Ứng dụng LSTM cho dự báo giá cổ phiếu

Introduction

- Việc dự báo giá cổ phiếu là rất quan trọng trong lĩnh vực đầu tư và tài chính. Những dự báo này giúp các nhà đầu tư đưa ra quyết định thông minh về việc mua, bán hoặc giữ cổ phiếu.
- Dự báo giá cổ phiếu có thể được sử dụng để định giá các công ty, đánh giá tính khả thi của các dự án đầu tư và đưa ra quyết định về cổ phiếu nên được giữ hoặc bán. Nó cũng giúp các nhà đầu tư có thể tìm ra những cổ phiếu có tiềm năng tăng giá mạnh trong tương lai.
- Tuy nhiên, việc dự báo giá cổ phiếu không phải là một công việc đơn giản. Thị trường chứng khoán thường xuyên bị ảnh hưởng bởi các yếu tố bên ngoài khác nhau, chẳng hạn như thay đổi trong chính sách kinh tế, biến động giá dầu và nhiều yếu tố khác. Điều này có nghĩa là dự báo giá cổ phiếu chỉ là một trong nhiều yếu tố được sử dụng để đưa ra quyết định đầu tư, và nó không phải là chìa khóa đơn độc để đạt được lợi nhuận.
- Vì vậy, việc dự báo giá cổ phiếu quan trọng, nhưng nó phải được kết hợp với các phân tích khác như phân tích cơ bản và kỹ thuật để đưa ra quyết định đầu tư chính xác.

- Ở trong báo cáo này, tôi sẽ đưa ra một mô hình dự đoán phù hợp dựa trên mạng học sâu cho việc dự báo giá cổ phiếu từ dữ liệu giá cổ phiếu FPT, MSN, VIC và PNJ. Trong bài này, tôi chỉ sử dụng mô hình để dự báo cho giá cổ phiếu FPT, sau đó các dữ liệu khác ta sẽ làm tương tự.

Dữ liệu, Inputs & Outputs

Dữ liệu

- Dữ liệu FPT
  - 97407 records từ 25/12/2018 - 22/12/2020
  - Trong mỗi ngày dữ liệu giá cổ phiếu được ghi lại theo từng phút từ 9h15 - 14h46.
  - Mỗi record bao gồm các thuộc tính sau:
    - Date/Time: thời gian record được ghi lại theo mỗi phút.
    - Ticker: Tên mã cổ phiếu
    - Open: giá mở cửa
    - High: giá cổ phiếu cao nhất lần ghi đó
    - Low: giá cổ phiếu thấp nhất trong lần ghi
    - Close: giá đóng cửa
    - Volume: tổng số cổ phiếu đã giao dịch trong khoảng thời gian ghi
    - Open Interest.
  - Toàn bộ dữ liệu cũng được cung cấp dưới dạng chuỗi dữ liệu theo từng phút, hàng tuần từ thứ 2 đến thứ 6.

Inputs:

- Đầu vào của mô hình là dữ liệu cổ phiếu close được MinMaxScale

Outputs:

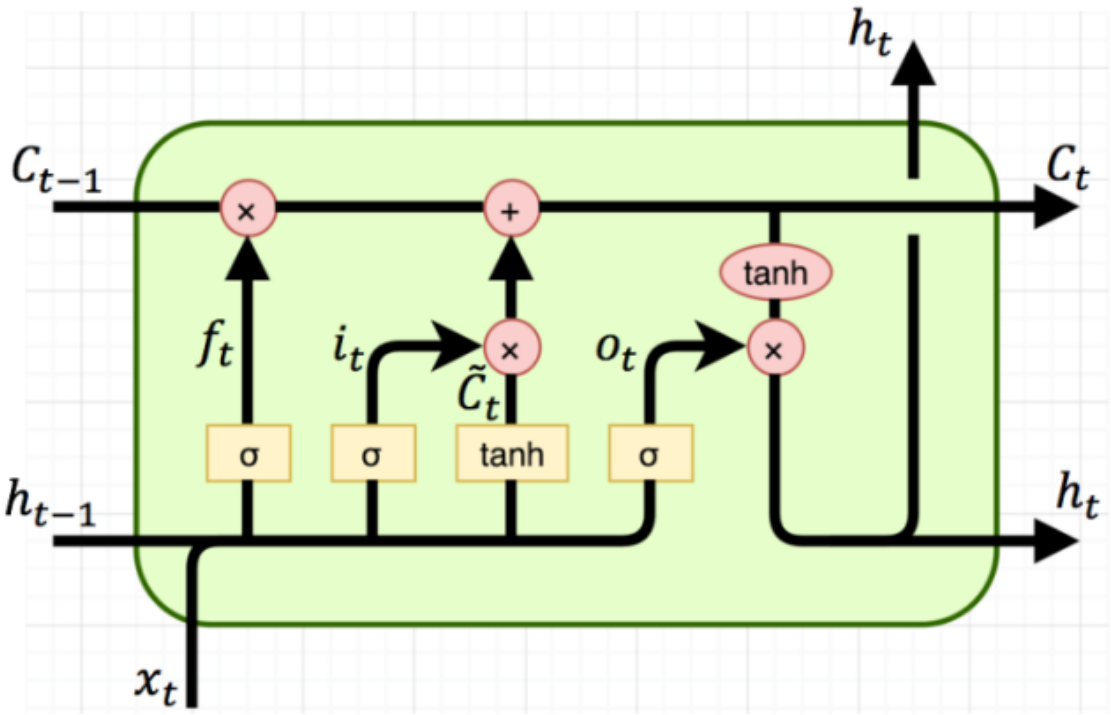
- Giá cổ phiếu closes dưới dạng đầu ra của MinMaxScale sau đó sẽ được inverse_transform

