ĐẠI HỌC BÁCH KHOA HÀ NỘI
KHOA TOÁN TIN

ĐỒ ÁN II

DỰ BÁO GIÁ CHỨNG KHOÁN DỰA TRÊN MÔ HÌNH RNN VÀ LSTM

LÊ NGỌC HÀ
ha.ln216922@sis.hust.edu.vn

Ngành Hệ thống thông tin quản lý

Giảng viên hướng dẫn: PGS.TS. Nguyễn Đình Hân _____

Chữ ký GVHD

Bộ môn: Toán Tin

Khoa: Toán Tin

HÀ NỘI, 12/2024

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

1. **Mục tiêu và nội dung của đồ án**:
   (a) Mục tiêu:
   (b) Nội dung:

2. **Kết quả đạt được**:
   (a)
   (b)

3. **Ý thức làm việc của sinh viên**:

Hà Nội, ngày 29 tháng 10 năm 2024
Giảng viên hướng dẫn

PGS.TS. Nguyễn Đình Hân

LỜI CẢM ƠN

"Chúng ta cần tìm thời điểm thích hợp để dừng lại và cảm ơn những người đã tạo nên sự khác biệt cho cuộc đời mình."

Lời đầu tiên của đồ án này, em muốn gửi lời cảm ơn và lòng biết ơn chân thành tới thầy **PGS.TS. Nguyễn Đình Hân**. Thầy là người trực tiếp hướng dẫn và hỗ trợ em trong suốt quá trình thực hiện đồ án. Không những là người thầy chỉ bảo em về kiến thức, về định hướng phát triển mà thầy còn là người truyền cảm hứng cho em rất nhiều. Sau một thời gian có cơ hội được làm việc cùng thầy, ngoài việc học hỏi được những kiến thức chuyên ngành, những kinh nghiệm về ngành học thì sự tận tâm, cách thầy quan tâm đến sinh viên thật sự khiến em cảm thấy vô cùng thân thuộc, gần gũi và đáng để học hỏi. Em chúc thầy thật nhiều sức khoẻ và những điều tuyệt vời nhất sẽ luôn đến với thầy.

Đồng thời, em cũng muốn gửi lời cảm ơn đến Khoa Toán Tin - Đại học Bách khoa Hà Nội, nơi đã tạo điều kiện cho em được học tập và phát triển bản thân trong một môi trường đầy năng động và thoả sức sáng tạo.

Cuối cùng, con cũng muốn bày tỏ sự biết ơn tới gia đình, bạn bè, đặc biệt là bố và mẹ đã không quản vất vả, cực nhọc để lo cho con được đầy đủ nhất, để con không cảm thấy thiệt thòi so với các bạn đồng trang lứa. Mọi người vừa là chỗ dựa của con, cũng là động lực to lớn để con cố gắng hơn mỗi ngày trên con đường học tập, tìm kiếm và chinh phục tri thức.

Do kiến thức và các kỹ năng của bản thân vẫn còn hạn hẹp nên đồ án của em không tránh khỏi những thiếu sót và sai sót. Bởi vậy, em rất mong nhận được những ý kiến đóng góp quý báu từ thầy cô để đồ án được hoàn thiện hơn.

Em xin chân thành cảm ơn!

TÓM TẮT ĐỒ ÁN

- Chương 1:
- Chương 2:
- Chương 3:
- Chương 4:

Hà Nội, ngày 06 tháng 12 năm 2024
Sinh viên thực hiện

Lê Ngọc Hà

Mục lục

Danh sách từ viết tắt

CSS Cascading Style Sheets

Danh sách hình vẽ

Danh sách bảng

Mở đầu

1. Lý do chọn đề tài

Trong bối cảnh phát triển mạnh mẽ của thị trường tài chính, dự báo giá chứng khoán trở thành nhu cầu cấp thiết không chỉ đối với các nhà đầu tư mà còn cho các doanh nghiệp và cơ quan quản lý. Biến động giá chứng khoán chịu ảnh hưởng bởi nhiều yếu tố phức tạp như các chỉ số kinh tế vĩ mô, tâm lý thị trường và yếu tố công ty cụ thể.

Việc sử dụng các mô hình hồi quy đa biến giúp khai thác và phân tích tác động của nhiều yếu tố này một cách hệ thống và hiệu quả hơn, từ đó cung cấp các dự báo chính xác, hỗ trợ đưa ra các quyết định đầu tư có căn cứ. Chính vì vậy, đề tài "Dự báo giá chứng khoán dựa trên mô hình hồi quy đa biến LSTM và RNN" nhằm nghiên cứu và ứng dụng các phương pháp phân tích dữ liệu để góp phần giải quyết bài toán phức tạp này.

2. Đối tượng và phạm vi nghiên cứu

Đề tài tập trung vào các công cụ và kỹ thuật dự báo giá chứng khoán của một số mã cổ phiếu được niêm yết trên thị trường chứng khoán Việt Nam.

- Đối tượng nghiên cứu: Các yếu tố tác động đến giá cổ phiếu, chẳng hạn như chỉ số lạm phát, tỷ giá, lãi suất cùng các yếu tố nội tại doanh nghiệp (lợi nhuận, doanh thu, giá trị tài sản).
- Phạm vi nghiên cứu: Giới hạn trong việc xây dựng và đánh giá mô hình hồi quy đa biến, sử dụng dữ liệu thực tế trong một khoảng thời gian nhất định để kiểm chứng tính chính xác và khả năng áp dụng của mô hình vào dự báo giá cổ phiếu.

