HIỆU QUẢ DỰ ĐOÁN GIÁ BITCOIN: SO SÁNH GIỮA 2 MÔ HÌNH VAR VÀ LSTM

Nguyễn Mạnh Đức[20521196]
Trường Đại học Công nghệ thông tin-ĐHQG Tp.HCM, Việt Nam
Email: 20521196@gm.uit.edu.vn

Văn Ngọc Nhật Huy[20521418]
Trường Đại học Công nghệ thông tin-ĐHQG Tp.HCM, Việt Nam
Email: 20521418@gm.uit.edu.vn

Nguyễn Đinh Việt Thắng[20521898]
Trường Đại học Công nghệ thông tin-ĐHQG Tp.HCM, Việt Nam
Email: 20521898@gm.uit.edu.vn

Tóm tắt. Bài báo này nhằm cung cấp so sánh về hiệu quả dự đoán giá Bitcoin trong tương lai bằng một số phương pháp phổ biến hiện nay. Phân tích một số yếu tố ảnh hưởng đến giá Bitcoin như tâm lý nhà đầu tư và chính sách tiền tệ. Với dữ liệu thu thập hàng ngày từ năm 2019 đến 2022, giá Bitcoin được dự đoán dựa trên 4 mô hình là mô hình vector tự hồi quy (VAR), mô hình mạng bộ nhớ ngắn hạn dài hạn (LSTM). Sử dụng 3 chỉ số là phần trăm sai số tuyệt đối trung bình (MAPE), trung bình sai số bình phương (MSE) và sai số tuyệt đối trung bình (MAE) để so sánh các mô hình.

Từ khóa: Dự đoán Bitcoin · LSTM · VAR · MAPE · MSE · MAE · Fear and greed · DXY

1 Giới thiệu nghiên cứu

Bitcoin là một đồng tiền điện tử lớn nhất trong thị trường Cryptocurrency với giá hiện tại là khoảng 20000 $ với vốn hóa tương đương 400 tỷ $. Với sự phát triển của công nghệ blockchain, thị trường cryptocurrency dần dần được mọi người khắp thế giới quan tâm và biết đến như một kênh đầu tư giúp tăng thêm lợi nhuận. Thị trường tài chính là thị trường tiềm ẩn nhiều rủi ro và đầy thách thức. Bitcoin được ví như là "vàng 4.0" và sự biến động giá tăng giảm tạo ra những tác động đến lợi nhuận cũng như sự thua lỗ của các nhà đầu tư. Dưới sự phát triển của công nghệ thông tin, các mô hình học máy, học sâu ngày càng phát triển, việc áp dụng và dự đoán sự biến động của giá Bitcoin dựa vào dữ liệu trong quá khứ đang trở thành mối quan tâm lớn đối giúp nhà đầu tư quyết định mua hay bán, điều chỉnh danh mục đầu tư hợp lý, từ đó tối ưu hóa lợi nhuận và đem lại khoản sinh lời lớn cho bản thân. Trong nghiên cứu này, chúng tôi sử dụng mô hình vector tự hồi quy (VAR), mô hình mạng bộ nhớ ngắn hạn dài hạn (LSTM) để dự đoán giá Bitcoin. Kết quả của các mô hình sẽ được so sánh với nhau qua các chỉ số phần trăm sai số tuyệt

đối trung bình (MAPE), sai số tuyệt đối trung bình (MAE) và trung bình sai số bình phương (MSE).

Trong những năm qua, các nhà nghiên cứu đã phát triển nhiều phương pháp có thể dùng để dự đoán giá Bitcoin. Một số phương pháp dự đoán biến động giá Bitcoin có thể kể đến như phương pháp phương sai có điều kiện của sai số tự thay đổi tự hồi quy (ARCH), phương pháp phương sai có điều kiện của sai số thay đổi tự hồi quy tổng quát (GARCH)...

Cùng với sự phát triển của trí tuệ nhân tạo, học máy và học sâu dần được sử dụng rộng rãi và phổ biến trong việc dự đoán sự biến động trong ngành tài chính kinh tế. Các nhà nghiên cứu đã phát triển các phương pháp dự đoán không dựa trên bất kì lý thuyết kinh tế nào. Trong nghiên cứu này, chúng tôi sử dụng mô hình bộ nhớ ngắn hạn dài hạn (LSTM) cho việc dự đoán sự biến động giá Bitcoin. Kết quả dự đoán sẽ được so sánh với mô hình VAR.