Phương pháp & Mô hình

- Ở đây, chúng tôi sẽ đi xây dựng một mô hình phục vụ cho việc dự báo giá cổ phiếu close của cổ phiếu FPT.
- Theo phương pháp này, tôi xây dựng một mô hình sử dụng việc ghi lại trong dữ liệu training và sau đó dự báo giá close của cổ phiếu FPT cho từng phút sau đó.
- Mô hình được đưa ra là mô hình Long-Short Term Memory (LSTM).

LSTM

- Là một trong những mô hình kiến trúc mạng neural sử dụng trong lĩnh vực xử lý dữ liệu chuỗi thời gian (time-series). Mô hình LSTM được phát triển để giải quyết vấn đề mất mát thông tin trong mô hình RNN (Recurrent Neural Network).



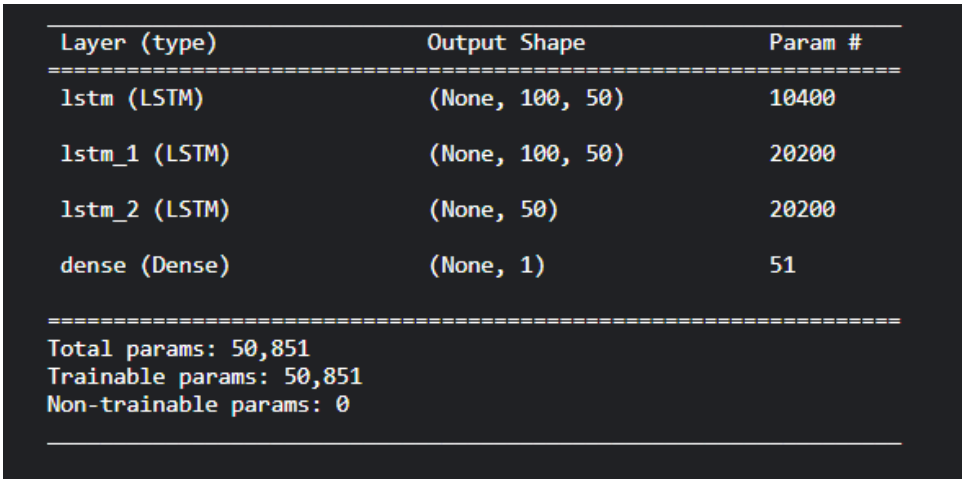
Mô hình LSTM

- Mô hình LSTM sử dụng các cổng để xử lý thông tin được truyền từ các cell trước đó. Các cổng này bao gồm:
  - Cổng quên (forget gate): Quên, loại bỏ thông tin không cần thiết
  - Cổng đầu vào (input gate): quyết định xem cần lấy bao nhiêu từ input của state và hidden layer của layer trước.

- Cổng đầu ra (output gate): quyết định xem cần lấy bao nhiêu từ cell state để trở thành output của hidden state, ngoài ra cũng được lấy để tính ra output cho state t.

LSTM cho bài toán

- Chúng tôi xây dựng mô hình LSTM cho bài toán như hình dưới đây:



Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 100, 50)	10400
lstm_1 (LSTM)	(None, 100, 50)	20200
lstm_2 (LSTM)	(None, 50)	20200
dense (Dense)	(None, 1)	51

Total params: 50,851
Trainable params: 50,851
Non-trainable params: 0

- Mô hình gồm:
  - 3 LSTM:
    - Các lớp LSTM đều gồm 50 đơn vị ẩn (hidden state), tuy nhiên để lấy kết quả của mỗi bước thời gian đầu vào nên tôi đã để return_sequence = True trong 2 lớp LSTM đầu tiên.
    - Sau đó lớp LSTM thứ 3 chúng tôi chỉ lấy giá trị kết quả của bước thời gian cuối cùng.
    - Cuối cùng để đưa ra 1 kết quả, Dense(1) đã được sử dụng.