3. Ý nghĩa khoa học và thực tiễn của đề tài

Đề tài có ý nghĩa khoa học khi nghiên cứu ứng dụng của mô hình hồi quy đa biến, một trong những kỹ thuật phân tích dữ liệu phổ biến trong dự báo vào lĩnh vực tài chính, góp phần mở rộng ứng dụng của thống kê và phân tích dữ liệu vào thị trường chứng khoán.

Về mặt thực tiễn, kết quả của đề tài giúp các doanh nghiệp, nhà đầu tư hiểu rõ hơn mối liên hệ giữa các yếu tố ảnh hưởng đến giá cổ phiếu, từ đó hỗ trợ trong việc đưa ra các quyết định đầu tư có cơ sở. Kết quả nghiên cứu cũng có thể là tài liệu tham khảo hữu ích cho các doanh nghiệp tài chính, các tổ chức đầu tư và các nhà nghiên cứu quan tâm đến dự báo tài chính.

Chương 1

Bài toán về dự báo giá chứng khoán

1.1 Các loại giá chứng khoán phổ biến

Năm 1602, thị trường chứng khoán đầu tiên trên thế giới được thành lập tại Amsterdam, Hà Lan với sự ra đời của Công ty Đông Ấn Hà Lan (Dutch East India Company). Đây cũng là công ty đầu tiên phát hành cổ phiếu ra công chúng. Giá cổ phiếu lúc này được xác định bởi nhu cầu của các nhà đầu tư mua và bán cổ phần, đặt nền móng cho khái niệm giá chứng khoán sau này.

Giá chứng khoán là giá trị mà tại đó một loại chứng khoán (cổ phiếu, trái phiếu hoặc các công cụ tài chính khác) được giao dịch trên thị trường. Giá này thường được xác định bởi cung và cầu của thị trường, dựa trên các yếu tố như hiệu quả hoạt động của công ty, điều kiện kinh tế vĩ mô, tâm lý thị trường...

Giá chứng khoán có một số loại quan trọng mà bất kỳ ai nghiên cứu về chứng khoán cũng cần biết:

- Giá mở cửa: Đây là giá giao dịch đầu tiên của chứng khoán trong phiên giao dịch, nó được hình thành dựa trên cung và cầu tại thời điểm thị trường mở cửa. => Đóng vai trò quan trọng vì giá mở cửa phản ánh những biến động về tin tức, sự kiện ngoài giờ giao dịch.
- Giá đóng cửa: Đây là giá giao dịch cuối cùng của chứng khoán trong phiên giao dịch, thường được sử dụng để tham chiếu cho phiên giao dịch tiếp theo.
- Giá IPO (Initial Public Offering Price): Đây là giá cổ phiếu lần đầu ra mắt công chúng, thường được xác định bởi các tổ chức bảo lãnh phát hành dựa trên giá trị của doanh nghiệp và kỳ vọng từ thị trường.

- Giá tham chiếu: Đây là giá cơ sở để tính toán biên độ dao động giá trong ngày giao dịch đó. Ở thị trường chứng khoán Việt Nam, giá tham chiếu thường là giá đóng cửa của ngày giao dịch trước, dùng để xác định giá trần và giá sàn.
- Giá trần và giá sàn: Lần lượt là mức giá cao nhất và thấp nhất mà chứng khoán có thể được giao dịch trong một ngày.
- Giá khớp lệnh: Mức giá mà tại đó một lệnh mua và một lệnh bán được khớp trên thị trường.
- Giá cao nhất và giá thấp nhất: Lần lượt là mức giá cao nhất và thấp nhất đạt được trong một phiên.
- Giá trung bình: Mức giá trung bình của tất cả các giao dịch được thực hiện trong một ngày giao dịch.
- Giá thị trường: Giá hiện tại của chứng khoán được giao dịch trên thị trường, phản ánh giá trị kỳ vọng của nhà đầu tư đối với chứng khoán tại thời điểm đó.
- Giá danh nghĩa: Mệnh giá của cổ phiếu được ghi trên giấy chứng nhận cổ phiếu. Ở Việt Nam, giá danh nghĩa phổ biến là 10.000 VNĐ/cổ phiếu.

Có thể thấy, mỗi loại giá chứng khoán cung cấp một góc nhìn cụ thể về diễn biến giao dịch, hỗ trợ nhà đầu tư trong việc phân tích và đưa ra quyết định phù hợp. Việc hiểu rõ từng loại giá chứng khoán sẽ giúp chúng ta tăng cường khả năng quản lý rủi ro và tối ưu hóa lợi nhuận.

1.2 Các yếu tố chi phối giá chứng khoán

Giá chứng khoán chịu sự chi phối bởi nhiều yếu tố bao gồm tình hình nội bộ của doanh nghiệp, yếu tố kinh tế vĩ mô, những tác động và sự kiện trên thế giới. Những yếu tố này không chỉ mang tính cục bộ mà còn thể hiện mối quan hệ mật thiết giữa chính trị, kinh tế và thị trường tài chính toàn cầu.

