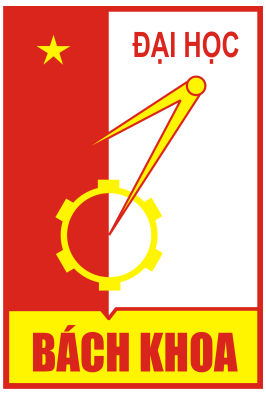
ĐẠI HỌC BÁCH KHOA HÀ NỘI
TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

—————o0o—————-



Bài tập lớn Nhập môn Học máy và khai phá dữ liệu

DỰ ĐOÁN CHUYỂN ĐỘNG GIÁ CỦA CÁC CỔ PHIẾU TRÊN THỊ TRƯỜNG CHỨNG KHOÁN VIỆT NAM

Giảng viên hướng dẫn: **PGS.TS. Thân Quang Khoát**

Sinh viên: **Nguyễn Trọng Bằng - 20190038**

Nguyễn Hải Dương - 20190044

Nguyễn Ngọc Bảo - 20193989

Lớp: **CTTN Khoa học máy tính K64**

Lời nói đầu

Trong thời đại công nghệ phát triển nhanh chóng như hiện nay, việc kết hợp và ứng dụng công nghệ vào trong đời sống càng ngày càng đóng vai trò vô cùng quan trọng, công nghệ không chỉ đem lại lợi ích hàng ngày cho con người mà còn là sự tích lũy cho sự phát triển về sau.

Cùng với sự phát triển của học máy, các ứng dụng của nó đang dần thay đổi rất nhiều mặt trong cuộc sống, đặc biệt là với các lĩnh vực của nền kinh tế. Thị trường chứng khoán cũng không nằm ngoài xu thế này. Việc nghiên cứu xu hướng biến động giá của các loại cổ phiếu không chỉ đem lại ý nghĩa trong kinh tế như dự đoán chính xác được sự ổn định của thị trường có thể đem tới cơ hội đầu tư cho các nhà đầu tư, mà còn có ý nghĩa lớn trong việc nghiên cứu và xây dựng các mô hình dự đoán tổng quát cho dự liệu dạng chuỗi thời gian nói chung, bởi dữ liệu cổ phiếu có thể thay đổi rất bất thường tùy thuộc vào thị trường. Chính vì vậy, nhóm đã quyết định lựa chọn đề tài về dự đoán giá cổ phiếu để có thể hiểu rõ hơn về cách thức các mô hình dự đoán hoạt động, bên cạnh đó nhóm có thể có những kiến thức nền tảng cho những dự án về sau.

Dù đã cố gắng nhưng do thời gian và kiến thức còn hạn chế nên bài báo cáo không thể tránh khỏi những sai sót. Chúng em rất mong được thầy góp ý để rút kinh nghiệm và sửa đổi cho nhưng lần báo cáo sau. Chúng em xin chân thành cảm ơn!

Mục lục

Chương 1

Giới thiệu

1.1 Thị trường chứng khoán Việt Nam

Thị trường chứng khoán là một tập hợp bao gồm những người mua và người bán cổ phiếu (hay chứng khoán), thứ đại diện cho quyền sở hữu của họ đối với một doanh nghiệp; chúng có thể bao gồm các cổ phiếu được niêm yết trên sàn giao dịch chứng khoán đại chúng, hoặc những cổ phiếu được giao dịch một cách không công khai, ví dụ như cổ phần của một công ty tư nhân được bán cho các nhà đầu tư thông qua các nền tảng gọi là vốn cộng đồng. Những khoản đầu tư trên thị trường chứng khoán hầu hết được thực hiện thông qua môi giới chứng khoán và nền tảng giao dịch điện tử.

Hình 1.1: Thị trường chứng khoán

Thị trường chứng khoán đã xuất hiện trên thế giới cách đây hàng thế kỷ, nhưng mới hình thành tại thị trường Việt Nam cách đây hơn 20 năm. Mở đầu

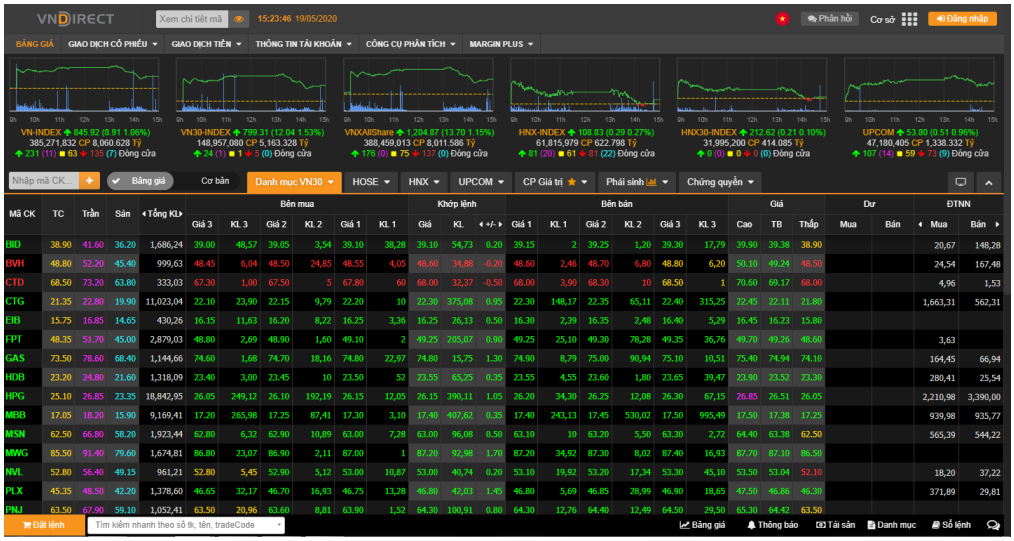
là sự kiện thành lập Ủy ban Chứng khoán Nhà nước Việt Nam, theo Nghị định số 75/CP của Chính phủ, vào ngày 28/11/1996. Hai năm sau, vào ngày 11/7/1998, dựa vào Nghị định số 48/CP của Chính phủ, thị trường chứng khoán Việt Nam chính thức được khai sinh. Lúc này, Trung tâm Giao dịch Chứng khoán TP. Hồ Chí Minh (tiền thân là Sở Giao dịch Chứng khoán Hồ Chí Minh – HOSE) được thành lập.

Năm 2005, Trung tâm lưu ký chứng khoán Việt Nam (VSD) được thành lập. Cùng với đó là sự ra đời của Trung tâm Giao dịch Chứng khoán Hà Nội (tiền thân của Sở Giao dịch Chứng khoán Hà Nội – HSX) vào ngày 8/3/2005. Tiếp đến, ngày 24/6/2009, sàn Upcom đi vào vận hành. Đây trở thành nơi giao dịch cổ phiếu lớn, tuy nhiên vẫn chưa đạt đủ tiêu chuẩn đi niêm yết trên 2 sàn giao dịch TP.HCM (HOSE) và Hà Nội (HNX).

Trải qua hơn 20 năm với những thăng trầm, tính đến thời điểm hiện tại, mức vốn hóa đã tăng nhanh chóng lên tới hơn 82% GDP. Thời gian giao dịch cũng được kéo dài tương tự. Thị trường chứng khoán được bổ sung các loại lệnh giao dịch mới như: lệnh thị trường, ATC... Ngoài ra, chu kỳ thanh toán T+3 được rút ngắn xuống còn T+2. Điều này một phần thể hiện được sự bùng nổ mạnh mẽ của thị trường chứng khoán. Với đà tăng trưởng nổi bật trong cả tiến trình của lịch sử hình thành thị trường chứng khoán Việt Nam, hứa hẹn trong thời gian tới, chứng khoán sẽ là kênh đầu tư bùng nổ hơn nữa.