Thử nghiệm & kết quả

Thử nghiệm

- Mô hình được huấn luyện với bộ dữ liệu training vơi 80% bộ dữ liệu đã cho và testing với 20% còn lại.
- Để đánh giá hiệu suất của mô hình, ta sử dụng MSE (mean squared error)
- Tối ưu hóa mô hình bằng thuật toán tối ưu “adam”

- Sau đó, mô hình được huấn luyện với:
  - 100 epochs
  - Batch size = 4
  - Ngoài ra để giảm thiểu overfitting và không mất thời gian khi training với nhiều epochs, tôi đã dùng early stopping.

Kết quả

- Sau khi sử dụng early stopping, việc huấn luyện đã dừng lại sau 15 epochs với kết quả như bảng sau:

```
Epoch 1/100
1216/1216 [==============================] - 31s 17ms/step - loss: 7.7652e-04 - val_loss: 2.6280e-04
Epoch 2/100
1216/1216 [==============================] - 20s 16ms/step - loss: 3.8950e-05 - val_loss: 9.0852e-05
Epoch 3/100
1216/1216 [==============================] - 19s 16ms/step - loss: 3.9233e-05 - val_loss: 5.1438e-05
Epoch 4/100
1216/1216 [==============================] - 19s 15ms/step - loss: 3.5134e-05 - val_loss: 3.4571e-05
Epoch 5/100
1216/1216 [==============================] - 20s 16ms/step - loss: 2.8143e-05 - val_loss: 1.6380e-04
Epoch 6/100
1216/1216 [==============================] - 19s 16ms/step - loss: 2.4746e-05 - val_loss: 6.6874e-05
Epoch 7/100
1216/1216 [==============================] - 19s 15ms/step - loss: 2.0169e-05 - val_loss: 1.3925e-05
Epoch 8/100
1216/1216 [==============================] - 20s 17ms/step - loss: 1.5863e-05 - val_loss: 1.8543e-05
Epoch 9/100
1216/1216 [==============================] - 19s 16ms/step - loss: 1.3021e-05 - val_loss: 3.0572e-05
Epoch 10/100
1216/1216 [==============================] - 19s 16ms/step - loss: 9.9438e-06 - val_loss: 8.0776e-06
Epoch 11/100
1216/1216 [==============================] - 19s 16ms/step - loss: 1.0259e-05 - val_loss: 1.8484e-05
Epoch 12/100
1216/1216 [==============================] - 19s 15ms/step - loss: 8.2013e-06 - val_loss: 3.0992e-05
Epoch 13/100
1216/1216 [==============================] - 19s 16ms/step - loss: 8.9940e-06 - val_loss: 2.7817e-05
Epoch 14/100
1216/1216 [==============================] - 20s 16ms/step - loss: 8.4798e-06 - val_loss: 9.9429e-06
Epoch 15/100
1216/1216 [==============================] - 19s 15ms/step - loss: 8.7114e-06 - val_loss: 1.3904e-05
```

- Khi training xong mô hình LSTM, tôi đã dùng MSE để đánh giá hiệu suất của mô hình, và kết quả thu được trên tập training và test là:

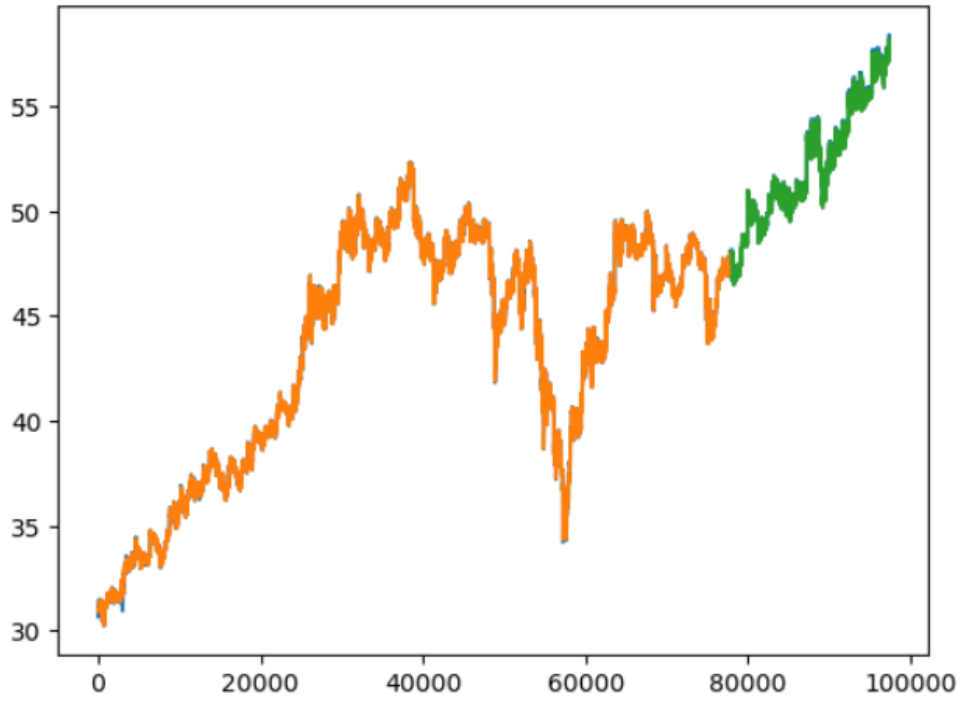
```
math.sqrt(mean_squared_error(y_train, train_pred))
```

```
43.16155795962248
```

```
math.sqrt(mean_squared_error(y_test,test_pred))
```

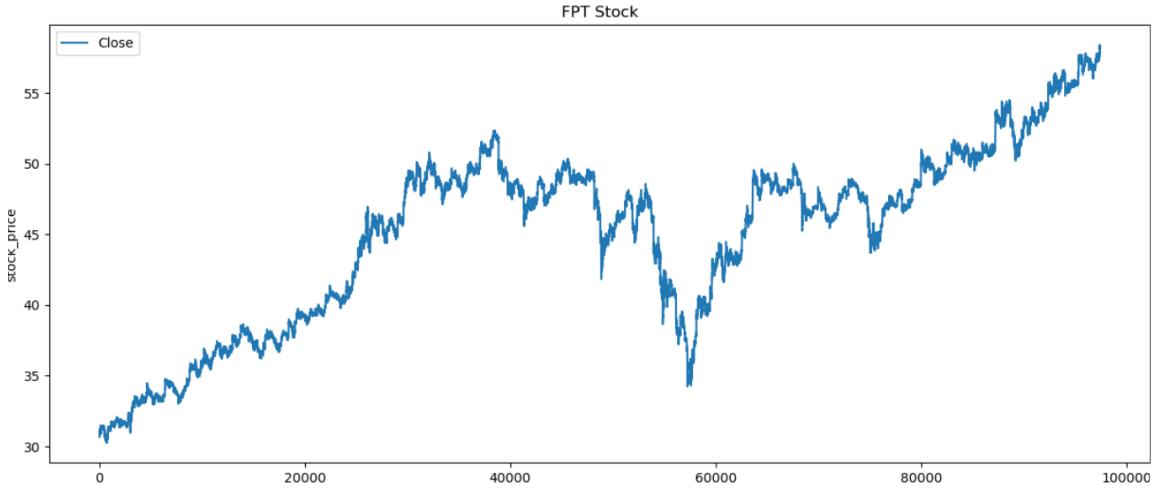
```
51.52582201057387
```

- So sánh kết quả dự đoán trên tập training và test so với kết quả thực tế:

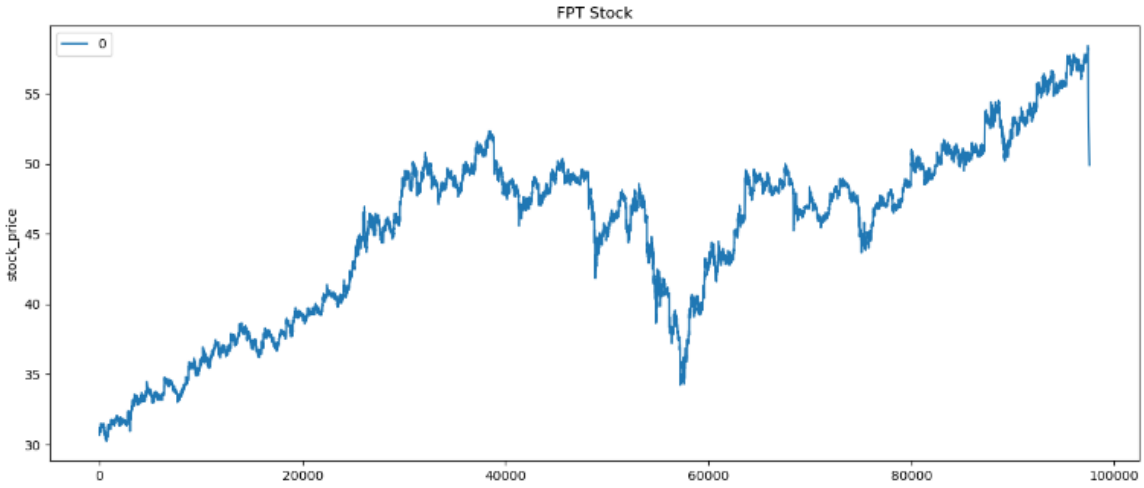


- Như chúng ta thấy như hình trên, giá trị dự đoán trên tập training là phần màu cam, trong khi đó giá trị dự đoán trên tập test là màu xanh lá. Còn phần xanh dương là giá trị thực tế. Do giá trị dự đoán khá khớp nên phần giá trị thực tế dường như ta chỉ thấy thấp thoáng một màu xanh dương ở sau.
- Cuối cùng phần quan trọng nhất đó là dự đoán giá trị cổ phiếu close trong n phút tiếp theo (ở bài báo cáo này, tôi sẽ đi dự đoán theo đơn vị thời gian là phút).
    - Theo bộ dữ liệu được cung cấp, ta thấy trong 1 ngày, giá cổ phiếu được ghi lại bắt đầu từ 9h15 cho tới 14h46 và được ghi 134 lần.

- Do đó nếu để dự đoán giá cổ phiếu vào 9h15 của ngày tiếp theo ta chỉ cần cộng time với 134. Và tương tự với các mốc thời gian khác.
- Và trong notebook, tôi đã dùng mô hình để dự báo giá cổ phiếu trong 134 phút tiếp theo.

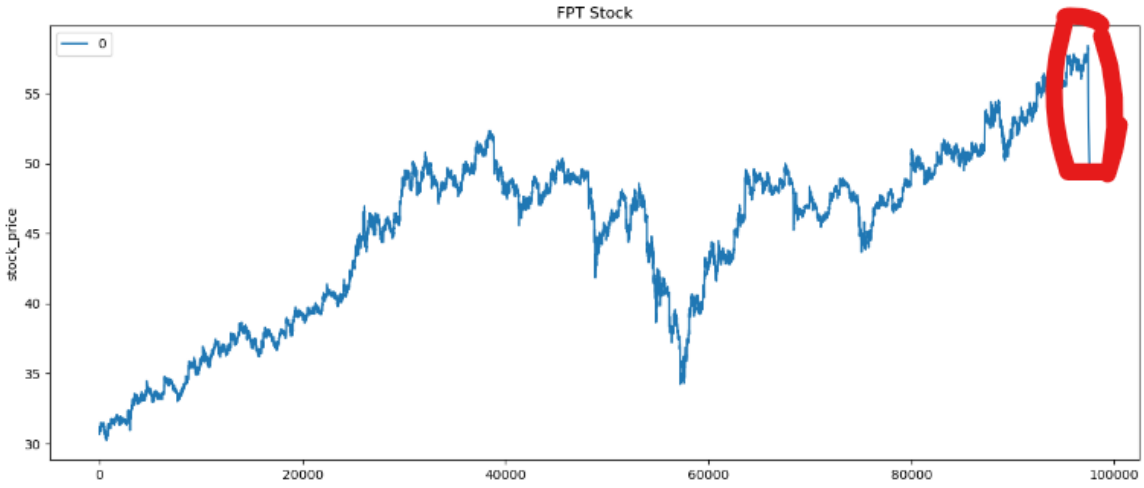


Giá cổ phiếu ban đầu



Giá trị cổ phiếu trong ngày tới

- Từ hai hình trên, ta thấy được giá cổ phiếu tiếp theo sẽ bị giảm khoảng 4 đơn vị trong ngày tiếp theo.



- Như vậy để giải quyết vấn đề của bài toán dự báo sự giao động của giá cổ phiểu sau n phút. Thì ta chỉ cần lấy giá cổ phiếu sau n phút mà ta đã tính được ở trên trừ đi giá cổ phiếu hiện tại.

Kết luận

Sau khi mô hình trên được huấn luyện, ta có thể áp dụng chúng cho các giá cổ phiếu khác như VIC, MSN và PNJ. Và cũng có thể áp dụng cho việc dự báo giá open, high và low.

Tham khảo

- https://www.analyticsvidhya.com/blog/2021/01/understanding-architecture-of-lstm/
- https://nttuan8.com/bai-14-long-short-term-memory-lstm/