dao động của dự báo tốt nhất chỉ ở khoảng 0.9 - 2.13%. Nhưng dù vậy mô hình vẫn chưa tốt trong khả năng dự báo xu hướng tăng hay giảm khi chỉ số MDA đạt được các lần tốt nhất là dữ liệu BID với STB đều đạt được 52.86% ở tỉ lệ chia 6-3-1.

J. GRU

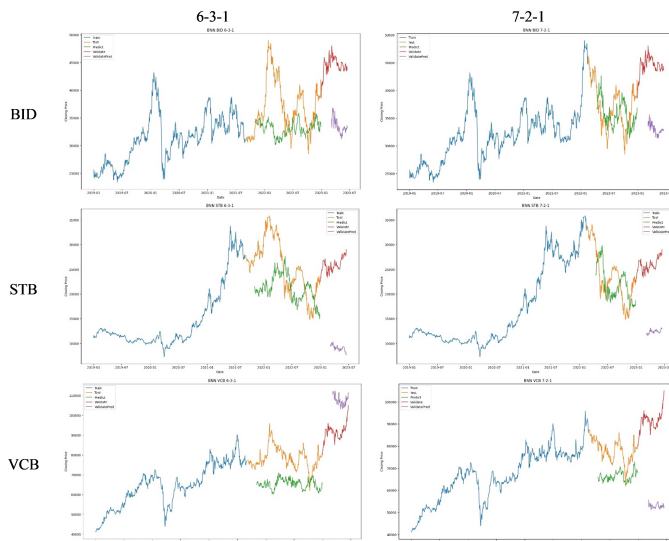
Từ kết quả dự đoán của Hình 18 và Bảng II, ta nhận thấy mô hình cho kết quả dự đoán rất sát với thực tế trên tất cả các tỉ lệ và tập dữ liệu. Với chỉ số MAPE dao động từ 1.34 cho đến 2.81 trên tập Test và RMSE đều tốt ở các tỉ lệ và tập dữ liệu, GRU là một trong hai mô hình cho kết quả dự đoán tốt nhất. Với tập Validation, GRU cũng cho ra các nhận định tương tự.

K. BNN

Kết hợp kết quả dự báo ở Hình 15 và chi tiết ở Bảng II, ta nhận thấy giá trị dự báo của BNN lệch khá nhiều so với thực tế với chỉ số RMSE lên tới 14520.91 (VCB 6-3-1) trên tập Test. Tuy kết quả MDA cao nhưng chỉ số RMSE và MAPE khi so sánh với các mô hình khác là thấp hơn đáng kể. Ở cả 3 tập dữ liệu, BNN đều cho kết quả tốt hơn ở tỉ lệ 7-2-1 và có xu hướng dự đoán giảm trên hầu hết các trường hợp (trừ tập Val của VCB 6-3-1).

L. LSTM

Mô hình LSTM cho kết quả như hình 16 và bảng II. Mô hình dự báo tốt nhất ở tỉ lệ 7-2-1 đối với dữ liệu BID, STB cho kết quả RMSE lần lượt là 1472.26, 1545.31. Đối với tập dữ liệu VCB thì LSTM cho ra kết quả tốt nhất ở tỉ lệ 6-3-1 với RMSE là 2280.42. LSTM cho ra kết quả dự báo khá tương đồng đối với thực tế.



Hình 15. Kết quả dự báo mô hình BNN



Hình 16. Kết quả dự báo mô hình LSTM

M. CNN-LSTM

Từ Hình 17, mô hình cho kết quả RMSE thấp nhất ở tỉ lệ 7-2-1 trên tập VCB là 3455.59, STB là 2912.51 và BID là 2576.79. Tương đồng với độ đo MAPE cũng tốt nhất trên tỉ lệ 7-2-1. Mô hình được mong chờ sẽ có thể tốt hơn so với mô hình LSTM đơn lẻ, tuy nhiên việc xây dựng các lớp chưa hợp lý có thể đã dẫn đến việc mô hình có độ nhạy với các biến động dữ liệu bị hạn chế, các dự báo về các mức giảm/tăng mạnh thường xảy ra trễ hơn thực tế dẫn đến các chỉ số RMSE, MAPE phụ thuộc vào mức chênh lệch dự báo có kết quả tệ hơn so với các mô hình khác. Tuy nhiên, khi xem xét đơn thuần việc dự báo tăng/giảm của giá cổ phiếu, MDA của CNN-LSTM cho ra khá tốt (49.14 ở VCB 6-3-1, 55.56 ở