Một số khái niệm về thị trường chứng khoán được sử dụng trong bài báo cáo này:

- Sàn chứng khoán: Sàn giao dịch chứng khoán là một nền tảng giao dịch các loại chứng khoán trên thị trường. Nhà đầu tư có thể thực hiện việc trao đổi, mua bán, cũng như chuyển nhượng các chứng khoán khác nhau tại đây. 3 sàn giao dịch lớn tại Việt Nam hiện tại là HOSE (Hochiminh Stock Exchange), HNX (Hanoi Stock Exchange), UPCOM (Unlisted Public Company Market).
- Ngày giao dịch: Tại thời điểm hiện tại, một ngày giao dịch được tính từ 9h sáng đến 11h30 sáng và 1h chiều đến 3h chiều các ngày trong tuần trừ thứ bảy, chủ nhật và các ngày lễ theo quy định của bộ Lao Động.



Hình 1.2: Bảng điện tử chứa các thông tin giao dịch

- Biên độ giá: Là biên độ dao động tối đa về giá của một cổ phiếu trong một ngày giao dịch. Các sàn giao dịch khác nhau sẽ có biên độ giao dịch khác nhau: HOSE: 7%, HNX: 10%, UPCOM: 15%.

- VN-INDEX: VN-Index là chỉ số đại diện cho Sở HoSE từ khi thị trường chứng khoán đi vào hoạt động, đại diện cho tất cả cổ phiếu được niêm yết và giao dịch trên HOSE. Chỉ số này được tính theo phương pháp trọng số giá trị thị trường, dựa vào mức độ chi phối của từng cổ phiếu được sử dụng. VN-Index được tính theo công thức:

  VN-Index = (Tổng giá trị thị trường của các cổ phiếu niêm yết hiện tại / Tổng giá trị của các cổ phiếu niêm yết cơ sở) x 100

Hình 1.3: VN Index

- Giá mở cửa: Là giá mở đầu của một ngày giao dịch, được xác định bằng phiên ATO.
- Giá đóng cửa: Là giá kết thúc của một ngày giao dịch, được xác định bằng phiên ATC.
- Giá cao nhất: Là giá cao nhất mà cổ phiếu đó được giao dịch trong một ngày giao dịch.
- Giá thấp nhất: Là giá thấp nhất mà cổ phiếu đó được giao dịch trong một ngày giao dịch.

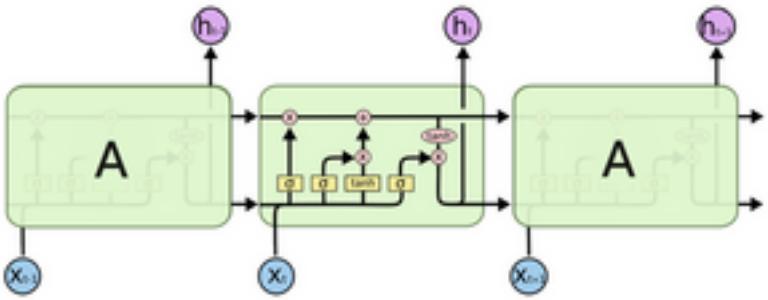
1.2 Bài toán dự đoán giá cổ phiếu

Vấn đề dự báo dữ liệu chuỗi thời gian tài chính, mà cụ thể là dự báo giá cổ phiếu hiện nay chủ yếu được tiếp cận dưới hai dạng, đó là dự báo giá cổ phiếu hoặc xu hướng của giá cổ phiếu sau n ngày. Có thể nói rõ hơn đó là dựa vào thông tin giá cổ phiếu trong quá khứ và hiện tại để dự báo xu hướng hoặc giá cổ phiếu trong tương lai sau n ngày. Trong thực tế vấn đề dự báo giá cổ phiếu được các chuyên gia phân tích tài chính dự đoán dựa vào rất nhiều yếu tố, ví dự như: giá cổ phiếu trong quá khứ, tình hình kinh tế vĩ mô, tình hình chính trị, tình hình hoạt động doanh nghiệp, chu kỳ tăng trưởng, chu kỳ trả lãi. Tuy nhiên, trong báo cáo này chúng em chỉ giới hạn lựa chọn thông tin giá cổ phiếu trong quá khứ và hiện tại để phân tích dự báo. Dự đoán giá cổ phiếu là một bài toán thú vị thu hút được sự quan tâm

của cả các nhà nghiên cứu lẫn các nhà đầu tư. Tuy nhiên, đây cũng là một bài toán rất khó bởi lẽ giá chứng khoán thường rất phức tạp và nhiễu loạn. Đã có nhiều cố gắng dự đoán thị trường tài chính bằng phương pháp phân tích truyền thống cho đến kỹ thuật trí tuệ nhân tạo đặc biệt là mạng nơ ron nhân tạo (ANN). ANN là kỹ thuật được sử dụng nhiều trong lĩnh vực này bởi nó có thể mô tả được mối quan hệ phi tuyến giữa đầu vào với đầu ra. Các bài toán dự báo ngắn hạn thường tại thời điểm t sẽ dự báo thời điểm ngay sau đó là (t + 1), trong khi đối với thị trường chứng khoán Việt Nam nhà đầu tư cần phải dự báo đến thời điểm (t + 2), thời điểm nhà đầu tư mới có thể giao dịch cổ phiếu mới mua.

1.3 Long short term memory (LSTM)

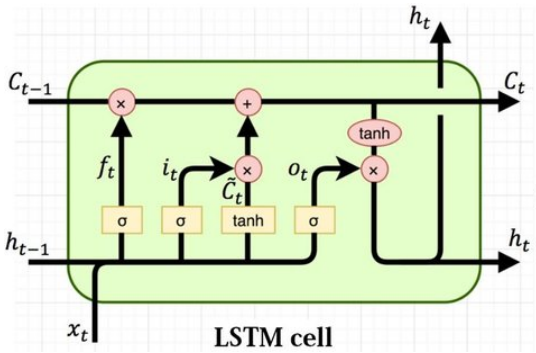
Trong những năm gần đây, các mô hình máy học đã được ứng dụng vào bài toán này để hỗ trợ các nhà đầu tư tạo ra lợi nhuận, tuy nhiên với những mô hình máy học truyền thống thì độ chính xác vẫn còn những hạn chế nhất định. Tuy nhiên với sự phát triển của những mô hình học sâu, việc nhận dạng được những mẩu phi tuyến tính trong chuỗi thời gian của chứng khoán đã trở nên dễ tiếp cận hơn bao giờ hết. Một hướng tiếp cận khá phổ biến và hiệu quả trong những năm gần đây cho bài toán dự đoán chuỗi thời gian là sử dụng mô hình LSTM. Đây là mô hình học sâu thu hút được nhiều sự quan tâm của các nhà nghiên cứu trong và ngoài nước. LSTM được sử dụng rất nhiều cho các bài toán có dữ liệu thời gian hay tuần tự như dịch máy, nhận diện giọng nói, dự báo thời tiết với độ chính xác cao.