Đầu tiên, giá chứng khoán bị chi phối bởi tình hình nội bộ của doanh nghiệp. Mỗi doanh nghiệp sẽ có cách vận hành, tổ chức khác nhau, chính vì đặc điểm riêng biệt này sẽ là yếu tố quyết định tới việc xác định giá trị cổ phiếu:

- Các công ty sở hữu tài sản lớn thường có tính cạnh tranh cao hơn trên thị trường chứng khoán, điều này giúp góp phần nâng cao giá chứng khoán.

- Doanh thu và lợi nhuận càng cao thì giá cổ phiếu sẽ có xu hướng tăng nhờ lấy được niềm tin của các nhà đầu tư vào sự phát triển bền vững của doanh nghiệp.
- Các công ty chi trả cổ tức hấp dẫn giúp tăng sự thu hút trên thị trường, qua đó giá cổ phiếu cũng được đẩy lên cao hơn.
- Mức thu nhập trên mỗi cổ phần (EPS) càng cao sẽ là minh chứng cho hoạt động tiềm năng của doanh nghiệp đó, tạo ra tác động tích cực đến giá cổ phiếu.

Tiếp theo, yếu tố thứ hai chi phối giá chứng khoán là yếu tố kinh tế vĩ mô. Những biến động của nền kinh tế có tác động không hề nhỏ đến toàn bộ thị trường chứng khoán nói chung và giá chứng khoán nói riêng. Tiêu biểu có thể kể đến như:

- Việc tăng trưởng GDP mạnh mẽ giúp doanh nghiệp đạt được lợi nhuận cao từ đó tạo điều kiện thuận lợi cho giá cổ phiếu tăng theo.
- Tình trạng lạm phát cao có thể làm giảm sức mua và giá trị thực của doanh nghiệp, dẫn đến tác động tiêu cực lên giá cổ phiếu.
- Lãi suất vay vốn cao khiến chi phí tài chính tăng, làm giảm lợi nhuận và kéo giá cổ phiếu đi xuống.
- Các chính sách tiền tệ ảnh hưởng trực tiếp đến thanh khoản và lãi suất, tạo ra những thay đổi lớn đối với thị trường chứng khoán.

Cuối cùng, những tác động trên thế giới và sự kiện mang tính toàn cầu cũng là yếu tố chi phối giá chứng khoán trong nước. Khi các giai đoạn tăng trưởng hoặc suy thoái kinh tế trên thế giới có thể gây ảnh hưởng trực tiếp đến các thị trường trong nước. Bên cạnh đó, biến động giá dầu cũng là một yếu tố quan trọng, bởi giá dầu thay đổi sẽ tác động đến chi phí sản xuất và lợi nhuận của nhiều ngành công nghiệp, đặc biệt là các doanh nghiệp phụ thuộc vào nguyên liệu đầu vào. Một yếu tố toàn cầu khác là lợi suất trái phiếu quốc tế có thể ảnh hưởng đến dòng vốn đầu tư và gián tiếp tác động đến giá cổ phiếu. Ngoài ra, giá cổ phiếu còn bị chi phối bởi các yếu tố khác như giá trị USD và vàng, bởi những biến động của chúng thường ảnh hưởng đến tâm lý nhà đầu tư và quyết định đầu tư. Chính sách của ngân hàng trung ương với các biện pháp như điều chỉnh lãi suất hoặc cung ứng tiền tệ có khả năng định hướng xu hướng thị trường chứng khoán.

Việc hiểu rõ và phân tích tỉ mỉ về các yếu tố ảnh hưởng này sẽ giúp nhà đầu tư,

doanh nghiệp đưa ra quyết định hiệu quả hơn, nắm bắt được tình hình giá chứng khoán lên xuống ra sao, đồng thời tránh được các rủi ro không mong muốn.

1.3 Các phương pháp dự báo giá chứng khoán

a) Phân tích cơ bản

Phương pháp phân tích cơ bản dựa trên việc phân tích và đánh giá cổ phiếu thông qua các yếu tố kinh tế và tài chính bao gồm: Xem xét khả năng sinh lời thông qua doanh thu và lợi nhuận, sử dụng các chỉ số tài chính như EPS, P/E, P/B, vị thế cạnh tranh của công ty và các yêu tố vĩ mô (lãi suất, GDP, tỷ lệ thất nghiệp...).

b) Phân tích kỹ thuật

Phương pháp phân tích kỹ thuật tập trung vào hành vi giá, giả định rằng các yếu tố về giá và khối lượng giao dịch sẽ có ảnh hưởng trực tiếp đến giá chứng khoán. Phương pháp phân tích kỹ thuật thường sử dụng các công cụ phân tích bao gồm:

- Biểu đồ giá giúp hiển thị các biến động của giá chứng khoán theo thời gian để giúp các nhà đầu tư, doanh nghiệp có cái nhìn tổng quan nhất.
- Đường trung bình để tính giá chứng khoán trung bình trong một khoảng thời gian cụ thể (theo giờ, theo ngày, theo tháng...) để từ đó xác định xu hướng giá trong tương lai.
- Các chỉ số kỹ thuật như chỉ số RSI (Relative Strength Index) để đo lường mức độ dao động của giá, chỉ số MACD (Moving Average Convergence Diverge) giúp cung cấp các biến động của thị trường giá.

c) Mô hình chuỗi thời gian

Mô hình chuỗi thời gian (Time Series Models) là một trong số những phương pháp phổ biến để dự đoán giá chứng khoán dựa trên dữ liệu lịch sử. Một số mô hình tiêu biểu có thể kể đến như ARIMA, SARIMA, ARCH, GARCH... Các mô hình này phân tích mối quan hệ giữa các giá trị dữ liệu thông qua thời gian để từ đó dự đoán xu hướng, biến động giá trong tương lai.