Bên cạnh việc ứng dụng các phương pháp dự đoán giá Bitcoin, không giống như các nghiên cứu trước đây thường dự đoán giá Bitcoin dựa vào các giá Bitcoin trong quá khứ, nghiên cứu này xây dựng dữ liệu để dự đoán giá Bitcoin dựa trên mức độ quan tâm và tâm lý của nhà đầu tư, tiền tệ liên quan đến các thị trường tài chính khác liên quan đến Bitcoin. Chúng tôi sẽ thảo luận về tác động của các yếu tố này đến giá Bitcoin trong các phần sau.

2 Các mô hình dự báo giá Bitcoin và đánh giá mô hình

2.1 Mô hình VAR (Vector Auto-Regressive Model)

Mô hình VAR (Vector Auto-Regressive model) là một trong những mô hình thành công nhất, linh hoạt và dễ sử dụng nhất để phân tích chuỗi thời gian đa biến. Mô hình VAR đã được chứng minh là có tác động hữu ích để dự đoán hành vi của chuỗi thời gian kinh tế và tài chính.

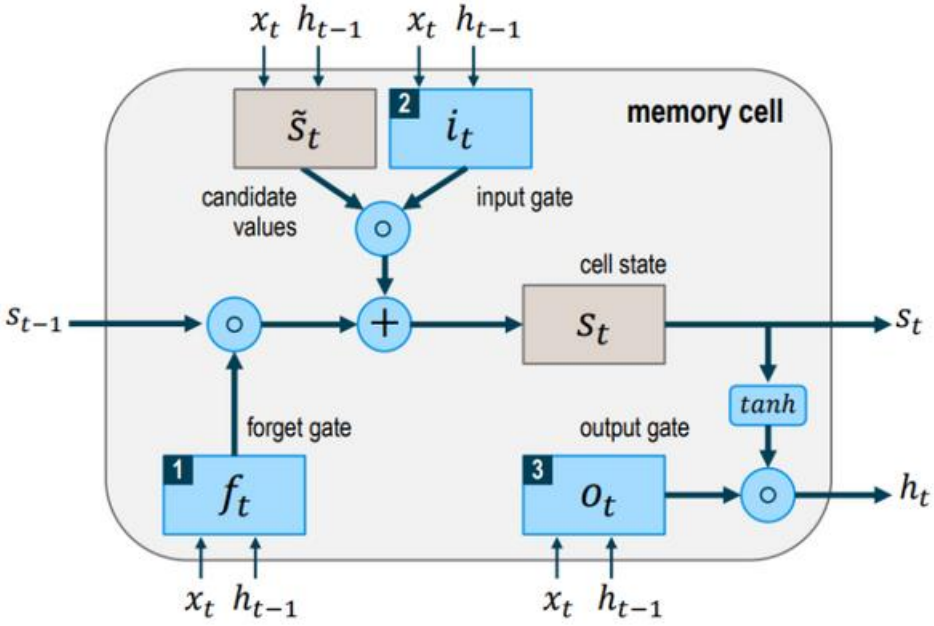
Mô hình vector tự hồi quy với p độ trễ (VAR(p)) có dạng như sau:

$$Yt = c + A1Yt-1 + A2Yt-2+ \ldots + ApYt-p + \varepsilon t$$

Yêu cầu quan trọng nhất để ước lượng mô hình VAR là các chuỗi thời gian phải có tính dừng. Một quá trình ngẫu nhiên được gọi là có tính dừng nếu giá trị trung bình và phươn sai của nó không thay đổi theo thời gian, và giá trị hiệp phương sai giữa hai đoạn chỉ phụ thuộc vào khoảng cách và thời gian trễ giữa hai giai đoạn. Hai khoảng này không phụ thuộc vào phép đo hiệp phương sai thực tế. Trong nghiên cứu này việc kiểm định tính dừng của các chuỗi thời gian bằng cách thực hiện kiểm định Augmented DickeyFuller (ADF), việc tìm ra độ trễ tối ưu cho mô hình VAR được thực hiện dựa trên các tiêu chí như thống kê LR, tiêu chí thông tin Akaike (AIC).

2.2 Mô hình LSTM

LSTM là một phiên bản mở rộng của mạng RNN, được đề xuất vào năm 1997 bởi Sepp Hochreiter và Jurgen Schmidhuber. LSTM được thiết kế để giải quyết các bài toán phụ thuộc xa trong mạng RNN do ảnh hưởng bởi vấn đề gradient biến mất [1].