Hình 1.4: Cấu trúc LSTM

LSTM lần đầu tiên được đề xuất bởi Hochreiter và Schmidhuber vào năm 1997 [**?**], và sau đó trở nên rất phổ biến, đặc biệt là để giải quyết các vấn

đề dự đoán chuỗi thời gian. Là một biến thể của phương pháp RNN, LSTM hoạt động tốt trên nhiều vấn đề và hiện đang được sử dụng rộng rãi. LSTM thu thập dữ liệu trong một khoảng thời gian, bằng cách sử các đơn vị cổng và ô nhớ trong thiết kế mạng nơ-ron. Các ô nhớ có các trạng thái ô lưu trữ dữ liệu trải nghiệm gần đây. Mỗi thời điểm thông tin đến một ô nhớ, kết quả được kiểm soát thông qua sự kết hợp của trạng thái ô, và sau đó, trạng thái ô được làm mới. Bây giờ, nếu ô nhớ nhận được bất kỳ thông tin nào khác, đầu ra sẽ được xử lý bằng cách sử dụng cả thông tin mới này và trạng thái ô được làm mới. LSTM có thể trích xuất hoặc loại trừ dữ liệu đến hoặc từ ô trước đó. Những dữ liệu này được quản lý một cách có chủ ý bởi các cấu trúc được gọi là cổng. Cổng là một cách tiếp cận để kiểm soát xem dữ liệu có thể đi vào trạng thái ô hay không. Gate là sự kết hợp của một hàm sigmoid và một quá trình nhân điểm. Hàm sigmoid có thể tạo ra bất kỳ số nào từ 0 đến 1. Giá trị này kiểm soát việc truyền dữ liệu như sau: số 0 biểu thị "không vượt qua bất cứ điều gì" trong khi giá trị 1 biểu thị "vượt qua mọi thứ". Đối với mô hình LSTM, các cổng khác nhau được sử dụng để truyền dữ liệu gần đây của từ ô này sang ô khác. Các cổng này được gọi là Update Gate, Forget gate và Output gate. Các ô này được sử dụng để điều khiển bộ nhớ của mô hình LSTM.



$$i_t = \sigma\left(x_t U^i + h_{t-1} W^i\right)$$
$$f_t = \sigma\left(x_t U^f + h_{t-1} W^f\right)$$
$$o_t = \sigma\left(x_t U^o + h_{t-1} W^o\right)$$
$$\tilde{C}_t = \tanh\left(x_t U^g + h_{t-1} W^g\right)$$
$$C_t = \sigma\left(f_t * C_{t-1} + i_t * \tilde{C}_t\right)$$
$$h_t = \tanh(C_t) * o_t$$

Hình 1.5: Tính toán luồng dữ liệu trong LSTM
Nguồn: ResearchGate - Structure of the LSTM cell

Chương 2

Mô hình bài toán học máy

Do các cổ phiếu cùng ngành hoặc có các mối quan hệ kinh tế mật thiết, lợi ích theo nhóm cổ phiếu trên thị trường chứng khoán nên trong bài báo cáo này chúng em hướng đến mô hình dự đoán giá của cổ phiếu mục tiêu dựa trên lịch sử thông tin giao dịch gần đây của chính cổ phiếu mục tiêu và 4 cố phiếu có độ tương quan cao với cổ phiếu mục tiêu. Yêu cầu của hệ thống học máy cho bài toán dự đoán giá cổ phiếu được phát biểu như sau:

- Đầu vào: Các thông tin giao dịch trong 14 ngày gần nhất của 4 cổ phiếu tương quan và cổ phiếu mục tiêu. Thông tin giao dịch một ngày bao gồm 4 yếu tố: Giá cao nhất (max^i_j), giá thấp nhất (min^i_j), giá mở cửa ($open^i_j$), giá đóng cửa ($close^i_j$), biên độ dao động d. Trong đó i thể hiện vị trí tương đối so với ngày giao dịch hiện tại. j để xác định cổ phiếu

$$\begin{aligned} input = [&\{max^{-13}_{target}, min^{-13}_{target}, open^{-13}_{target}, close^{-13}_{target}\}, \\ &\ldots, \{max^{0}_{target}, min^{0}_{target}, open^{0}_{target}, close^{0}_{target}\}, \\ &\{max^{-13}_{corr1}, min^{-13}_{corr1}, open^{-13}_{corr1}, close^{-13}_{corr1}\}, \\ &\ldots, \{max^{0}_{corr1}, min^{0}_{corr1}, open^{0}_{corr1}, close^{0}_{corr1}\} \\ &\{max^{-13}_{corr2}, min^{-13}_{corr2}, open^{-13}_{corr2}, close^{-13}_{corr2}\}, \\ &\ldots, \{max^{0}_{corr2}, min^{0}_{corr2}, open^{0}_{corr2}, close^{0}_{corr2}\} \\ &\{max^{-13}_{corr3}, min^{-13}_{corr3}, open^{-13}_{corr3}, close^{-13}_{corr3}\}, \\ &\ldots, \{max^{0}_{corr3}, min^{0}_{corr3}, open^{0}_{corr3}, close^{0}_{corr3}\} \\ &\{max^{-13}_{corr4}, min^{-13}_{corr4}, open^{-13}_{corr4}, close^{-13}_{corr4}\}, \\ &\ldots, \{max^{0}_{corr4}, min^{0}_{corr4}, open^{0}_{corr4}, close^{0}_{corr4}\}, d] \end{aligned} \tag{2.1}$$

- Đầu ra: Các thông tin giao dịch dự đoán trong 2 ngày tiếp theo. 2 ngày

là số ngày tối thiểu để cổ phiếu mua được có thể giao dịch, do đó khi thực hiện mua một cổ phiếu, ta có nhu cầu dự đoán giá cổ phiếu trong 2 ngày sắp tới để tránh mua vào đợt giá giảm.

$$output = [\{pmax^1_{target}, pmin^1_{target}, popen^1_{target}, pclose^1_{target}\}, \{pmax^2_{target}, pmin^2_{target}, popen^2_{target}, pclose^2_{target}\}] \tag{2.2}$$

- Ràng buộc: Giá cao nhất, giá thấp nhất, giá đóng cửa không được vượt quá biên độ giao động so với giá mở cửa.

$$|\frac{pmax^i_{target}}{popen^i_{target}} - 1| \leq d \tag{2.3}$$

$$|\frac{pmin^i_{target}}{popen^i_{target}} - 1| \leq d \tag{2.4}$$

$$|\frac{pclose^i_{target}}{popen^i_{target}} - 1| \leq d \tag{2.5}$$

Chương 3

Dữ liệu huấn luyện

3.1 Thu thập dữ liệu

Dữ liệu chứng khoán được thu thập từ trang web https://cafef.vn bao gồm lịch sử giao dịch của tất cả các cổ phiếu và thông tin về các chỉ số thị trường (VN-Index, VN30, ...) từ 28/7/2000 đến 20/5/2022. Dữ liệu bao gồm 3 file .csv là thông tin giao dịch từng ngày của các cổ phiếu ở 3 sàn lớn ở Việt Name (Upcom, HNX, HSX) và 1 file .csv là thông tin qua từng ngày của các chỉ số thị trường (VNINDEX, VN30). Mỗi file .csv bao gồm 7 feature:

- 'Ticker': Mã cổ phiếu hoặc tên chỉ số thị trường.
- 'Open': Giá mở cửa
- 'Close': Giá đóng cửa
- 'High': Giá cao nhất trong ngày
- 'Low': Giá thấp nhất trong ngày
- 'YYYYMMDD': Ngày giao dịch
- 'Volume': Khối lượng giao dịch

3.2 Lọc tách dữ liệu

Do các dữ liệu được lấy từ nguồn miễn phí cũng như do các chiêu trò, những điểm đen trên một thị trường non trẻ như thị trường Việt Nam, dữ liệu sau khi thu thập về cần được xử lí nhằm giải quyết các vấn đề sau:

1. Vấn đề lớn nhất với dữ liệu là sự tồn tại của những cổ phiếu bẩn hay còn gọi là những cổ phiếu bị lái quá mạnh, cổ phiếu bơm thổi trong khi tiềm năng phát triển của công ty là không có. Những cổ phiếu này từng một thời làm thị trường Việt Nam điêu đứng, làm mất niềm tin nơi nhà đầu tư trong nước cũng như nước ngoài. Đặc điểm của những cổ phiếu này là yếu tố nội tại doanh nghiệp rất bình thường hoặc có phần yếu kém nhưng lại có những đợt tăng trần hàng chục phiên. Và theo sau đó là những đợt xuống giá không phanh. Điển hình nhất mà cũng là vết nhơ nhất của thị trường chứng khoán Việt Nam là trường hợp của cổ phiếu ROS là công ty con của tập đoàn FLC của ông Trịnh Văn Quyết. Do đó, việc đầu tiên trong việc xử lí dữ liệu là lọc bỏ những cổ phiếu này ra khỏi dữ liệu huấn luyện.