d) Phương pháp học máy

Phương pháp học máy (Machine Learning) thuộc một nhánh của trí tuệ nhân tạo tập trung vào việc phát triển các thuật toán và mô hình để máy tính có thể tự học từ dữ liệu và cải thiện hiệu suất. Thay vì được lập trình theo các quy tắc cụ thể,

phương pháp học máy sẽ học từ dữ liệu và điều chỉnh mô hình để đạt được kết quả tốt nhất. Một số thuật toán phổ biến có thể kể đến như hồi quy tuyến tính, cây quyết định, SVM, phân cụm K-means, PCA, LSTM, RNN...

Ngoài ra, chúng ta có thể kết hợp nhiều mô hình khác nhau để đưa ra dự đoán tối ưu và chính xác nhất. Trong bài báo cáo này, em đã sử dụng mô hình RNN và LSTM để dự báo giá chứng khoán để giúp các công ty, doanh nghiệp có cái nhìn tổng quan và đưa ra quyết định dễ dàng hơn.

1.4 Vai trò của việc dự báo giá chứng khoán

Dự báo giá chứng khoán giúp nhà đầu tư có thể dự đoán sự biến động của thị trường và các cổ phiếu, từ đó đưa ra quyết định đầu tư hợp lý nhằm tối đa hóa lợi nhuận. Đồng thời, dự báo giúp phát hiện sớm các dấu hiệu của sự thay đổi trong xu hướng thị trường, giúp giảm thiểu rủi ro cho nhà đầu tư.

Việc dự báo giá chứng khoán cung cấp thông tin hữu ích để xây dựng và điều chỉnh danh mục đầu tư, giúp nhà đầu tư phân bổ vốn một cách hợp lý giữa các loại tài sản, giảm thiểu sự phụ thuộc vào một ngành hoặc cổ phiếu duy nhất.

Dự báo giá chứng khoán cũng giúp nhà đầu tư xác định thời điểm mua vào hoặc bán ra cổ phiếu, từ đó tối ưu hóa lợi nhuận. Các chiến lược giao dịch theo xu hướng hoặc theo phân tích kỹ thuật có thể được áp dụng dựa trên dự báo chính xác về giá cổ phiếu.

1.5 Tổng quan về tình hình giá chứng khoán

1.5.1 Tình hình giá chứng khoán trên thế giới

Tính đến tháng 10/2024, giá chứng khoán toàn cầu ghi nhận kết quả tích cực nhất trong 5 năm qua, với chỉ số MSCI toàn cầu tăng 7,7%, mức tăng cao nhất kể từ năm 2019. Sự tăng trưởng này chủ yếu nhờ vào nền kinh tế Mỹ và sự bùng nổ của công nghệ Trí tuệ nhân tạo (AI). Đặc biệt, sự tăng trưởng mạnh mẽ của Nvidia, với giá trị vốn hóa thị trường tăng hơn 1.000 tỷ USD trong ba tháng đầu năm đã thúc đẩy đà đi lên của giá chứng khoán.

Mặc dù lạm phát ở Mỹ tăng cao bất ngờ trong tháng 1/2024 và tháng 2/2024 nhưng

giá chứng khoán vẫn tiếp tục duy trì đà tăng. Ngoài ra, các thị trường chứng khoán ở châu Âu và châu Á cũng có sự tăng trưởng vượt trội. Nhật Bản là nước dẫn đầu nhờ vào sự phục hồi mạnh mẽ của ngành công nghệ và niềm tin vào nền kinh tế trong nước.

Giá chứng khoán toàn cầu có sự tăng trưởng ấn tượng nhưng nếu suy thoái kinh tế diễn ra hoặc tỷ lệ thất nghiệp ở Mỹ tăng đột ngột thì những mức tăng trưởng này có thể bị ảnh hưởng. Mặc dù vậy, mức giá chứng khoán hiện tại có thể kéo dài đến năm 2029 hoặc thậm chí 2033 nếu không có yếu tố bất ngờ nào. [1]

1.5.2 Tình hình giá chứng khoán tại Việt Nam

Xét đến tháng 10/2024, giá chứng khoán Việt Nam ghi nhận nhiều tín hiệu tích cực nhưng vẫn còn đối mặt với những thách thức đáng kể từ cả yếu tố trong nước và tác động của kinh tế toàn cầu. Trong tháng 9/2024, chỉ số VN-Index (chỉ số đại diện cho thị trường chứng khoán Việt Nam) có nhiều biến động đáng chú ý. Sau khi chạm mốc 1.300 điểm, VN-Index không thể duy trì được mức này nhưng vẫn kết thúc tháng với một đợt tăng nhẹ.

Với sự tăng mạnh trở lại của tỷ giá USD/VND, áp lực thoái vốn từ các nhà đầu tư nước ngoài gia tăng, ảnh hưởng đến sự ổn định của giá chứng khoán. Tuy nhiên, nhờ các chính sách điều hành kinh tế vĩ mô ổn định giúp nguy cơ giảm sâu hơn của thị trường được đánh giá là không cao. Nhìn chung, thị trường chứng khoán Việt Nam nói chung và giá chứng khoán tại Việt Nam nói riêng vẫn còn rất nhiều tiềm năng để các nhà đầu tư, doanh nghiệp có thể khai thác và hy vọng trong tương lai.