Hình 1. Sơ đồ biểu diễn kiến trúc bên trong một tế bào LSTM

Ý tưởng của LSTM là bổ sung thêm trạng thái bên trong tế bào (cell internal state) s_t và ba cổng sàng lọc các thông tin đầu vào và đầu ra cho tế bào gồm forget gate f_t, input gate i_t và output gate o_t. Tại mỗi bước thời gian t, các cổng đều lần lượt nhận giá trị đầu vào x_t (đại diện cho một phần tử trong chuỗi đầu vào) và giá trị h_{t-1} có được từ đầu ra của memory cell từ bước thời gian trước đó t – 1. Các cổng đều đóng vai trò có nhiệm vụ sàng lọc thông tin với mỗi mục đích khác nhau:

Forget gate: Có nhiệm vụ loại bỏ những thông tin không cần thiết nhận đượckhỏi cell internal state.

Input gate: Có nhiệm vụ chọn lọc những thông tin cần thiết nào được thêm vào cell internal state

Output gate: Có nhiệm vụ xác định những thông tin nào từ cell internal state được sử dụng như đầu ra

Trong quá trình lan truyền xuôi (forward pass), cell internal state s_t và giá trị đầu ra h_t được tính như sau:

$$f_t = \sigma(W_{f,x}x_t + W_{f,h}h_{t-1} + b_f)$$
$$\tilde{s}_t = \tanh(W_{\tilde{s},x}x_t + W_{\tilde{s},h}h_{t-1} + b_{\tilde{s}})$$
$$i_t = tanh(W_{i,x}x_t + W_{i,h}h_{t-1} + b_i)$$
$$s_t = f_t \circ s_{t-1} + i_t \circ \tilde{s}_t$$
$$o_t = \sigma(W_{o,x}x_t + W_{o,h}h_{t-1} + b_o)$$
$$h_t = o_t \circ \tanh(s_t)$$

Trong đó:

x_t: là $vector$ đầu vào tại mỗi bước thời gian t

$W_{f,x}, W_{f,h}, W_{\tilde{s},x}, W_{\tilde{s},h}, W_{i,x}, W_{i,h}, W_{o,x}, W_{o,h}$: là các ma trận trọng số trong mỗi

tế bào LSTM
b_f , $b_{\tilde{s}}$, b_i , b_o : *là các vector bias*
f_t , i_t , o_t: *lần lượt chứa các giá trị kích hoạt lần lượt cho các cổng forget gate, input gate, output gate*
$s_t, \tilde{s}$: *lần lượt là các vector đại diện cho cell internal state và candidate value*
h_t: *là giá trị đầu ra tế bào LSTM*

2.3 Đánh giá mô hình

Trong thống kê, Mean Squared Error (MSE - sai số bình phương trung bình) của công cụ ước tính (của thủ tục ước tính số lượng không quan sát được) đo trung bình bình phương của các lỗi – nghĩa là chênh lệch bình phương trung bình giữa các giá trị dự đoán và giá trị gốc[13]. MSE là một hàm rủi ro, tương ứng với giá trị dự kiến của mất lỗi bình phương. Công thức tính MSE:

$$MSE = \frac{\sum_{t=1}^{n} \varepsilon_t^2}{n} = \frac{\sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2}{n}$$

Mean Absolute Error (MAE - sai số tuyệt đối trung bình) là một phương pháp đo lường sự khác biệt giữa hai biến liên tục. Chúng ta có độ đo MAE được tính theo công thức sau:

$$MAE = \frac{\sum_{t=1}^{n} |\varepsilon_t|}{n} = \frac{\sum_{t=1}^{n} |Y_t - \hat{Y}_t|}{n}$$

Mean Absolute Percent Error (MAPE - sai số tỷ lệ phần trăm tuyệt đối trung bình) là trung bình của các lỗi tuyệt đối chia cho các giá trị quan sát thực tế. Chúng ta có độ đo MAPE được tính theo công thức sau:

$$MAPE = \frac{\sum_{t=1}^{n} \frac{|\varepsilon_t|}{Y_t}}{n} = \frac{\sum_{t=1}^{n} \frac{|Y_t - \hat{Y}_t|}{Y_t}}{n}$$

3 Dữ liệu nghiên cứu

Nghiên cứu này dự báo giá Bitcoin dựa trên các yếu tố mô hình ảnh hưởng đến giá Bitcoin. Như đã nhấn mạnh ở phần giới thiệu, các yếu tố thúc đẩy được tiếp cận trên 2 khía cạnh là chính sách tiền tệ và tình trạng cảm xúc.