Hình 3.1: Cổ phiếu ROS tăng lên hơn 200000 VND một cổ phiếu và giảm về 3000 VND

2. Ngoài những cổ phiếu bị thao túng, một số cổ phiếu nằm trong diện bị hạn chế giao dịch do thời gian dài có vốn chủ sở hữu âm cũng cần lọc bỏ vì chúng có thời gian giao dịch và chuyển động giá khác với mặt bằng chung. Những chứng chỉ quỹ và các hợp đồng tương lai cũng lọc bỏ vì chúng có chuyển động giá khác, sớm pha hoặc chậm pha so với thị trường cơ sở. Ngoài ra những cổ phiếu không có thanh khoản quá

một tuần cũng được loại bỏ. Những cổ phiếu không còn được niêm yết tại thời điểm hiện tại cũng được loại bỏ khỏi danh sách.

3. Ở giai đoạn đầu của thị trường chứng khoán từ 2000 - 2010, số lượng cổ phiếu niêm yết ít cũng như vốn hóa thị trường nhỏ, số lượng nhà đầu tư tham gia thị trường ít nên các giao dịch thường hạn chế và bất quy tắc. Do đó giai đoạn này cũng được cắt đi nhằm tránh những nhiễu dữ liệu có thể xảy ra.

4. Do nguồn dữ liệu là miễn phí nên không tránh khỏi những dữ liệu NULL, NAN. Việc điền các liệu này được thực hiện như sau:

   - Nếu dữ liệu lỗi thuộc feature 'Open' thì sẽ được gán bằng giá 'Close' của ngày hôm trước.
   - Nếu dữ liệu feature 'High' lỗi thì sẽ được gán bằng max('Close', 'Open')
   - Nếu dữ liệu feature 'Low' lỗi thì sẽ được gán bằng min('Close', 'Open')
   - Nếu dữ liệu lỗi thuộc feature 'Close thì sẽ được gán bằng 'Open'. Coi như cổ phiếu đó giữ nguyên giá sau phiên giao dịch.
   - Với một cổ phiếu lỗi hơn 10% dữ liệu sẽ bị loại bỏ khỏi tập dữ liệu.

5. Với những trường giá trị vi phạm ràng buộc 2.3, 2.4, 2.5, chúng sẽ được điều chỉnh đúng bằng cận của ràng buộc.

6. Đối với những cổ phiếu thực hiện chia tách cổ phiếu hoặc trả thưởng cổ tức, giá cổ phiếu sẽ được điều chỉnh chung theo công thức

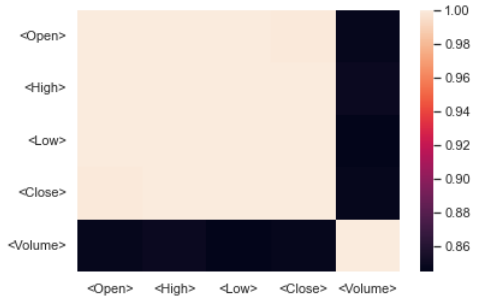
$$P' = \frac{P + P_\alpha \times \alpha - C}{1 + \alpha + \beta} \tag{3.1}$$

   Với P là giá hiện tại, P' là giá tại ngày giao dịch không hưởng quyền, P_α là giá cổ phiếu phát hành thêm, α là tỷ lệ cổ phiếu phát hành thêm, β là tỷ lệ cổ phiếu thưởng, C là cổ tức tiền

3.3 Phân tích dữ liệu

Sau quá trình tách và lọc dữ liệu, chúng em còn 20 cổ phiếu đạt tiêu chuẩn về chất lượng và độ minh bạch của dữ liệu trong số hơn 400 cổ phiếu của

thị trường chứng khoán Việt Nam. Dữ liệu của 20 cổ phiếu này được đưa vào đánh giá độ tương quan giữa các feature cũng như tương quan với chính chúng trong quá khứ. Qua hình 3.2 ta thấy độ tương quan của Volume với các thuộc tính về giá khá nhỏ so với độ tương quan giữa các thuộc tính giá với nhau nên trong báo cáo này chúng em bỏ qua Volume trong quá trình dự đoán giá. Hình 3.3 thể hiện sự tương quan với quá khứ của các thuộc tính giá. Ta có thể thấy thuộc tính Close có độ tương quan thấp nhất. Điều đó trên thực tế là chuẩn xác vì ở thị trường mới nổi như Việt Nam, các chiêu trò ở phiên ATC đã làm lũng đoạn giá đóng cửa của thị trường khá nhiều. Điều đó cũng đặt ra thách thức khi dự đoán feature Close.

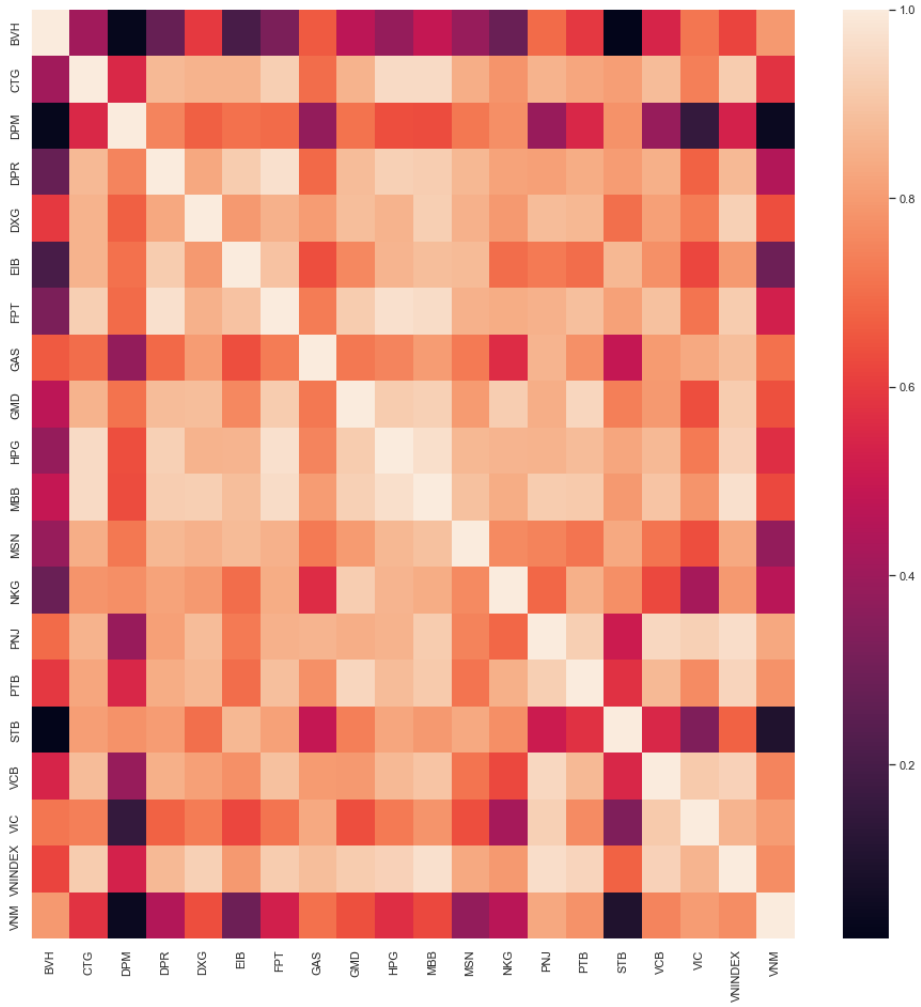


Hình 3.2: Độ tương quan giữa các feature

variables	VIF
<Open>	4186.163045
<High>	3842.119505
<Low>	4175.044501
<Close>	3238.355838