BID 7-2-1, 51.67 ở STB 7-2-1), đều ở mức cao hơn so với các mô hình khác cho thấy tiềm năng của mô hình ở bài toán này. Các nhận định cũng đúng trên tập validation.



Hình 17. Kết quả dự báo mô hình CNN-LSTM

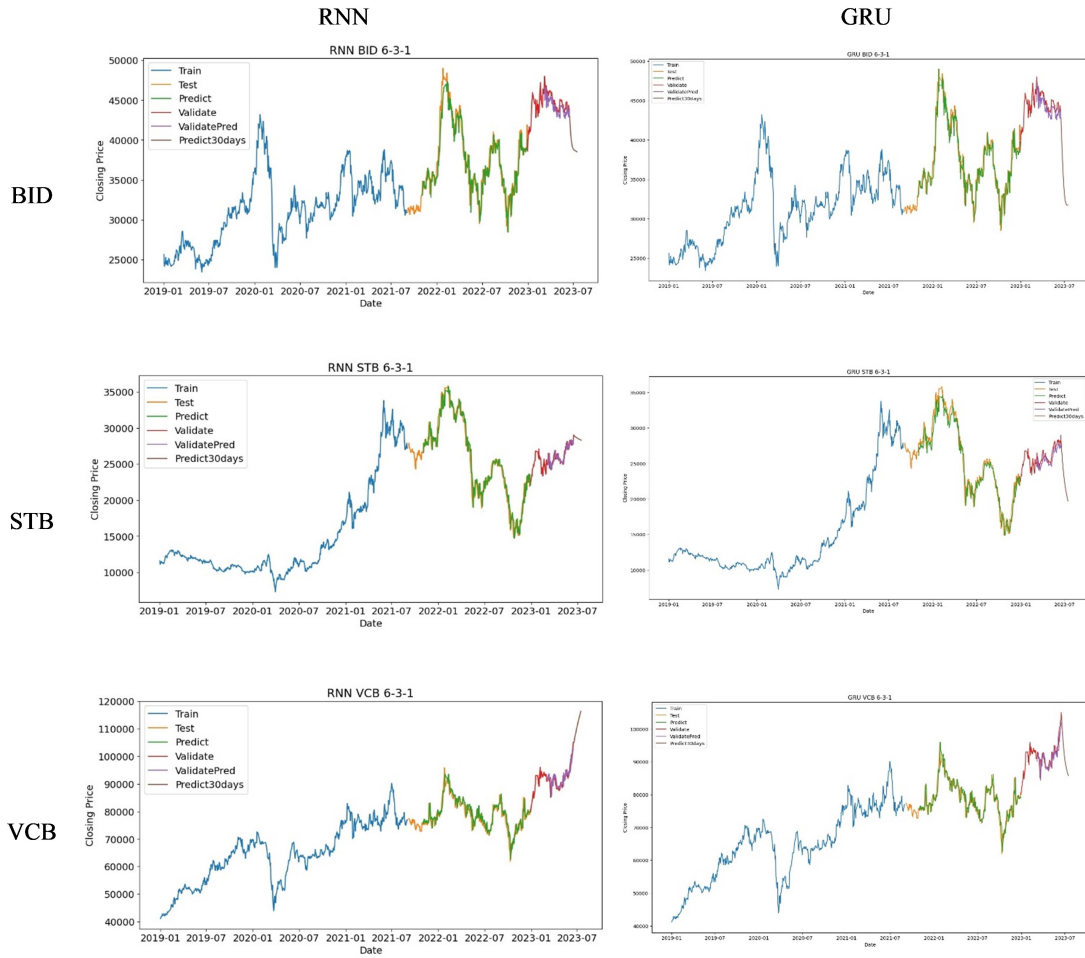
N. Dự báo 30 ngày tiếp theo

Sau khi so sánh độ chính xác của các mô hình và có bảng kết quả như Bảng II. Hai mô hình có độ chính xác tốt nhất theo độ đo RMSE được dùng để dự báo giá cổ phiếu 30 ngày tiếp theo của 3 tập dữ liệu BID, STB, VCB trên tỉ lệ (6-3-1). Kết quả dự báo được thể hiện ở Hình 18. .

Từ Hình 18, RNN và GRU cho kết quả dự báo tương đồng ở hai tập dữ liệu BID và STB khi có cùng xu hướng giảm. Còn tại VCB có sự đối nghịch, trong khi RNN dự báo giá vẫn ổn định khi đạt đỉnh thì GRU lại cho dự báo giảm mạnh sau khi giá đạt đỉnh. Ngoài ra, trên cả ba tập dữ liệu, GRU đều cho kết quả biến động mạnh hơn nhiều so với RNN.

VI. TỔNG KẾT VÀ NHẬN ĐỊNH

Từ kết quả so sánh của Bảng II, trên tập dữ liệu test, dựa vào độ đo RMSE và MAPE có thể thấy mô hình RNN có kết quả tốt nhất trong các mô hình ở trên cả hai tỉ lệ 6-3-1 và 7-2-1 ở bộ dữ liệu VCB và BID. Ở cổ phiếu STB, dù GRU ở tỉ lệ 6-3-1 tốt hơn RNN tuy nhiên tổng lại RNN vẫn cho kết quả tốt nhất ở đa số trường hợp. Qua đó có thể thấy mô hình RNN khi được xây dựng hợp lý cho kết quả rất tốt và phù hợp với bài toán dự báo giá cổ phiếu. Xem xét qua độ đo MDA, mô hình RNN không còn là mô hình tốt nhất mà ARIMAX là tốt nhất ở mọi trường hợp. Chọn mô hình tốt nhất sẽ tùy thuộc vào mục đích, khi xét trên độ đo MDA, nhóm sẽ tập trung tìm những mô hình có khả năng dự báo xu