Chính sách tiền tệ được coi là tác động từ phía cầu đến giá Bitcoin. Nó có ảnh hưởng đáng kể đến thị trường hàng hóa và các thị trường đầu tư tài chính. Chỉ số DXY là chỉ số theo dõi hoạt động của USD và đo lường giá trị của USD so với 6 loại tiền tệ pháp định khác. USD cũng là quy ước chính để đổi qua các đồng tiền điện tử như USDT, BUSD để niêm yết giá Bitcoin trên các sàn giao dịch điện tử. Ngoài ra, chỉ số trạng thái tâm lý của thị trường crypto – Fear & Greed Index giúp ghi nhận trạng thái tâm lý sợ hãi và tham lam của các nhà đầu tư Bitcoin nói riêng và cả thị trường crypto nói chung. Fear & Greed Index được xác định bằng cách phân tích

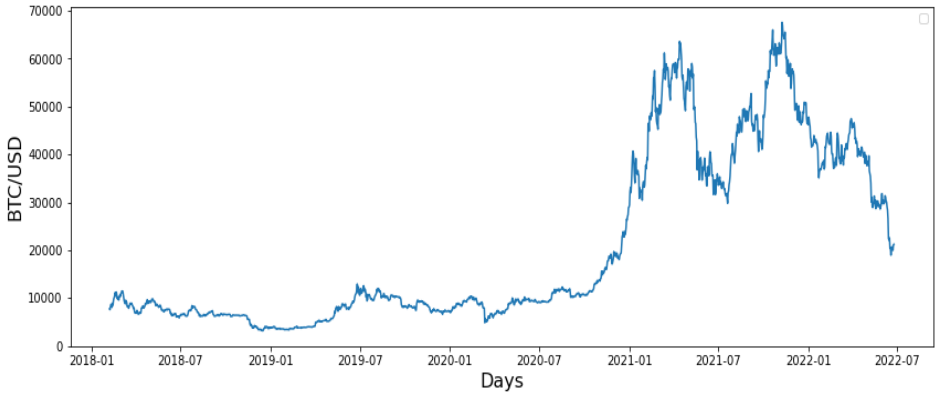
một bộ dữ liệu về Bitcoin được tổng hợp từ nhiều gồm, 6 yếu tố quyết định đến chỉ số là: valotility (sự biến động giá), market Momentum/Volume (chỉ báo động lượng thị trường / Khối lượng giao dịch), Social Media (Truyền thông xã hội), Surveys (khảo sát), Dominance (tỷ trọng vốn hóa của đồng coin so với toàn bộ thị trường) và Google Trends (xu hướng tìm kiếm google). Do đó, bộ dữ liệu được sử dụng để dự đoán như sau:

Bảng 1. Các chỉ số ảnh hưởng tới giá Bitcoin.

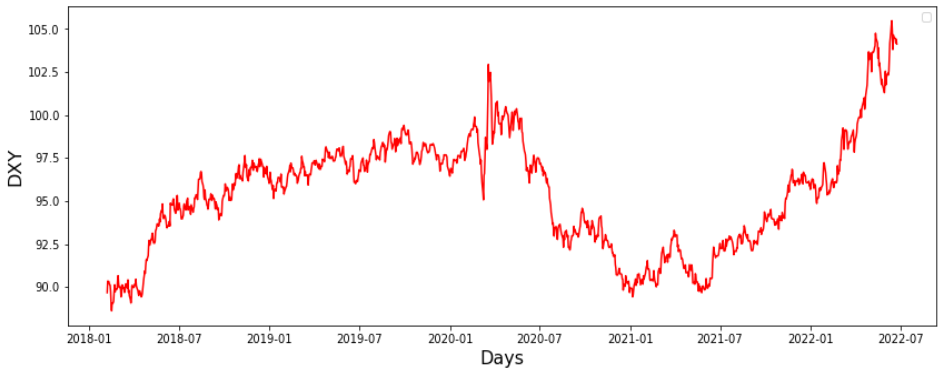
Chỉ số	Mô tả	Nguồn
BTC/USD	Giá cột "Close" của Bitcoin theo từng ngày	Trading view Track All Market
DXY	Chỉ số theo dõi hoạt động của USD và đo lường giá trị của USD so với 6 loại tiền tệ pháp định khác.	Trading view Track All Market
Fear and Greed Index	Là một chỉ số đo lường cảm tính thị trường.	Alternative