Chương 2

Cơ sở lý thuyết

2.1 Tìm hiểu về chuỗi thời gian

2.1.1 Khái niệm chuỗi thời gian

Định nghĩa 2.1. Mô hình chuỗi thời gian cho dữ liệu được quan sát $\{x_t\}$ là một chuỗi các biến ngẫu nhiên $\{X_t\}$, hay còn gọi là quá trình ngẫu nhiên với phân phối xác suất (hoặc các giá trị trung bình và tương quan giữa các biến) xác định nhận $\{x_t\}$ là giá trị đại diện tại thời điểm quan sát.

Chuỗi thời gian được hiểu đơn giản là một tập hợp các giá trị quan sát $\{x_t\}$, ở đó mỗi quan sát x_t được ghi lại tại thời gian xác định $t \in T$, với T là miền giá trị thực nào đó. Nếu T là tập rời rạc (thường là tập con của tập hợp số tự nhiên) thì ta có chuỗi thời gian rời rạc. Khi T là một khoảng hay hợp của nhiều khoảng nào đó trên $\mathbb{R}$ thì ta nói $\{x_t\}_{t\in T}$ là chuỗi thời gian liên tục.

Ví dụ 2.1. Chuỗi thời gian rời rạc: Giá cổ phiếu của một công ty được ghi lại vào cuối mỗi ngày giao dịch, số lượng khách hàng đến một cửa hàng được đo đếm vào cuối mỗi tháng...

Ví dụ 2.2. Chuỗi thời gian liên tục: Nhiệt độ được đo liên tục trong một khoảng thời gian hàng giây hoặc hàng phút trong suốt một ngày, tín hiệu điện tim được ghi lại liên tục trong một khoảng thời gian để theo dõi hoạt động của tim...

2.1.2 Đặc điểm chuỗi thời gian

Dữ liệu ở dạng chuỗi thời gian thường có một số đặc điểm sau đây, mỗi đặc điểm người ta còn xem là một thành phần của chuỗi.

- Tính xu hướng (Trend): Tính chất thể hiện xu hướng thay đổi tăng hay giảm của dữ liệu trong một khoảng thời gian dài. Đây là một đặc trưng khá phổ biến của nhiều dữ liệu chuỗi thời gian.
- Tính mùa vụ (Seasonality): Tính chất thể hiện tính chất thay đổi tăng hoặc giảm, lặp đi lặp lại một cách đều đặn của chuỗi dữ liệu trong một khoảng thời gian. Các qui luật mùa vụ sẽ lặp lại theo một số chu kỳ phổ biến chẳng hạn như: năm (như GDP, kim ngạch xuất nhập khẩu), tháng (chuỗi liên quan tới doanh thu, doanh số, du lịch, dịch vụ), ngày (chuỗi liên quan tới qui luật mua sắm, tiêu dùng, vui chơi giải trí).
- Tính chu kỳ (Cyclical): Tính chất thể hiện sự thay đổi lên xuống lặp lại, hoặc những thay đổi định kỳ, có thể kéo dài trong nhiều năm và chuyển từ giai đoạn này sang giai đoạn khác.
- Tính chất bất quy tắc (Irregular): Tính chất thể hiện sự thay đổi bất thường của dữ liệu, xảy ra hoàn toàn ngẫu nhiên và khó có thể dự đoán. Sự bất thường này có thể đến do sai số trong đo lường dữ liệu, những sự kiện phát sinh làm ảnh hưởng tới giá trị của dữ liệu tại những thời điểm quan sát.

2.2 Các chỉ số đánh giá mô hình

2.2.1 Mean Squared Error (MSE)

Mean Squared Error (MSE) là một trong những số liệu phổ biến nhất được sử dụng cho các bài toán hồi quy. Nó tìm thấy sai số bình phương trung bình giữa các giá trị được dự đoán và thực tế. MSE là thước đo chất lượng của một công cụ ước tính, nó luôn không âm và các giá trị càng gần 0 càng tốt. Ta có công thức MSE cho bài toán dự báo giá chứng khoán như sau:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2 \tag{2.1}$$

Trong đó:

- N: Số lượng mẫu.
- x_i: Giá trị thực tế của mẫu thứ i.
- $\hat{x}_i$: Giá trị dự đoán tương ứng với mẫu thứ i.

2.2.2 Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) tương tự MSE và được tính bằng cách lấy căn bậc hai của MSE:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2} \tag{2.2}$$

2.2.3 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) là sai số tuyệt đối trung bình, một chỉ số đánh giá hiệu quả của mô hình dự đoán trong học máy (machine learning) và thống kê. Chỉ số này đo lường độ lớn trung bình của sai số dự đoán, tức là mức độ chênh lệch giữa các giá trị thực tế và giá trị dự đoán mà không xét dấu. Ta có công thức tính MAE như sau:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{x_i - \hat{x}_i}{x_i}\right| \tag{2.3}$$

2.2.4 Mean Absolute Percentage Error (MAPE)

MAPE là thước đo thường được sử dụng để đánh giá hiệu suất của mô hình hồi quy. Nó đo chênh lệch tỷ lệ phần trăm trung bình giữa giá trị dự đoán và giá trị thực tế. Với n là cỡ của dữ liệu, MAPE được xác định như sau:

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{x_i - \hat{x}_i}{x_i}\right| \times 100\% \tag{2.4}$$

2.2.5 Coefficient Of Determination ($\mathbf{R}^2$)

Coefficient of Determination (R^2) là một chỉ số đánh giá hiệu suất của mô hình hồi quy. Nó đo lường mức độ phù hợp của mô hình dự đoán với dữ liệu thực tế, cho biết tỷ lệ phương sai của biến phụ thuộc được giải thích bởi các biến độc lập trong mô hình. Công thức tính R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{2.5}$$

Trong đó: $\bar{x}$ là giá trị trung bình của các giá trị thực tế.