hướng nhị phân (tăng-giảm) của ngày kế tiếp tốt nhất mà không quan tâm đến mức tăng giảm thực tế. Ngược lại, xét độ đo RMSE hoặc MAPE đồng nghĩa việc sẽ quan tâm đến mức tăng giảm, mức độ chênh lệch giữa giá trị dự đoán và thực tế của cổ phiếu mà



Hình 18. Kết quả dự báo 30 ngày của RNN và GRU trên tỉ lệ (6-3-1)

hầu hết các nghiên cứu liên quan đều xem xét kết quả thực nghiệm trên khía cạnh này. Do đó, nhóm cũng sẽ xem xét trên độ đo RMSE và kết luận hai mô hình cho kết quả tốt nhất là RNN và GRU nhằm báo giá 30 ngày tiếp theo của cả ba cổ phiếu.

VII. HƯỚNG PHÁT TRIỂN TRONG TƯƠNG LAI

Trong quá trình thực nghiệm, dù gặp một vài khó khăn trong việc tìm ra các kiến trúc lớp phù hợp cho các mô hình học sâu cũng như việc đánh giá kết quả thực nghiệm số lượng lớn các mô hình trải dài từ mô hình thông kê, học máy đến học sâu nhưng nhìn chung kết quả thực nghiệm đã đạt được ngoài các mong đợi ban đầu của bài nghiên cứu. Với bộ dữ liệu cổ phiếu có nhiều biến động bất ngờ nhất là dưới ảnh hưởng của dịch Covid, việc dự báo được chính xác các giá trị tương lai có thể gặp nhiều bất lợi, tuy nhiên nếu lựa chọn các bộ siêu tham số, xây dựng kiến trúc mô hình cũng như chọn lựa mô hình hợp lý mà ở đây RNN và GRU đã cho ra kết quả rất tốt thì nhóm hy vọng có thể cải tiến và phát huy các ưu điểm hiện tại về độ đa dạng của các mô hình được sử dụng để có được kết quả khách quan và bao quát nhất. Việc có thêm độ đo MDA cũng cho bài thực nghiệm thêm một phương cách

đánh giá hiệu quả mô hình thông qua tỉ lệ khớp về xu hướng dữ liệu, khá mới mẻ so với các bài nghiên cứu khác trước đây.