Chuỗi thời gian được thu thập hàng ngày từ ngày 06/02/2018 đến ngày 26/06/2022. Như vậy, chuỗi thời gian gồm 1600 quan sát. Nhóm sẽ xử lý dữ liệu để tạo ứng với mỗi dữ liệu đầu vào là chỉ số của 3 chuỗi thời gian của 30 ngày liên tiếp nhau, sẽ dự đoán ra giá Bitcoin của 3 ngày, 7 ngày, 14 ngày tiếp theo từ ngày cuối cùng của dữ liệu đầu vào. Ứng với mỗi khoảng dự đoán khác nhau, nhóm sẽ so sánh mức hiệu quả của mô hình. Và nhóm cũng tạo dữ liệu dữ liệu đầu vào chỉ gồm chỉ số giá Bitcoin của 30 ngày liên tiếp nhau để so sánh độ hiệu quả dự đoán với mô hình sử dụng cả 3 chuỗi thời gian.

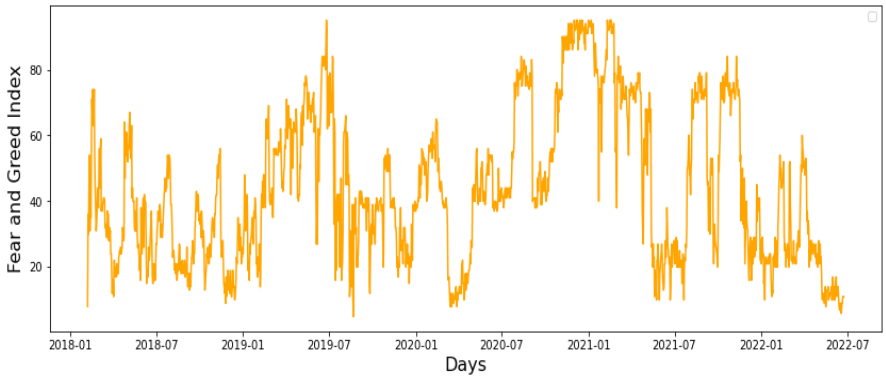
Dưới đây là hình ảnh trực quan 3 chuỗi thời gian mà nhóm sử dụng:



Hình 2. Chỉ số BTC/USD



Hình 3. Chỉ số DXY



Hình 4. Chỉ số tâm lí

4 Kết quả nghiên cứu

4.1 Mô hình VAR (Vector Auto-Regressive Model)

4.1.1 Kiểm định tính dừng

Kiểm định Augmented Dickey-Fuller (ADF) trên bộ dữ liệu thu được kết quả như sau:

	BTC/USD	**DXY**	**Fear and Greed Index**
ADF Statistic	-1.276189	-1,257011	-4,584949
n_lags	0.639994	0.648603	0.000137
p-value	0.639994	0.648603	0.000137

Critial Values			
1%	-3.434508	-3.434466	-3.434451
5%	-2.863376	-2.863358	-2.863351
10%	-2.567747	-2.567738	-2.567734

Kết quả kiểm định tính dừng cho thấy chuỗi thời gian BTC/USD và DXY không có tính dừng ở chuỗi gốc (giá trị p-value > 0.05) và chuỗi thời gian Fear and Greed Index cho thấy chuỗi có tính dừng (giá trị p-value < 0.05). Tuy nhiên, giá trị p-value của kiểm định ADF cho thấy BTC/USD và DXY cho thấy chuỗi dừng ở sai phân bậc 1. Nhóm sẽ để chuỗi ở dạng sai phân bậc 1 để đưa vào mô hình VAR, sau khi dự đoán sẽ đưa lại về giá trị gốc.

4.1.2 Kiểm định tác động giữa các chuỗi thời gian

Nhóm sử dụng Granger Causality test để kiểm tra chuỗi thời gian có hữu ích cho dự đoán chuỗi thời gian khác không, cụ thể ở đây, nhóm kiểm tra chuỗi thời gian DXY và Fear and Greed Index liệu có hữu ích đối với việc dự đoán chuỗi thời gian BTC/USD hay không.