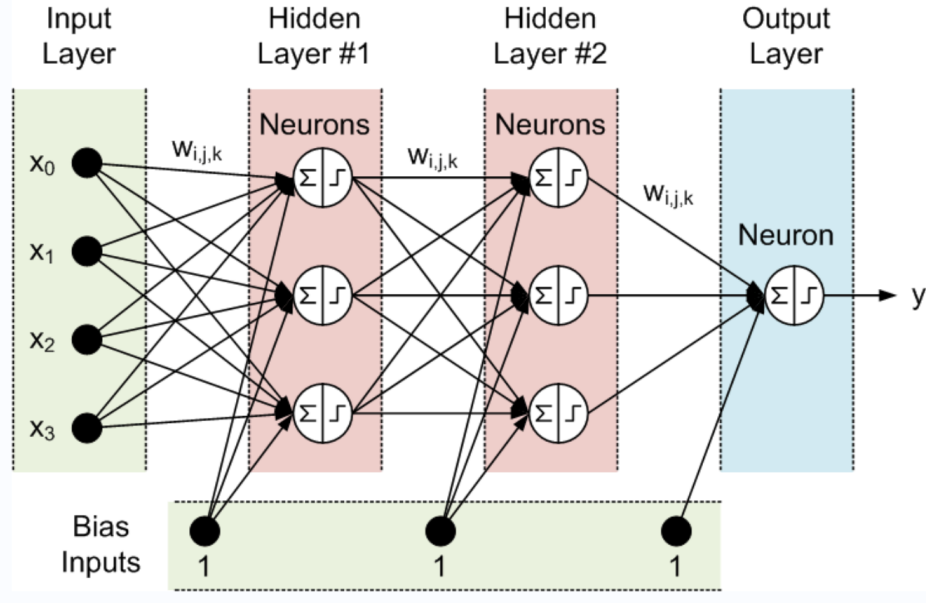
2.3 Các mô hình dựa trên chuỗi thời gian

2.3.1 Mô hình Neural Network

Neural Network (NN) về bản chất là tập hợp của nhiều nơ-ron được liên kết với nhau theo một trật tư, quy luật nhất định. Mạng nơ-ron là một mô hình tính toán được lấy cảm hứng từ cách hoạt động của não bộ con người, được tạo thành từ các nơ-ron liên kết với nhau thành từng lớp.

Mỗi nơ-ron nhận tín hiệu đầu vào từ các nơ ron khác, thực hiện tính toán rồi truyền kết quả đầu ra cho các nơ-ron tiếp theo. Nhờ vào việc điều chỉnh các nơ-ron liên kết trong quá trình huấn luyện mà mạng nơ-ron có thể học hỏi và giải quyết các vấn đề phức tạp trong thực tế như nhận diện hình ảnh, dự đoán ngôn ngữ, ký hiệu, dự đoán nhu cầu, xu hướng...

a) Cấu tạo của mô hình Neural Network



Hình 2.1: Cấu tạo của mô hình Neural Network

Cấu tạo của mô hình Neural Network được thể hiện trong hình (2.1) gồm các thành phần sau:

- Input Layer: Đây là layer đầu tiên chứa các giá trị đầu vào.
- Hidden Layer: Đây là layer ở giữa, layer này có thể gồm nhiều layer khác nhau hoặc không cần có. Mô hình càng có nhiều lớp ẩn thì càng phức tạp.

- Output Layer: Đây là layer cuối chứa các giá trị đầu ra.
  => Tổng số layer trong mô hình được quy ước là số layer - 1 (không bao gồm input layer).
- Node: Node là các hình tròn, mỗi node có hệ số bias b riêng. Mỗi node trong hidden layer và output layer sẽ liên kết với tất cả các node ở layer trước đó, mỗi liên kết có hệ số w riêng.

b) Thuật toán của mô hình Neural Network

Bước 1: Tại input layer sẽ chứa các giá trị x_i đầu vào tương ứng với một node.

Bước 2: Tính tổng Linear, đây là bước kết hợp thông tin từ các node ở layer trước giúp tổng hợp thông tin theo cách tuyến tính để chuẩn bị cho bước xử lý phi tuyến tiếp theo.

$$z_j^{(l)} = \sum_{i=1}^{h^{(l-1)}} w_{ji}^{(l)} a_i^{(l-1)} + b_j^{(l)} \tag{2.6}$$

Trong đó:

- $z_j^{(l)}$: Giá trị trước khi áp dụng activation function tại node thứ j, layer thứ l.
- $h^{(l-1)}$: Số lượng node trong layer trước đó $l-1$.
- $w_{ji}^{(l)}$: Trọng số weight dùng để nối từ node thứ i trong layer $l-1$ sang node thứ j trong layer l, mỗi kết nối giữa hai node có một w đặc trưng.
- $a_i^{(l-1)}$: Giá trị đầu ra từ node thứ i của layer $l-1$, tức là đầu ra sau khi áp dụng activation function ở layer trước.
- $b_j^{(l)}$: Hệ số điều chỉnh bias tại node thứ j trong layer l.