Trong tương lai, nhóm sẽ tập trung cải thiện hiệu suất của các mô hình thông qua việc tinh chỉnh các bộ siêu tham số cũng như kiến trúc lớp của các mô hình học sâu như LSTM, CNN-LSTM, RNN để tận dụng được sức mạnh của các mô hình học sâu.

Nhóm cũng sẽ thực nghiệm thêm với những mô hình mới và thêm nhiều bộ dữ liệu, nhất là với những ngân hàng xảy ra nhiều biến động tỉ giá do các yếu tố bất ngờ (kinh doanh thua lỗ, nợ xấu) để tăng tính thực tế cho mô hình huấn luyện.

VIII. LỜI CẢM ƠN

Cuối cùng, nhóm nghiên cứu xin gửi lời cảm ơn đến thầy Nguyễn Đình Thuận và anh Nguyễn Minh Nhật đã luôn hỗ trợ và góp ý để nhóm có thể hoàn thành bài nghiên cứu một cách tốt nhất.

TÀI LIỆU

- [1] N. H. Tien, R. J. S. Jose, S. E. Ullah, and H. V. Thang, "The impact of world market on ho chi minh city

Tên Model		VCB		STB		BID	
		6-3-1	7-2-1	6-3-1	7-2-1	6-3-1	7-2-1
ARIMA	RMSE	5209.76	10354.49	8934.60	14764.77	7411.00	9797.20
	MAPE	4.90	12.17	33.16	64.16	15.00	25.52
	MDA	48.19	43.89	48.38	45.37	51.20	46.15
ARIMAX	RMSE	10713.74	10713.03	4207.05	4207.05	10173.45	5300.10
	MAPE	11.31	11.23	9.12	17.50	21.85	10.83
	MDA	64.76	68.78	68.67	71.04	69.58	73.76
SARIMAX	RMSE	10591.88	10550.70	2632.29	4201.35	10222.94	5302.74
	MAPE	11.18	11.05	9.62	17.48	21.94	10.83
	MDA	46.89	46.98	47.86	49.13	49.35	50.77
RNN	RMSE	1085.73	1522.23	873.23	820.93	1052.78	1051.21
	MAPE	1.44	1.53	2.81	3.15	2.13	2.21
	MDA	48.45	43.33	48.80	47.78	43.64	46.67
LSTM	RMSE	2280.42	2404.30	2891.13	1545.31	1529.73	1472.26
	MAPE	2.20	2.41	8.95	5.83	3.14	3.24
	MDA	47.14	45.71	50.00	48.57	47.14	42.86
GRU	RMSE	1420.37	1573.78	926.10	755.53	1043.37	1017.18
	MAPE	1.34	1.62	2.96	2.81	2.05	2.15
	MDA	41.43	43.89	48.11	51.11	49.83	46.11
CNN_LSTM	RMSE	3487.44	3455.59	2912.51	1855.21	3136.44	2576.79
	MAPE	3.50	3.57	9.39	6.37	5.86	5.58
	MDA	49.14	46.67	45.70	51.67	50.52	55.56
BNN	RMSE	14520.91	10761.95	6936.26	4419.06	6282.86	3841.02
	MAPE	16.36	13.22	22.92	18.27	11.71	9.29
	MDA	47.42	48.33	44.67	44.44	52.58	47.78
RF	RMSE	1851.62	1617.63	2636.88	1797.28	2018.98	1062.12
	MAPE	1.70	1.61	8.51	6.61	3.29	2.37
	MDA	45.70	46.11	50.17	49.44	48.45	43.33
LINEAR	RMSE	12346.95	12268.16	7218.50	10228.29	4843.72	4584.93
	MAPE	13.83	14.09	28.08	44.37	11.23	10.6
	MDA	45.18	45.70	46.08	47.51	50.6	54.75
XGBOOST	RMSE	2007.07	2041.93	2973.20	1996.40	2781.04	1202.68
	MAPE	1.91	2.06	10.01	8.29	4.44	2.73
	MDA	48.11	46.11	49.14	47.22	46.05	48.33

