ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC MÁY TÍNH

□---------🗐---------□

MÔN HỌC: MÁY HỌC

BÁO CÁO ĐỒ ÁN: DỰ ĐOÁN GIÁ CỦA BITCOIN

Giảng viên hướng dẫn	:	Lê Đình Duy
		Phạm Nguyễn Trường An
Lớp	:	CS114.O11.KHCL
Nhóm sinh viên thực hiện	:	Phạm Trần Xuân Khôi – 21521014
		Lê Văn Quân - 21522491

Thành phố Hồ Chí Minh, Tháng 01 năm 2024

TÓM TẮT ĐỒ ÁN

Đồ án này tập trung vào việc dự đoán giá Bitcoin sử dụng mô hình Long Short-term Memory (LSTM). Các thành viên trong nhóm đã huấn luyện mô hình này dựa trên dữ liệu giá Bitcoin từ 18/7/2010. Mặc dù giá Bitcoin có thể biến động do nhiều yếu tố và khó dự đoán, nhưng mô hình đã cho kết quả khả quan trên tập thử nghiệm và trong thực tế. Kết quả này có thể cung cấp một nguồn tham khảo hữu ích giúp các nhà đầu tư và nhà giao dịch trong việc dự đoán giá Bitcoin. Đồ án này mở ra hướng tiếp cận mới trong việc sử dụng công nghệ AI để phân tích và dự đoán thị trường tiền mã hóa.

- Repository github: Khoi19112003/CS114.O11.KHCL-21521014 (github.com)

NHẬN XÉT CỦA GIẢNG VIÊN

TP.HCM, ngày … tháng … năm 2024

Ký tên

Nội dung

Phần 0: UPDATE SAU KHI VẤN ĐÁP

I. Tóm tắt:

Các phần nhóm làm thêm sau khi vấn đáp bao gồm:

- Chia dataset thành tập train và test (80/20) và chỉ sử dụng tập train cho huấn luyện mô hình. Ban đầu, nhóm sử dụng toàn bộ data cho việc huấn luyện, dẫn đến việc đánh giá trở nên không khách quan.
- Đầu vào cho mô hình ban đầu không bao gồm Giá đóng, bây giờ đầu vào cho mô hình có cả Giá đóng.

II. Cụ thể:

Data: Dataset được chia thành 2 phần, phần dùng để train bao gồm 80% thời gian đầu tiên có trong data, 20% còn lại dùng để test.

Đầu vào cho mô hình bây giờ bao gồm 8 thông tin: Giá đóng, Giá mở, Giá cao nhất, Giá thấp nhất, Khối lượng giao dịch trong ngày, Biến động giá trong ngày, Lãi suất, Giá vàng của 30 ngày trước. Đầu ra cho mô hình là Giá đóng của ngày tiếp theo.

Data sẽ không được chuẩn hoá.

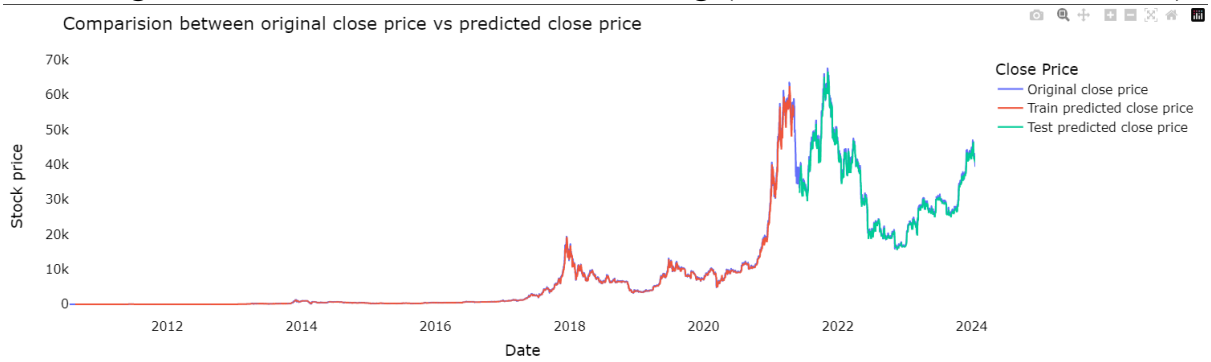
Mô hình sử dụng có thay đổi các siêu tham số so với ban đầu:

- Epochs: 100
- Batch size: 4
- Validation_split: 0.2

Kết quả:

Số đơn vị LSTM	50	100
MAPE	2.61%	5.57%

So sánh giá trị dự đoán và thực tế của Giá đóng (mô hình với 50 đơn vị LSTM):



*Giải thích lý do tại sao mô hình của nhóm không thể dự đoán nhiều ngày trong tương lai: giả sử hiện tại là ngày 0, để dự đoán Giá đóng ngày tiếp theo (ngày 1), ta cần dữ liệu của n ngày trước đó (trong đồ án này, n = 30 ngày), mô hình của nhóm có 8 features đầu vào và chỉ có 1 output (Giá đóng), do đó để dự đoán ngày kế tiếp (ngày 2) nữa, mô hình cần có thông tin về Giá mở, Giá cao nhất, Giá thấp nhất, Khối lượng giao dịch trong ngày, Biến động giá trong ngày, Lãi suất, Giá vàng của ngày vừa dự đoán (ngày 1). Mà ở hiện tại (ngày 0) không thể có các thông tin trên ở ngày 1, do đó việc dự đoán nhiều ngày là bất khả thi. Các nghiên cứu khác có thể dự đoán nhiều ngày là do họ sử dụng 1 feature đầu vào của 30 ngày trước để dự đoán ngày tiếp theo. Sau khi dự đoán, họ thêm dự đoán của ngày đó vào làm input cho việc dự đoán ngày kế tiếp.

Phần I: TỔNG QUAN

I. Tổng quan về đề tài

- Mục đích:
    - Có thể đưa ra dự đoán chính xác về giá Bitcoin của ngày tiếp theo, hỗ trợ nhà đầu tư và nhà giao dịch đưa ra quyết định mua hoặc bán.
- Ứng dụng:
    - Đầu tư và Thương mại: Bitcoin và các loại tiền điện tử khác đã trở thành một phần quan trọng của thị trường tài chính toàn cầu. Việc dự đoán giá Bitcoin có thể giúp nhà đầu tư và nhà giao dịch đưa ra quyết định mua hoặc bán.
    - Hiểu rõ hơn về thị trường: Giúp chúng ta hiểu rõ hơn về cách thức hoạt động của thị trường tiền điện tử, cũng như các yếu tố ảnh hưởng đến giá cả.
    - Học hỏi và nghiên cứu: Cung cấp một cơ hội tuyệt vời để học hỏi và thử nghiệm các kỹ thuật học máy và phân tích dữ liệu.

II. Giới thiệu về bài toán

- Input: dữ liệu về Bitcoin ngày hôm qua.
- Output: Giá Bitcoin hôm nay.

Phần II: XÂY DỰNG BỘ DỮ LIỆU

I. Thu thập dữ liệu

- Thu thập dữ liệu về giá Bitcoin, bao gồm các thông tin: Giá đóng, Giá mở, Giá cao nhất, Giá thấp nhất, Khối lượng giao dịch trong ngày, Biến động giá trong ngày với khung thời gian hàng ngày từ 18/07/2010-22/01/2024.
- Dữ liệu được lấy từ Investing.

Date	Price	Open	High	Low	Vol.	Change %
22/01/2024	39,556.4	41,581.7	41,684.9	39,468.4	85.05K	-4.87%
21/01/2024	41,583.2	41,695.4	41,878.0	41,504.5	16.11K	-0.27%
20/01/2024	41,695.4	41,647.6	41,858.0	41,449.5	22.27K	0.11%
19/01/2024	41,648.0	41,293.8	42,164.6	40,305.4	72.64K	0.86%
18/01/2024	41,292.7	42,763.5	42,908.0	40,682.6	70.35K	-3.45%

- Dựa theo nghiên cứu của Shinta Amalina Hazrati Havidz, Viendya Ervina Karman và Indra Yudha Mambea (Link), đề xuất thêm 2 yếu tố ảnh hưởng giá Bitcoin là Lãi suất và Giá vàng.
- Giá vàng lấy từ Investing (hàng ngày), Lãi xuất lấy từ Federal Reserve Bank (hàng tháng).