Bước 3: Áp dụng các hàm kích hoạt (Activation Function) để đưa tính phi tuyến lên tổng tuyến tính để tính đầu ra của node hiện tại, giúp mạng nơ-ron học các quan hệ phi tuyến phức tạp hơn đồng thời kiểm soát tín hiệu lan truyền và chuẩn hoá đầu ra vào khoảng giá trị cụ thể. Ở đây, đầu ra của hàm σ luôn nằm trong khoảng [0,1].

$$a_j^{(l)} = \sigma(z_j^{(l)}) \tag{2.7}$$

Trong đó:

- $a_j^{(l)}$: Giá trị đầu ra của node thứ j trong layer l sau khi áp dụng activation function.
- $\sigma(z_j^{(l)})$: Activation Function (Hàm kích hoạt) áp dụng lên giá trị $z_j^{(l)}$.

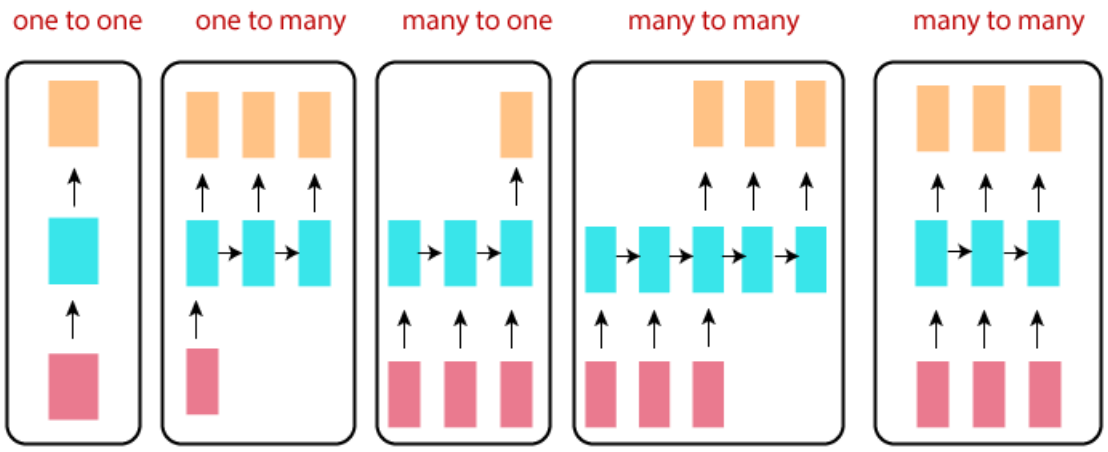
Bước 4: Output là giá trị dự đoán y của mô hình.

2.3.2 Mô hình Recurrent Neural Network

Recurrent Neural Network (RNN) được giới thiệu lần đầu tiên vào năm 1986 [2], đây là một loại mạng nơ-ron mà trong đó các kết nối giữa các node tạo thành một chu trình, cho phép thông tin được duy trì và xử lý qua các bước thời gian liên tiếp. RNN có thể xử lý dữ liệu dạng chuỗi hoặc tuần tự, khác biệt so với các mạng nơ-ron truyền thống [3]:

- Mỗi node trong mạng RNN không chỉ nhận thông tin từ đầu vào mà còn từ trạng thái ẩn được cập nhật liên tục qua các bước thời gian.
- Mô hình RNN giúp mạng nơ-ron có khả năng ghi nhớ ngữ cảnh của dữ liệu trước và từ đó có thể đưa ra các dự đoán về hiện tại và tương lai. Chính vì thế, RNN được sử dụng phù hợp với các bài toán như dự đoán chuỗi, phân loại thời gian thực, dịch máy...

a) Phân loại mô hình Recurrent Neural Network Phân loại mô hình Recurrent



Hình 2.2: Phân loại mô hình Recurrent Neural Network

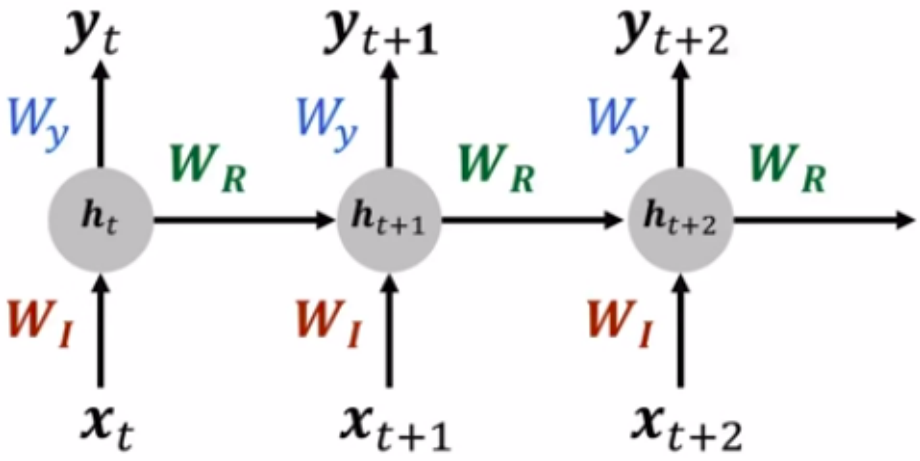
Neural Network được thể hiện trong hình (2.2) như sau:

- One - One: Đây là kiến trúc đơn giản nhất chỉ gồm một đầu vào và một đầu

ra, thường dùng để phân loại hình ảnh.