Bảng II
BẢNG SO SÁNH HIỆU QUẢ MÔ HÌNH TRÊN TẬP DỮ LIỆU TEST

- stock exchange in context of covid-19 pandemic,” *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 14, pp. 4252–4264, 2021.
- [2] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo, “Stock price prediction using the arima model,” in *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, 2014, pp. 106–112. DOI: 10.1109/UKSim.2014.67.
- [3] M. L. Challa, V. Malepati, and S. N. R. Kolusu, “S& BSE sensex and s& BSE IT return forecasting using ARIMA,” *Financial Innovation*, vol. 6, no. 1, Nov. 2020. DOI: 10.1186/s40854-020-00201-5. [Online]. Available: <https://doi.org/10.1186/s40854-020-00201-5>.
- [4] I. A. Rahmayanti, C. Andreas, and S. M. Ulyah, “Does US-china trade war affect the brent crude oil price? an ARIMAX forecasting approach,” in *INTERNATIONAL CONFERENCE ON MATHEMATICS, COMPUTATIONAL SCIENCES AND STATISTICS 2020*, AIP Publishing, 2021. DOI: 10.1063/5.0042359. [Online]. Available: <https://doi.org/10.1063/5.0042359>.
- [5] N. S. Arunraj, D. Ahrens, and M. Fernandes, “Application of sarimax model to forecast daily sales in food retail industry,” *International Journal of Operations Research and Information Systems (IJORIS)*, vol. 7, no. 2, pp. 1–21, 2016.
- [6] S. Selvin, R. Vinayakumar, E. Gopalakrishnan, V. K. Menon, and K. Soman, “Stock price prediction using lstm, rnn and cnn-sliding window model,” in *2017 international conference on advances in computing, communications and informatics (icacci)*, IEEE, 2017, pp. 1643–1647.
- [7] A. B. Omar, S. Huang, A. A. Salameh, H. Khurram, and M. Fareed, “Stock market forecasting using the random forest and deep neural network models before and during the covid-19 period,” *Frontiers in Environmental Science*, vol. 10, 2022, ISSN: 2296-665X. DOI: 10.3389/fenvs.2022.917047. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fenvs.2022.917047>.
- [8] Y. Guo, “Stock price prediction based on LSTM neural network: The effectiveness of news sentiment analysis,” in *2020 2nd International Conference on Economic Management and Model Engineering (ICEMME)*, IEEE, Nov. 2020. DOI: 10.1109/icemme51517.2020.00206. [Online]. Available: <https://doi.org/10.1109/icemme51517.2020.00206>.
- [9] J. M.-T. Wu, Z. Li, N. Herencsar, B. Vo, and J. C.-W. Lin, “A graph-based CNN-LSTM stock price prediction algorithm with leading indicators,” *Multimedia Systems*,