Date	Gold
22/1/2024	2026,75
21/1/2024	2029,624
20/1/2024	2029,624
19/1/2024	2029,624
18/1/2024	2023,104
17/1/2024	2006,33
16/1/2024	2028,445
15/1/2024	2054,772
12/1/2024	2049,174

Date	FEDFUNDS
18/7/2010	0,18
1/8/2010	0,19
1/9/2010	0,19
1/10/2010	0,19
1/11/2010	0,19
1/12/2010	0,18
1/1/2011	0,17
1/2/2011	0,16
1/3/2011	0,14
1/4/2011	0,10
1/5/2011	0,09
1/6/2011	0,09
1/7/2011	0,07

II. Xử lý dữ liệu

- Chuyển các giá trị Giá đóng, Giá mở, Giá cao nhất, Giá thấp nhất trong ngày thành dạng số thực:

42,835.9	46,348.1	46,503.2	41,857.9
42,851.3	42,836.7	43,248.6	42,443.3
41,746.1	42,851.3	43,069.4	41,739.6
42,510.7	41,747.6	43,348.9	41,719.2
43,145.5	42,515.2	43,563.7	42,093.1
42,768.7	43,139.1	43,192.3	42,211.8
41,292.7	42,763.5	42,908.0	40,682.6
41,648.0	41,293.8	42,164.6	40,305.4
41,695.4	41,647.6	41,858.0	41,449.5
41,583.2	41,695.4	41,878.0	41,504.5
39,556.4	41,581.7	41,684.9	39,468.4

42835,9	46348,1	46503,2	41857,9
42851,3	42836,7	43248,6	42443,3
41746,1	42851,3	43069,4	41739,6
42510,7	41747,6	43348,9	41719,2
43145,5	42515,2	43563,7	42093,1
42768,7	43139,1	43192,3	42211,8
41292,7	42763,5	42908	40682,6
41648	41293,8	42164,6	40305,4
41695,4	41647,6	41858	41449,5
41583,2	41695,4	41878	41504,5
39556,4	41581,7	41684,9	39468,4

- Xử lý định dạng của Khối lượng giao dịch trong ngày:

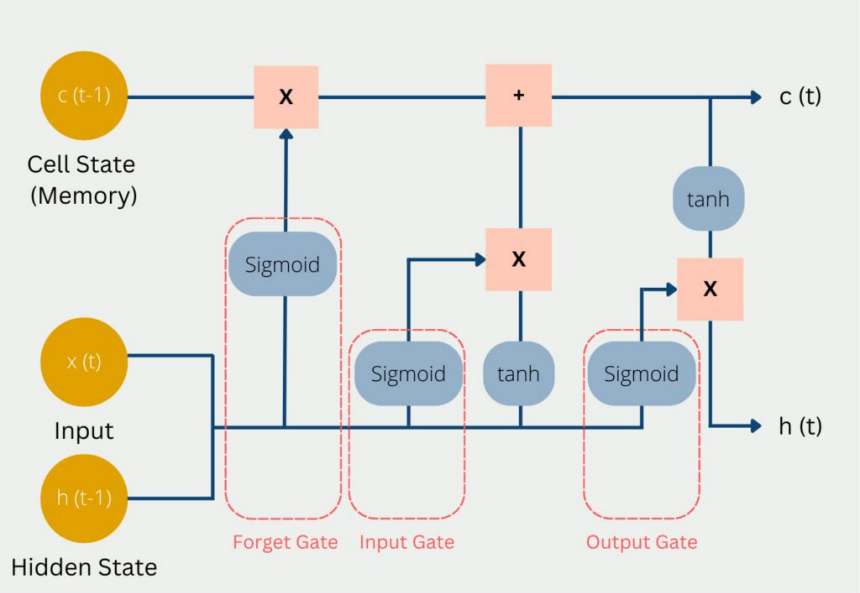
136.92K	136,92
48.18K	48,18
37.14K	37,14
52.08K	52,08
63.93K	63,93
50.44K	50,44
70.35K	70,35
72.64K	72,64
22.27K	22,27
16.11K	16,11
85.05K	85,05

- Giá vàng chỉ có thông tin về các ngày trong tuần.
    - Đặt Giá vàng ở các ngày nghỉ bằng giá vàng của ngày Thứ 6 trước đó.
- Tương tự, Lãi suất cũng chỉ có thông tin ngày đầu của tháng.
    - Đặt Lãi suất của các ngày trong tháng bằng với ngày đầu của tháng.
- Sau đó định dạng chúng lại thành số thực.
- Sử dụng các thông tin Giá mở, Giá cao nhất, Giá thấp nhất, Khối lượng giao dịch trong ngày, Biến động giá trong ngày, Lãi suất, Giá vàng làm đầu vào cho mô hình.
- Sử dụng Giá đóng là đầu ra cho mô hình.
- Phương pháp chia dữ liệu: Sử dụng TimeSeriesSplit với n_split = 30.