Hình 3.3: Độ tương quan của các feature với chính nó trong quá khứ

Các cổ phiếu được chọn được đưa và phân tích dữ liệu để tìm ra những cổ phiếu có độ tương quan cao nhất. Dựa vào bảng tương quan, chúng em có thể thấy một số tương quan giữa các ngành đã được thể hiện vào giá cổ phiếu của chúng. Chẳng hạn ngành bảo hiểm (BVH) và ngành ngân hàng (CTG, EIB, VCB, MBB) có tương quan ít với nhau do khi nền kinh tế trong giai đoạn tăng trưởng, nhu cầu tín dụng của xã hội cao dẫn đến kì vọng của ngành ngân hàng cao, còn ngành bảo hiểm lại có nhiều kì vọng phát triển trong các giai đoạn chững lại của nền kinh tế. VNM và DPM cũng có độ tương quan thấp do VNM là doanh nghiệp với ngành nghề chính là các sản phẩm liên quan đến sữa, còn DPM là doanh nghiệp hóa chất phân bón, giá của DPM tăng mạnh khi có những dấu hiệu tăng giá phân bón, đó lại là nguồn đầu vào của VNM dẫn đến sự đối lập phát triển của hai ngành. Giữa các ngành chúng ta có thể thấy ngành ngân hàng (CTG, EIB, VCB, MBB, STB) có độ tương quan lẫn nhau cao và đồng đều nên trong báo cáo lần này chúng em sẽ tập trung phân tích và dự đoán 5 mã cổ phiếu này.



Hình 3.4: Độ tương quan giữa 20 cổ phiếu được chọn

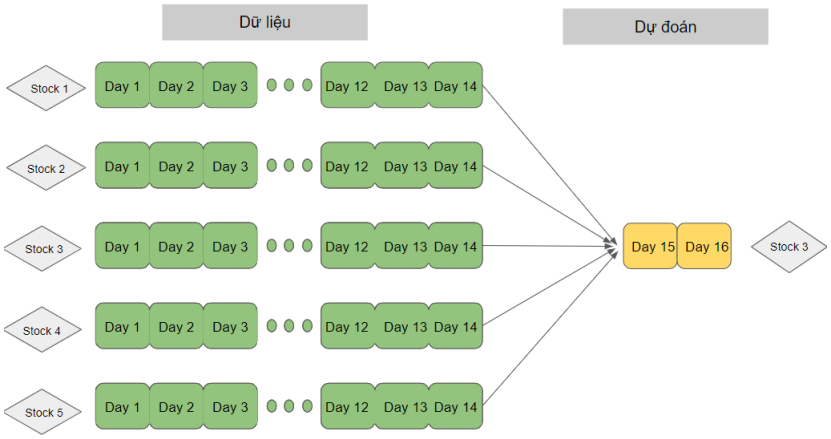
Chương 4

Phương pháp dề xuất

4.1 Mô hình đề xuất

Việc sử dụng dữ liệu quá khứ để dự đoán giá cho chính cổ phiếu đó trong tương lai là khả thi, tuy nhiên, trong thực tế, có nhiều mã cổ phiếu có liên quan và có sự ảnh hưởng nhất định đến nhau trên thị trường, do đó khi sử dụng thêm các cổ phiếu khác trong việc dự đoán một cổ phiếu mục tiêu có thể làm tăng khả năng dự đoán chính xác hơn. Bên cạnh đó, việc lựa chọn các cổ phiếu có tương quan cao để sử dụng là vô cùng quan trọng, các cổ phiếu này giúp mô hình nắm bắt được xu hướng thay đổi của cổ phiếu mục tiêu tốt hơn.

Trong phương pháp đề xuất, nhóm lựa chọn 5 cổ phiếu có tương quan cao, trong đó có 1 cổ phiếu đóng vai trò là mục tiêu dự đoán. Nhóm sử dụng thông tin giá trong 14 ngày của các cổ phiếu, nhằm dự đoán giá 2 ngày tiếp theo của cổ phiếu mục tiêu. Minh họa như trong hình dưới đây:

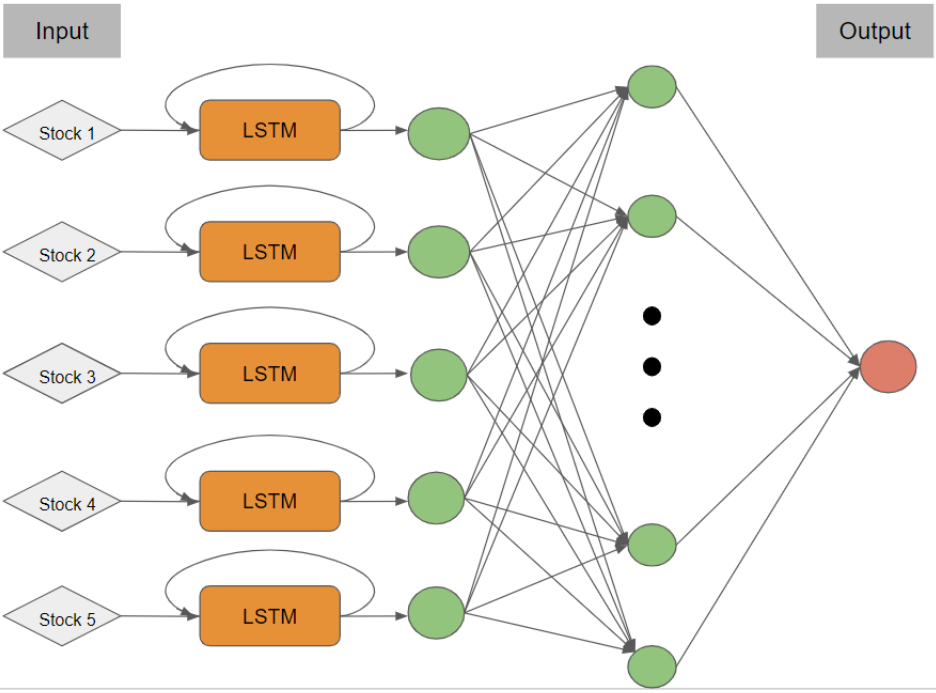


Hình 4.1: Minh họa bài toán với Stock 3 là cổ phiếu mục tiêu

4.2 Kiến trúc mô hình

Nhóm đề xuất mô hình deep-learning sử dụng các khối LSTM cho từng cổ phiếu riêng biệt sau khi các cổ phiếu này đã được phân nhóm với độ tương quan cao, sau đó sử dụng các lớp neural-network để tổng hợp thông tin tất cả các cổ phiếu này. Việc sử dụng LSTM cho từng cổ phiếu sẽ giúp mô hình nằm bắt được sự thay đổi theo thời gian của cổ phiếu này, sau đó, lớp Fully Connected sẽ tổng hợp thông tin của tất cả các cổ phiếu để thực hiện dự đoán.

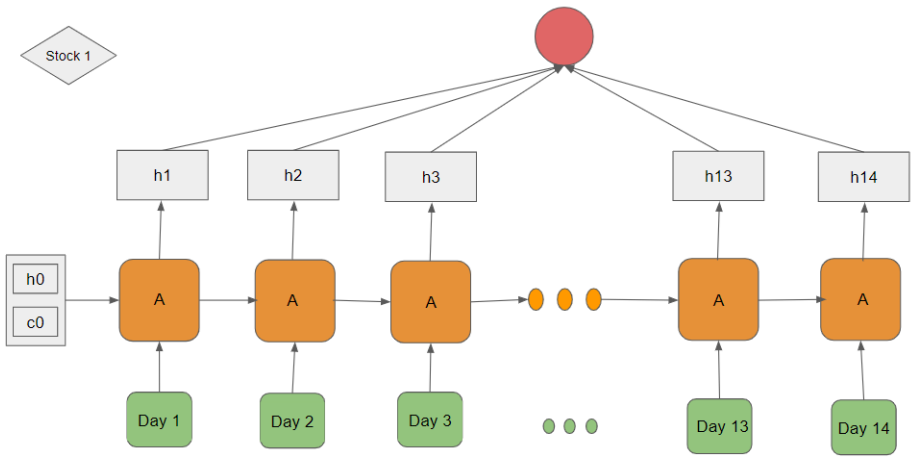
Kiến trúc tổng quát của mô hình như sau:



Hình 4.2: Kiến trúc tổng quát mô hình

4.3 Kiến trúc LSTM

Trong mỗi khối LSTM, mô hình sử dụng lần lượt dữ liệu trong 14 ngày để làm đầu vào, đầu ra của mô LSTM ở tất cả các bước được tổng hợp lại qua một lớp Fully Connected để làm đầu ra cuối cùng của LSTM.