- Feb. 2021. DOI: 10.1007/s00530-021-00758-w. [Online]. Available: <https://doi.org/10.1007/s00530-021-00758-w>.
- [10] I. E. Livieris, E. Pintelas, and P. Pintelas, "A CNN-LSTM model for gold price time-series forecasting," *Neural Computing and Applications*, vol. 32, no. 23, pp. 17351–17360, Apr. 2020. DOI: 10.1007/s00521-020-04867-x. [Online]. Available: <https://doi.org/10.1007/s00521-020-04867-x>.
- [11] H. Zhang, "Stock price prediction using linear regression and LSTM neural network," in *2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)*, IEEE, Aug. 2022. DOI: 10.1109/mlise57402.2022.00043. [Online]. Available: <https://doi.org/10.1109/mlise57402.2022.00043>.
- [12] L. Jidong and Z. Ran, "Dynamic weighting multi factor stock selection strategy based on XGboost machine learning algorithm," in *2018 IEEE International Conference of Safety Produce Informatization (IICSPI)*, IEEE, Dec. 2018. DOI: 10.1109/iicspi.2018.8690416. [Online]. Available: <https://doi.org/10.1109/iicspi.2018.8690416>.
- [13] D. Lien Minh, A. Sadeghi-Niaraki, H. D. Huy, K. Min, and H. Moon, "Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network," *IEEE Access*, vol. 6, pp. 55392–55404, 2018. DOI: 10.1109/ACCESS.2018.2868970.
- [14] R. Chandra and Y. He, "Bayesian neural networks for stock price forecasting before and during covid-19 pandemic," *PLOS ONE*, vol. 16, no. 7, pp. 1–32, Jul. 2021. DOI: 10.1371/journal.pone.0253217. [Online]. Available: <https://doi.org/10.1371/journal.pone.0253217>.
- [15] A. C. Harvey, "ARIMA models," in *Time Series and Statistics*, Palgrave Macmillan UK, 1990, pp. 22–24. DOI: 10.1007/978-1-349-20865-4_2. [Online]. Available: https://doi.org/10.1007/978-1-349-20865-4_2.
- [16] W. K. Adu, P. Appiahene, and S. Afrifa, "VAR, ARIMA and ARIMA models for nowcasting unemployment rate in ghana using google trends," *Journal of Electrical Systems and Information Technology*, vol. 10, no. 1, Feb. 2023. DOI: 10.1186/s43067-023-00078-1. [Online]. Available: <https://doi.org/10.1186/s43067-023-00078-1>.
- [17] S. I. Vagropoulos, G. I. Chouliaras, E. G. Kardakos, C. K. Simoglou, and A. G. Bakirtzis, "Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting," in *2016 IEEE International Energy Conference (ENERGYCON)*, IEEE, Apr. 2016. DOI: 10.1109/energycon.2016.7514029. [Online]. Available: <https://doi.org/10.1109/energycon.2016.7514029>.
- [18] M. M. Fathi, A. G. Awadallah, A. M. Abdelbaki, and M. Haggag, "A new budyko framework extension using time series SARIMAX model," *Journal of Hydrology*, vol. 570, pp. 827–838, Mar. 2019. DOI: 10.1016/j.jhydrol.2019.01.037. [Online]. Available: <https://doi.org/10.1016/j.jhydrol.2019.01.037>.
- [19] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, Mar. 1990. DOI: 10.1207/s15516709cog1402_1. [Online]. Available: https://doi.org/10.1207/s15516709cog1402_1.
- [20] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently recurrent neural network (indrnn): Building a longer and deeper rnn," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.
- [21] W. Chen, Y. Zhang, C. K. Yeo, C. T. Lau, and B. S. Lee, "Stock market prediction using neural network through news on online social networks," in *2017 International Smart Cities Conference (ISC2)*, 2017, pp. 1–6. DOI: 10.1109/ISC2.2017.8090834.
- [22] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16, San Francisco, California, USA: ACM, 2016, pp. 785–794, ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. [Online]. Available: <http://doi.acm.org/10.1145/2939672.2939785>.
- [23] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, Oct. 2001. DOI: 10.1023/A:1010950718922.
- [24] D. Steinberg, "Cart: Classification and regression trees," 2009.
- [25] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. DOI: 10.1109/5.726791.
- [26] W. Lu, J. Li, Y. Li, A. Sun, and J. Wang, "A CNN-LSTM-based model to forecast stock prices," *Complexity*, vol. 2020, A. E. I.-B. Hassanien, Ed., pp. 1–10, Nov. 2020. DOI: 10.1155/2020/6622927. [Online]. Available: <https://doi.org/10.1155/2020/6622927>.
- [27] H. S. Hota, R. Handa, and A. K. Shrivastava, "Time series data prediction using sliding window based rbf neural network," 2017.