Phần III: CHỌN MÔ HÌNH VÀ HUẤN LUYỆN

I. Long Short-term Memory

- LSTM là một mạng cải tiến của RNN, được đề xuất năm 1997 bởi Sepp Hochreiter và Jurgen Schmidhuber
- Được thiết kế để giải quyết vấn đề phụ thuộc dài(long-term dependencies) trong mạng RNN do bị ảnh hưởng bởi vấn đề vanishing gradient.
- Kiến trúc LSTM:



  - Gồm 4 thành phần chính: cell, input gate, forget gate, output gate.
  - LSTM giải quyết vấn đề "vanishing gradient" bằng cách sử dụng cơ chế "gating". Cơ chế này cho phép mô hình kiểm soát thông tin nào được giữ lại và thông tin nào được loại bỏ, giúp mô hình học được các phụ thuộc dài hạn.

➢ Việc lựa chọn LSTM là phù hợp vì khả năng xử lý dữ liệu chuỗi thời gian, học các mối quan hệ phụ thuộc dài và xử lý dữ liệu không tuyến tính.

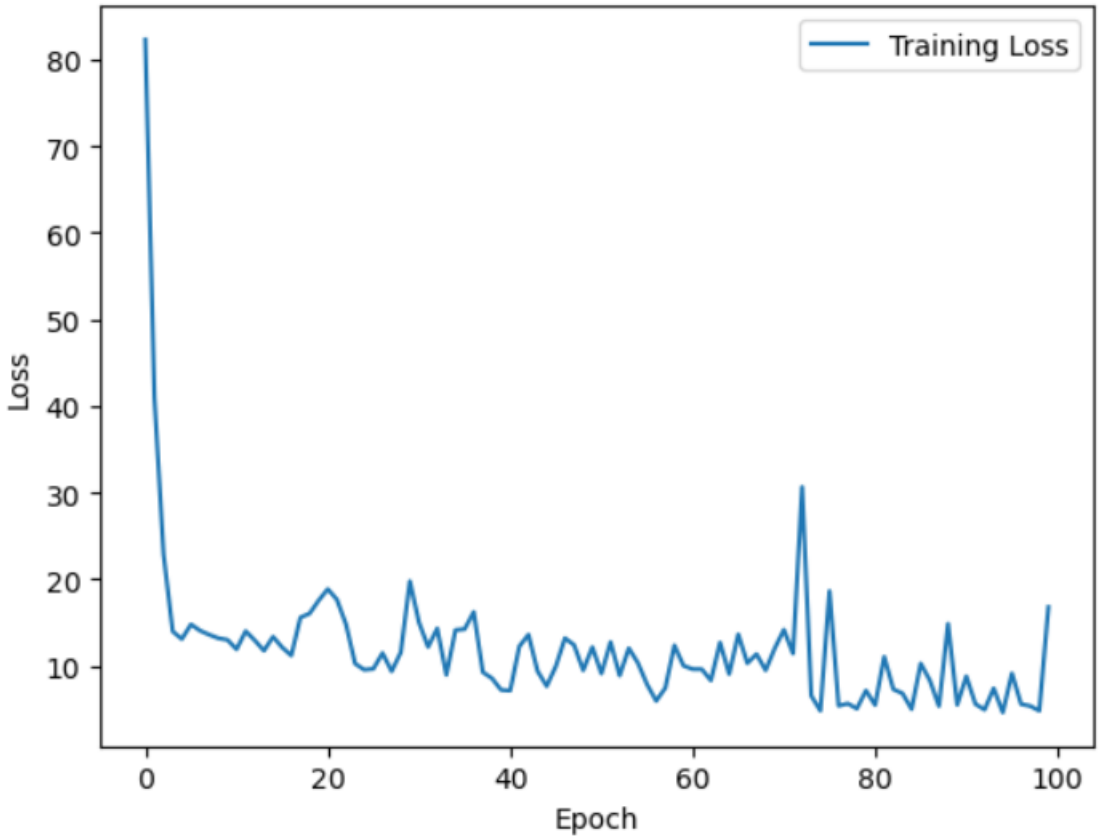
Link: [1402.1128] Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition (arxiv.org)

II. Huấn luyện mô hình

- Build LSTM:
  - Layer LSTM với 50/100 đơn vị

- Optimizer: Adam
- Loss function: Mean absolute percentage error

- Huấn luyện mô hình với thông số:
  - Epochs: 100
  - Batch size: 2



Phần IV: ĐÁNH GIÁ

- Độ đo đánh giá:

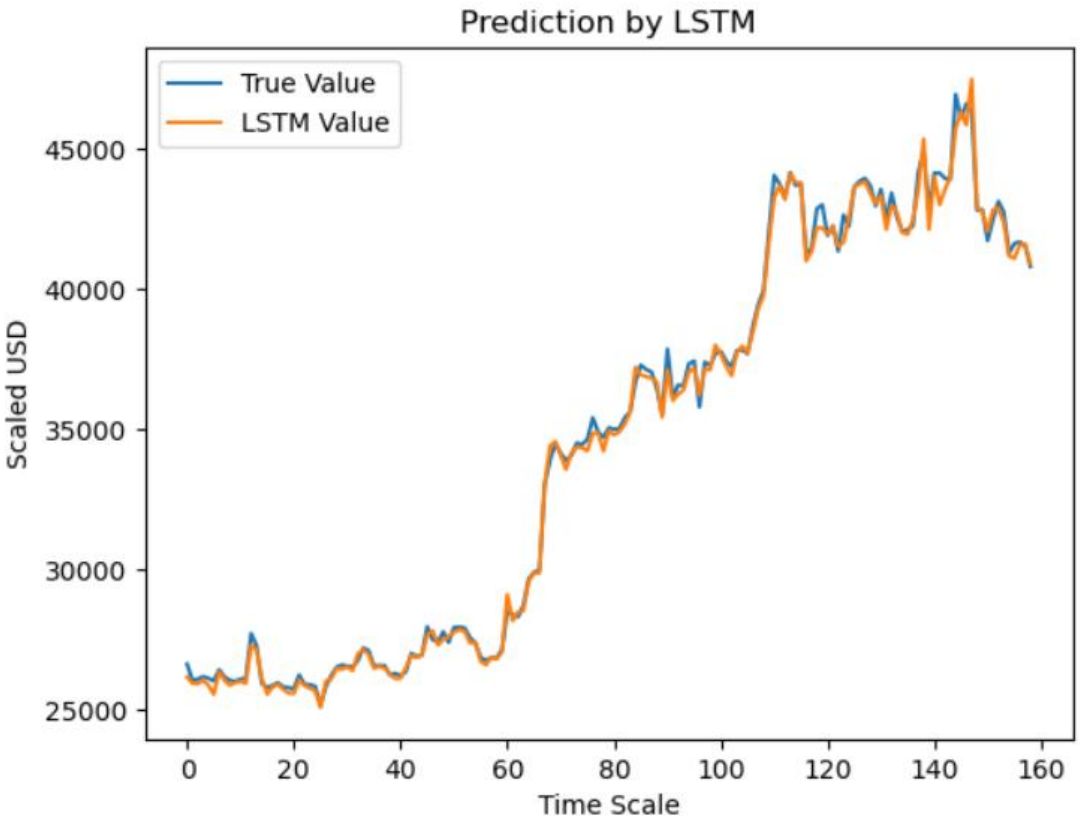
  Tỉ lệ sai số tuyệt đối trung bình MAPE (Mean Absolute Percentage Error)

$$MAPE = \frac{\sum \left(\frac{abs(y_i - f_i)}{y_i} \right)}{n}$$

- Kết quả:

Số đơn vị LSTM	50	100
Áp dụng Minmaxscaler	0.796%	0.787%
Không áp dụng Minmaxscaler	0.525%	0.455%

- So sánh giá trị dự đoán và thực tế của Giá đóng:



Phần V: TỔNG KẾT VÀ HƯỚNG PHÁT TRIỂN

- Đánh giá phương pháp:
  - Dự đoán khá chính xác trên tập thử nghiệm.
  - Phương pháp này chỉ dự đoán 1 ngày tiếp theo.
- Hướng phát triển:
  - Thay vì sử dụng Giá mở, Giá cao nhất, Giá thấp nhất, Khối lượng giao dịch trong ngày, Biến động giá trong ngày, Lãi suất, Giá vàng làm đầu vào cho mô hình, Giá đóng làm đầu ra cho mô hình thì sử dụng tất cả làm đầu vào cho mô hình, và đầu ra sẽ là tất cả thông tin này của ngày tiếp theo.

  ➢ Mô hình sẽ có thể dự đoán nhiều ngày trong tương lai (sử dụng ngày vừa dự đoán thêm vào input để dự đoán ngày kế tiếp).