H_0: Chuỗi thời gian x không là chuỗi thời gian nguyên nhân Granger y

H_1: Chuỗi thời gian x là chuỗi thời gian nguyên nhân Granger y.

"Nguyên nhân Granger" nghĩa là biết giá trị của chuỗi thời gian x ở một độ trễ nhất định rất hữu ích để dự đoán giá trị của chuỗi thời gian y tại một khoảng thời gian sau đó [3] Kết quả thu được trên 3 chuỗi thời gian là:

```
BTC/USD causes DXY?
------------------

Granger Causality
number of lags (no zero) 1
ssr based F test:         F=5.4931  , p=0.0192  , df_denom=1596, df_num=1
ssr based chi2 test:   chi2=5.5034  , p=0.0190  , df=1
likelihood ratio test: chi2=5.4939  , p=0.0191  , df=1
parameter F test:         F=5.4931  , p=0.0192  , df_denom=1596, df_num=1

DXY causes BTC/USD?
------------------

Granger Causality
number of lags (no zero) 1
ssr based F test:         F=0.8783  , p=0.3488  , df_denom=1596, df_num=1
ssr based chi2 test:   chi2=0.8799  , p=0.3482  , df=1
likelihood ratio test: chi2=0.8797  , p=0.3483  , df=1
parameter F test:         F=0.8783  , p=0.3488  , df_denom=1596, df_num=1

BTC/USD causes Fear and Greed Index?
------------------

Granger Causality
number of lags (no zero) 1
ssr based F test:         F=5.3762  , p=0.0205  , df_denom=1596, df_num=1
ssr based chi2 test:   chi2=5.3864  , p=0.0203  , df=1
likelihood ratio test: chi2=5.3773  , p=0.0204  , df=1
parameter F test:         F=5.3762  , p=0.0205  , df_denom=1596, df_num=1

Fear and Greed Index causes BTC/USD?
------------------

Granger Causality
number of lags (no zero) 1
ssr based F test:         F=1.4219  , p=0.2333  , df_denom=1596, df_num=1
ssr based chi2 test:   chi2=1.4246  , p=0.2326  , df=1
likelihood ratio test: chi2=1.4240  , p=0.2327  , df=1
parameter F test:         F=1.4219  , p=0.2333  , df_denom=1596, df_num=1
```

Hình 5. Kết quả thu được của 3 chuỗi thời gian với Granger Causality test

Kết quả cho thấy DXY và Fear and Greed Index đối với BTC/USD cho giá trị p-value lần lượt là 0.0192 và 0.0205 đều nhỏ hơn 0.05, ta có thể bác bỏ giả thuyết H_0 của thử nghiệm và kết luận rằng 2 chỉ số DXY và Fear and Greed Index rất hữu ích để dự đoán số lượng gà trong tương lai. Tuy nhiên, vẫn có trường hợp nhân quả ngược lại xảy ra, đó là chỉ số DXY và Fear and Greed Index thay đổi do chỉ số BTC/USD. Để loại trừ khả năng này, nhóm cũng đã kiểm tra thực hiện ngược lại, kết quả BTC/USD đối với DXY và Fear and Greed Index cho giá trị p-value lần lượt là 0.3488 và 0.2333, vì giá trị p-value không nhỏ hơn 0.05 nên ta không thể bỏ giả thuyết H_0, đó là chỉ số BTC/USD không hữu ích để dự đoán DXY và Fear and Greed Index. Do đó, nhóm kết luận rằng biết DXY và Fear and Greed Index rất hữu ích để dự đoán BTC/USD trong tương lai.

4.2 Kết quả dự đoán bằng mô hình VAR và LSTM

Bảng 2. Kết quả dự đoán của mô hình VAR và LSTM.