Hình 4.3: Kiến trúc của các khối LSTM

Kiến trúc cụ thể của các khối A trong LSTM trên hình đã được đề cập cụ thể ở 1.3.
Sự đổi mới của mô hình đề xuất có hai điểm

- Đầu tiên là các khối dự đoán LSTM sử dụng thông tin chuỗi thời gian từ một cổ phiếu để đưa ra dự đoán về xu hướng và giá trị cho cổ phiếu đó, các thông tin này sau đó được tổng hợp lại bởi các lớp fully-connected nhằm tổng hợp các thông tin từ các cổ phiếu cùng nhóm nhằm đưa ra dự đoán cho một cổ phiếu chung. Các khối dự đoán cho từng cổ phiếu và các khối tổng hợp lần lượt xử lý thông tin cục bộ của từng cổ phiếu và thông tin không gian chung của các cổ phiếu, dẫn đến hiệu quả và kết quả dự đoán có thể giải thích được

- Các thông tin về không gian được kết hợp giữa các cổ phiếu cùng nhóm ngành và có độ tương quan dữ liệu cao với cổ phiếu nhóm lựa chọn để dự đoán thay vì sử dụng dữ liệu từ tất cả các cổ phiếu trên thị trường, điều này vừa giúp giảm thiểu vấn đề về khả năng tính toán vừa tăng khả năng dự đoán cho mô hình.

Chương 5

Thực nghiệm

5.1 Hàm mất mát

MSE Loss: Nhóm sử dụng hàm Mean Square Error làm hàm mất mát.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \tag{5.1}$$

5.2 Huấn luyện mô hình

5.2.1 Tinh chỉnh mô hình

Nhóm sử dụng thuật toán tối ưu Adam và thực hiện tinh chỉnh 2 siêu tham số của mô hình gồm: số lượng nơ-ron tầng ẩn thuộc lớp Fully Connected trong hình 4.2 và tốc độ học.

Số lượng nơ-ron tầng ẩn

Bảng 1: Hiệu suất thử nghiệm trên tập valid

Số lượng nơ-ron	MAPE	MSE
32	2.27	0.12
64	1.65	0.075
128	1.85	0.083

Dựa trên kết quả thử nghiệm, nhóm lựa chọn số lượng nơ-ron tầng ẩn là 64 nơ-ron.

Learning rate

Bảng 2: Hiệu suất thử nghiệm trên tập valid

Số lượng nơ-ron	MAPE	MSE	Số epochs
0.01	1.75	0.08	100
0.001	1.65	0.075	105
0.0001	3.34	0.25	300

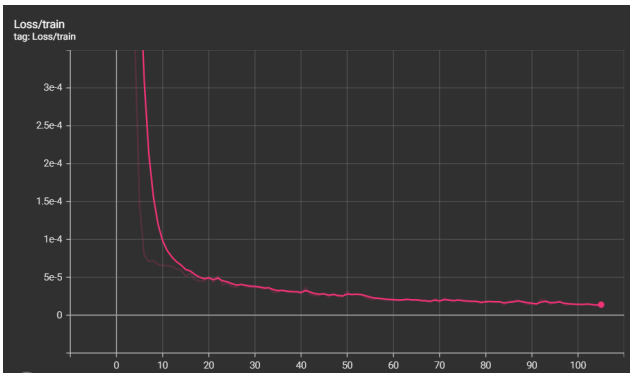
Quá trình thực nghiệm có sự khác nhau giữa số lượng epoch do nhóm sử dụng kỹ thuật Early Stopping. Dựa trên kết quả thử nghiệm, nhóm lựa chọn learning rate cho thuật toán tối ưu Adam là 0.001

5.2.2 Mô hình hoàn thiện

Thuật toán tối ưu Adam, với learning rate = 0,001.
Mô hình được huấn luyện với số epochs là 500, tuy nhiên nhóm đã sử dụng thêm cơ chế early-stopping trong quá trình học để giảm thiểu vấn đề overfitting, và mô hình huấn luyện của nhóm thường dừng học sớm sau 100 epochs, batch size = 32.
GPU: NVIDIA Tesla K80 (Google Colab)

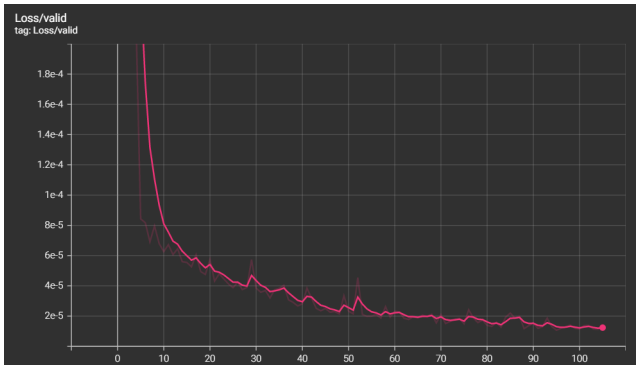
Lỗi trong quá trình huấn luyện

- Training Loss:



Hình 5.1: Lỗi trên tập huấn luyện

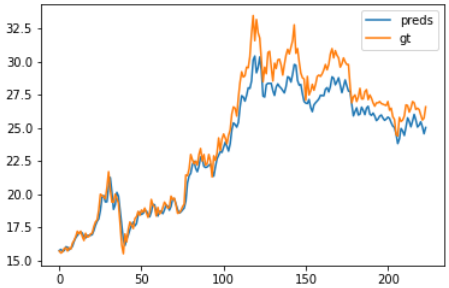
- Valid Loss:



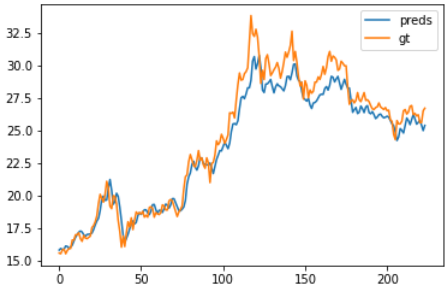
Hình 5.2: Lỗi trên tập valid

5.3 Kết quả

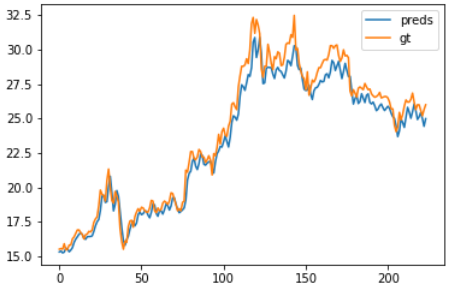
Kết quả dự đoán ngày 1:



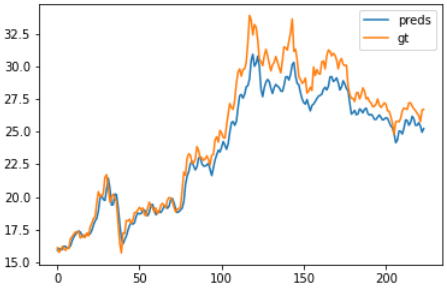
Hình 5.3: Open Day 1



Hình 5.4: Close Day 1

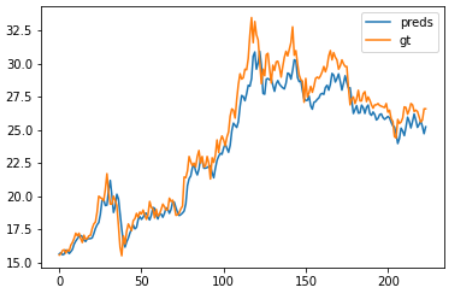


Hình 5.5: Low Day 1

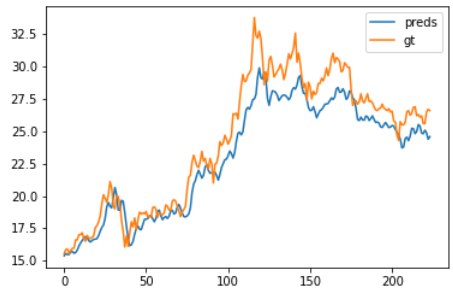


Hình 5.6: High Day 1

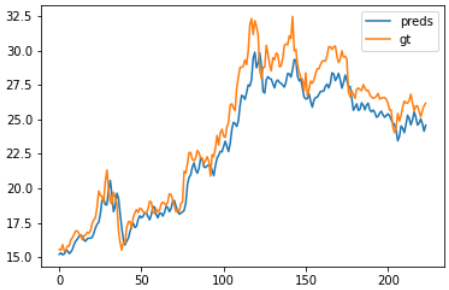
Kết quả dự đoán ngày 2:



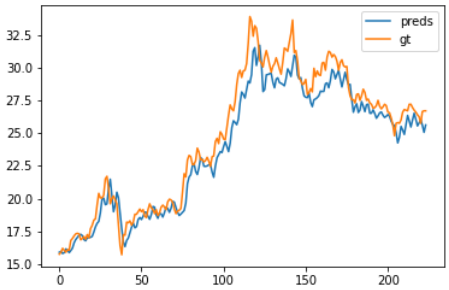
Hình 5.7: Open Day 2



Hình 5.8: Close Day 2



Hình 5.9: Low Day 2



Hình 5.10: High Day 2

Hiệu suất của mô hình trên tập test ứng với từng feature

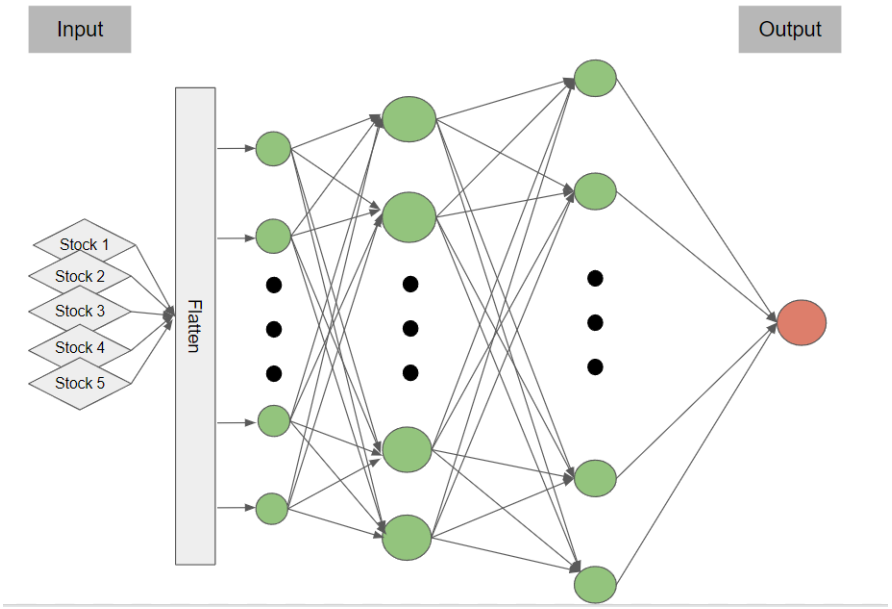
Bảng 3: Hiệu suất trên tập test của mô hình đề xuất.

Feature	MAE	MAPE	MSE	R2
Open	0.99	4.02	1.59	0.94
Close	1.16	4.81	2.24	0.91
Low	0.99	4.16	1.55	0.94
High	1.08	4.37	1.37	0.93

Kết quả trên cho thấy việc học với mô hình dự đoán có sử dụng kiến trúc như LSTM giúp giải quyết khá tốt đối với dữ liệu đầu vào dạng chuỗi thời gian do cơ chế ghi nhớ thông qua trạng thái của các cell trong LSTM. Việc dự đoán cũng cho thấy được xu hướng tăng giảm sau nhiều ngày dự đoán liên tiếp. Tuy nhiên, ở những ngày cuối trong tập kiểm tra, kết quả dự đoán có xu hướng sai lệch lớn hơn so với thực tế, nhất là với những ngày có giá cổ phiếu cao đột ngột.

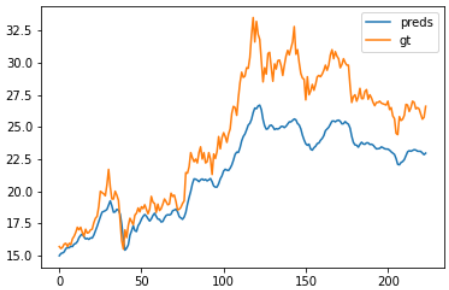
5.4 Thí nghiệm với mô hình không có các lớp LSTM

Trong thí nghiệm này, nhóm tiến hành loại bỏ các khối LSTM dùng để đưa ra dự đoán cục bộ cho từng cổ phiếu, chỉ giữ lại các lớp tổng hợp thông tin tuyến tính của các cổ phiếu và xem xét kết quả thu được từ việc dự đoán:

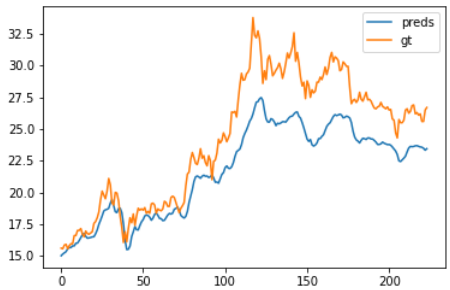


Hình 5.11: Kiến trúc của mô hình sau khi loại bỏ LSTM

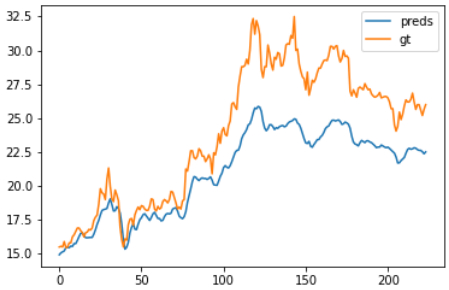
Kết quả dự đoán ngày 1:



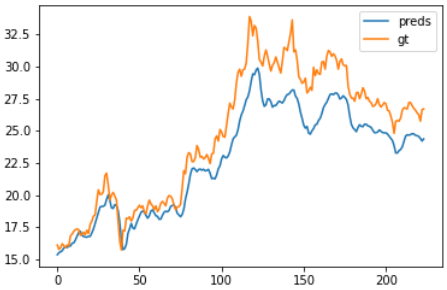
Hình 5.12: Open Day 1



Hình 5.13: Close Day 1

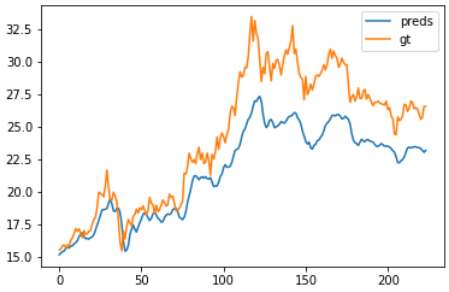


Hình 5.14: Low Day 1

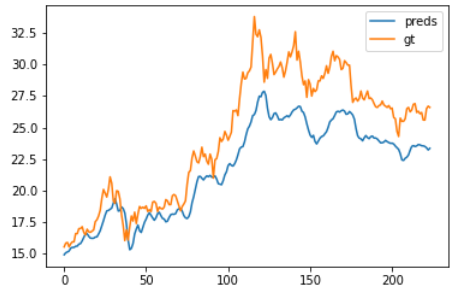


Hình 5.15: High Day 1

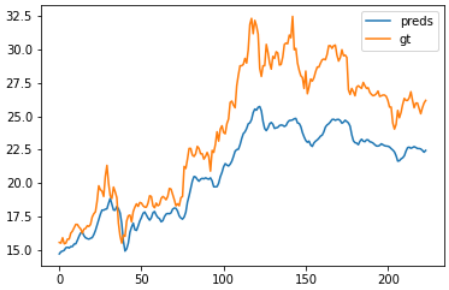
Kết quả dự đoán ngày 2:



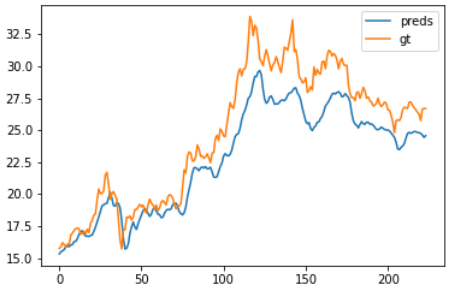
Hình 5.16: Open Day 2



Hình 5.17: Close Day 2



Hình 5.18: Low Day 2



Hình 5.19: High Day 2

Bảng 4: Hiệu suất trên tập test của mô hình thử nghiệm

Feature	MAE	MAPE	MSE	R2
Open	1.72	6.52	4.27	0.84
Close	2.26	8.84	7.05	0.74
Low	1.76	6.88	4.83	0.82
High	2.49	9.52	8.58	0.71

Từ kết quả dự đoán của mô hình sau khi loại bỏ đi các lớp LSTM, có thể thấy rằng việc dự đoán đã tệ hơn rất nhiều so với mô hình có sử dụng LSTM mà nhóm đưa ra trước đó, ngay cả khi nhóm đã thực hiện tinh chỉnh trên mô hình này. Kết quả dự đoán với những ngày đầu thấp hơn và càng về những ngày sau thì giá cổ phiếu được dự đoán càng thấp hơn rất nhiều so với thực tế. Mô hình được thí nghiệm cũng gần như không thể bắt được các giá trị cổ phiếu cao hơn rất nhiều so với bình thường. Sự chênh lệch này cho thấy rằng việc học bằng mô hình dự đoán chỉ sử dụng các lớp neuron tuyến tính

mà không có các cơ chế giúp ghi nhớ thông tin cần thiết xuyên suốt chuỗi thời gian đầu vào sẽ khó có thể giúp việc học được cải thiện, thậm chí có thể khiến việc học trở nên tệ đi rất nhiều.

5.5 Nhận xét chung

Việc dự đoán với mô hình sử dụng các khối LSTM cho từng loại cổ phiếu để đưa ra dự đoán cho một loại cổ phiếu mà nhóm đề xuất để giải quyết bài toán cho ra kết quả tốt (như đã trình bày ở 5.3). Mô hình dự đoán của nhóm đã dự đoán được khá chính xác với xu hướng cũng như giá trị thực tế nhờ vào việc sử dụng các cơ chế ghi nhớ những thông tin cần thiết xuyên suốt chuỗi thời gian của LSTM để có thể xem xét được sự tác động của giá những ngày trước đó đến giá hiện tại. Tuy nhiên vào những ngày cuối của tập test, kết quả dự đoán có xu hướng thấp hơn thực tế, nhất là với những ngày mà giá cổ phiếu lên cao nhanh chóng. Tìm hiểu về lý do gây nên vấn đề này, nhóm nhận thấy rằng dữ liệu cổ phiếu là dữ liệu chuỗi thời gian phức tạp và phụ thuộc rất nhiều yếu tố chứ không chỉ chịu sự tác động từ các loại cổ phiếu khác, với cổ phiếu mà nhóm lựa chọn để dự đoán là STB thì giá của cổ phiếu này có xu hướng tăng nhanh vào những năm gần đây, việc huấn luyện và kiểm thử trên các dữ liệu quá khứ có thể không kịp nắm bắt được xu hướng tăng của giá. Bên cạnh đó, khi xem xét với các loại dữ liệu chuỗi thời gian khác như chất lượng không khí thì gần như dữ liệu chứng khoán không có tính chu kỳ, việc này gây ra rất nhiều khó khăn trong dự đoán.

Như vậy, qua những thí nghiệm mà nhóm đã trình bày ở trên cũng như những nhận xét chung mà nhóm đã rút ra, đã giúp cho nhóm hiểu rõ hơn về bài toán cũng như việc chọn lựa mô hình với các tham số phù hợp để đưa vào giải quyết bài toán của nhóm.

Chương 6

Kết luận

Trong bài báo cáo này, nhóm chúng em đã trình bày về các bước tiếp cận cho bài toán dự đoán giá cổ phiếu. Đầu tiên, nhóm đã tiến hành phân tích tính tương quan trong thị trường nói chung và trong dữ liệu nói riêng của các loại cổ phiếu khác nhau, từ đó nhóm đã phân nhóm các loại cổ phiếu theo tính tương quan trên để có thể dự đưa ra dự đoán tốt cho cổ phiếu mà nhóm đã lựa chọn. Việc lựa chọn nhiều loại cổ phiếu để tiến hành dự đoán là khá cần thiết bởi bản thân mỗi loại cổ phiếu đều chịu rất nhiều tác động từ thị trường, một trong số đó chính là giá của các loại cổ phiếu khác. Về mô hình dự đoán mà nhóm đã sử dụng, được đề xuất dựa trên một bài báo đã có trước đó, kiến trúc của mô hình tuy chỉ là các mô hình cổ điển nhưng phù hợp với bài toán mà nhóm giải quyết. Kiến trúc được nhóm sử dụng bao gồm các khối LSTM để trích xuất thông tin từ từng chuỗi giá cổ phiếu, sau đó là các lớp tuyến tính để từ thông tin trích xuất của các cổ phiếu có tính tương quan cao đưa ra dự đoán cho cổ phiếu cần dự đoán. Nhóm cũng đã tiến hành việc thử nghiệm tính toán hiệu năng của mô hình dự đoán và đưa ra kết quả ở Chương 5.

Qua việc tìm hiểu và thực hiện bài tập lớn cho môn học,nhóm chúng em đã có thêm nhiều kinh nghiệm thực tế trong ứng dụng các phương pháp học máy để giải quyết các bài toán thực tiễn nói chung. Trong quá trình thực hiện bài tập lớn vẫn còn nhiều sai sót, nhóm chúng em rất mong nhận được lời nhận xét để hoàn thiện hơn các dự án mà chúng em sẽ tham gia về sau.

Tài liệu tham khảo

1. Long short-term memory-Fully connected (LSTM-FC) neural network for PM2. 5 concentration prediction. Zhao, Jiachen and Deng, Fang and Cai, Yeyun and Chen, Jie. Chemosphere 2019

2. Long short-term memory. Hochreiter, Sepp and Schmidhuber, Jürgen. Neural computation 1997

3. Stock price prediction by using hybrid sequential generative adversarial networks. He, Bate and Kita, Eisuke. 2020 International Conference on Data Mining Workshops (ICDMW).

4. Stock Index Prediction Based on Time Series Decomposition and Hybrid Model. Pin Lv , Qinjuan Wu, Jia Xu. Entropy 2022

5. Online: Stock Market Price Trend Prediction Using Time Series Forecasting)

6. Online: https://neptune.ai/blog/predicting-stock-prices-using-machine-learning