VAR - 1 - 3days MSE : 429323.17084455636 MAE : 623.8008480806699 MAPE : 2.9807104243422518	**LSTM - 1 - 3days** MSE : 42511900.95569266 MAE : 5504.633682725696 MAPE : 13.683648852410036
VAR - 1 - 7days MSE : 686063.7706866184 MAE : 642.4538979112606 MAPE : 3.163201888253265	**LSTM - 1 - 7days** MSE : 84307364.66141309 MAE : 7488.822204044582 MAPE : 18.078746450099008
VAR - 1 - 14days MSE : 57968967.08841268 MAE : 7194.507888530047 MAPE : 34.43160752345044	**LSTM - 1 - 14days** MSE : 73058859.38150683 MAE : 6402.595334025676 MAPE : 15.025083067484138
VAR - 3 - 3days MSE : 257839.336129179 MAE : 501.00026205082395 MAPE : 2.42407174270	**LSTM - 3 - 3days** MSE : 17820452.857461475 MAE : 3404.084194155092 MAPE : 8.381001170319458
VAR - 3 - 7days MSE : 671454.7657601906 MAE : 556.6350659555309 MAPE : 2.8120392582737197	**LSTM - 3 - 7days** MSE : 36440689.52672639 MAE : 5166.873146181781 MAPE : 12.758948519053606
VAR - 3 - 14days MSE : 63290151.48775984 MAE : 7496.6339129578255 MAPE : 35.907138652925134	**LSTM - 3 - 14days** MSE : 85108256.2328191 MAE : 7706.636038530656 MAPE : 18.573857576124254

Ký hiệu X – a – b days: X là mô hình dùng để dự đoán, a là số lượng chuỗi thời gian sử dụng làm dữ liệu đầu vào, b days là số lượng b ngày dự đoán ở đầu ra. Dựa vào kết quả các chỉ số MSE, MAE, MAPE có thể thấy, khi dự đoán giá BTC/USD trong khung thời gian 3 ngày và 7 ngày, mô hình VAR hoạt động tốt hơn mô hình LSTM, 2 chuỗi thời gian DXY và Fear and Greed Index giúp cho mô hình hoạt động tốt hơn đối với cả LSTM và VAR (các chỉ số thu được thấp hơn so với chỉ dùng 1 chuỗi thời gian BTC/USD làm dữ liệu đầu vào). Mô hình VAR và mô hình LSTM đều hoạt động tốt nhất trên khung dự đoán 3 ngày tiếp theo và hoạt động kém nhất trên khung dự đoán 14 ngày tiếp theo. Có thể thấy khi dữ liệu đầu ra dự đoán càng nhiều ngày (càng xa bộ dữ liệu) thì mô hình hoạt động kém dần. Kết quả cho thấy trùng khớp với Granger Causality test khi kiểm tra sự tác động giữa các chuỗi thời gian. Có thể

thấy 3 dữ liệu chuỗi thời gian đều không có tính chu kỳ và thời vụ, dữ liệu chứa một số lượng lớn yếu tố ngẫu nhiên và bộ dữ liệu còn khá ít, vậy nên mô hình hoạt động khá kém hiệu quả trên bộ dữ liệu này.

5 Kết luận

Đóng góp của báo cáo này là cung cấp mô phỏng chi tiết về yếu tố tác động đến giá Bitcoin xuất phát từ khía cạnh cảm xúc của nhà đầu tư Crypto và chính sách tiền tệ, cụ thể là chỉ số DXY. Kết quả về sự tác động của chỉ số DXY và Fear and Greed

Index hữu ích để dự đoán giá Bitcoin đã được kiểm định bằng phương pháp Granger Causality test, và được khẳng định lần nữa bởi các chỉ số MSE, MAE, MAPE khi so sánh sự hiệu quả của mô hình khi có và không có 2 chỉ số trên trên 2 mô hình VAR và LSTM trên 3 khung thời gian dự đoán khác nhau.

Tài liệu tham khảo

1. Nguyen Truong Long, Giải thích chi tiết về mạng Long Short-term Memory (LSTM). https://nguyentruonglong.net/giai-thich-chi-tiet-ve-mang-long-shortterm-memory-lstm.html, Written on October 18, 2018.
2. Khoa học dữ liệu – Khanh's blogs https://phamdinhkhanh.github.io/2019/12/12/ARIMAmodel.html#:~:text=ARIMA%20model%20l%C3%A0%20vi%E1%BA%BFt%20t%E1%BA%AFt,regression%3A%20K%C3%AD%20hi%E1%BB%87u%20l%C3%A0%20AR.
3. Học máy thống kê – Trường đại học Công nghệ thông tin ĐHQG Tp. Hồ Chí Minh https://courses.uit.edu.vn/pluginfile.php/329459/mod_resource/content/1/CS114_K23_Buoi7-Ridge-lasso2.pdf
4. Nguồn dữ liệu: https://alternative.me/crypto/fear-and-greed-index/ https://www.tradingview.com/
5. https://www.machinelearningplus.com/time-series/vector-autoregression-examples-python/
6. https://www.statology.org/granger-causality-test-in-python/