- One - Many: Một đầu vào sẽ ánh xạ đến nhiều đầu ra như từ một bức ảnh có thể tạo ra một chuỗi văn bản.
- Many - One: Nhiều đầu vào ánh xạ đến một đầu ra duy nhất, thường dùng trong các bài toán phân loại cảm xúc, dự đoán chuỗi thời gian...
- Many - Many: Nhiều đầu vào ánh xạ đến nhiều đầu ra nghĩa là số lượng input luôn bằng số lượng output, thường được dùng để dịch ngôn ngữ ký hiệu...

b) Thuật toán của mô hình Recurrent Neural Network:



Hình 2.3: Mô hình Recurrent Neural Network

Thuật toán của mô hình Recurrent Neural Network được mô tả như trong hình (2.3) như sau:

Bước 1: Input là các giá trị x_t.

Mô hình RNN nhận một chuỗi dữ liệu đầu vào theo từng bước thời gian, các giá trị đầu vào x_t, x_{t+1}, x_{t+2} sẽ tương ứng với thời điểm t, $t+1$, $t+2$ và được gọi là các timestep. Mỗi bước thời gian này có thể là một từ trong câu, một giá trị trong chuỗi thời gian, một tín hiệu âm thanh tại một thời điểm...

Bước 2: Trạng thái ẩn h_t.

Tại mỗi thời điểm t thì RNN đều có một trạng thái ẩn h_t đóng vai trò như một "bộ

nhớ" để lưu trữ thông tin từ các bước thời gian trước đó trong chuỗi. Trạng thái ẩn h_t sẽ được tính toán dựa trên hai yếu tố đó là thông tin từ bước thời gian trước h_{t-1} và thông tin đầu vào hiện tại x_t. Trạng thái ẩn sẽ được cập nhật qua một hàm phi tuyến tính để giúp mô hình học hỏi được các quan hệ phức tạp trong dữ liệu. Quá trình này được tính như sau:

$$h_t = f(W_r h_{t-1} + W_i x_t + b) \tag{2.8}$$

Trong đó:

- f: Hàm kích hoạt.
- W_r: Trọng số giữa h_{t-1} và h_t.
- W_i: Trọng số giữa h_t và x_t.
- b: Hệ số bias riêng.

Vì f là hàm kích hoạt chẳng hạn như σ, $tanh$ hoặc $ReLU$... tuỳ thuộc vào từng bài toán cụ thể. Từ đó, ta sẽ có công thức:

$$h_t = tanh(W_r h_{t-1} + W_i x_t) \tag{2.9}$$

Bước 3: Output là các giá trị y_t.

Đầu ra của mô hình tại mỗi thời điểm t sẽ được tính từ trạng thái ẩn h_t. Công thức tính đầu ra như sau:

$$y_t = W_y h_t \tag{2.10}$$

(Với W_y là trọng số giữa h_t và y_t).

Tóm lại, mỗi bước trong RNN không chỉ xử lý đầu vào hiện tại mà còn ghi nhớ thông tin từ quá khứ trước đó để đưa ra dự đoán cuối cùng sau khi chuỗi dữ liệu hoàn thành.

c) Vấn đề Loss Function

Khi làm việc với chuỗi thời gian dài, RNN thường sẽ gặp phải hai vấn đề đó là vanishing gradient và exploding gradient. Khi huấn luyện RNN, hàm mất mát cần phải được lan truyền ngược qua thời gian (BPTT) để tính toán gradient và cập nhật

trọng số. Tuy nhiên, khi chuỗi dữ liệu dài, gradient có thể trở nên rất nhỏ, khiến việc cập nhật trọng số trở nên kém hiệu quả, đặc biệt đối với các bước thời gian sớm hơn trong chuỗi. Từ đó khiến cho mô hình sẽ không học được tốt các thông tin quan trọng từ các bước thời gian đầu của chuỗi.

Ngược lại với vanishing gradient, exploding gradient xảy ra khi các gradient trở nên quá lớn trong quá trình lan truyền ngược. Điều này thường xảy ra khi các trọng số của mô hình quá lớn, khiến cho các cập nhật trọng số bị phóng đại, dẫn đến sự mất ổn định trong quá trình huấn luyện.

2.3.3 Mô hình Long Short Term Memory

Mô hình Long Short - Term Memory (LSTM) được công bố lần đầu tiên bởi hai tác giả Sepp Hochreiter và Jürgen Schmidhuber vào năm 1997 [4]. LSTM được giới thiệu như một kiến trúc mạng RNN đặc biệt để giải quyết vấn đề vanishing gradient trong việc huấn luyện các mạng nơ-ron truyền thống qua các chuỗi dài.

Hiểu đơn giản, Long Short - Term Memory là một loại mạng nơ-ron hồi quy có cấu trúc đặc biệt với các cell states và các cổng (gates) điều khiển thông tin, cho phép mô hình học cách chọn lọc thông tin quan trọng để lưu trữ trong trạng thái "bộ nhớ" dài hạn và bỏ qua thông tin không cần thiết.

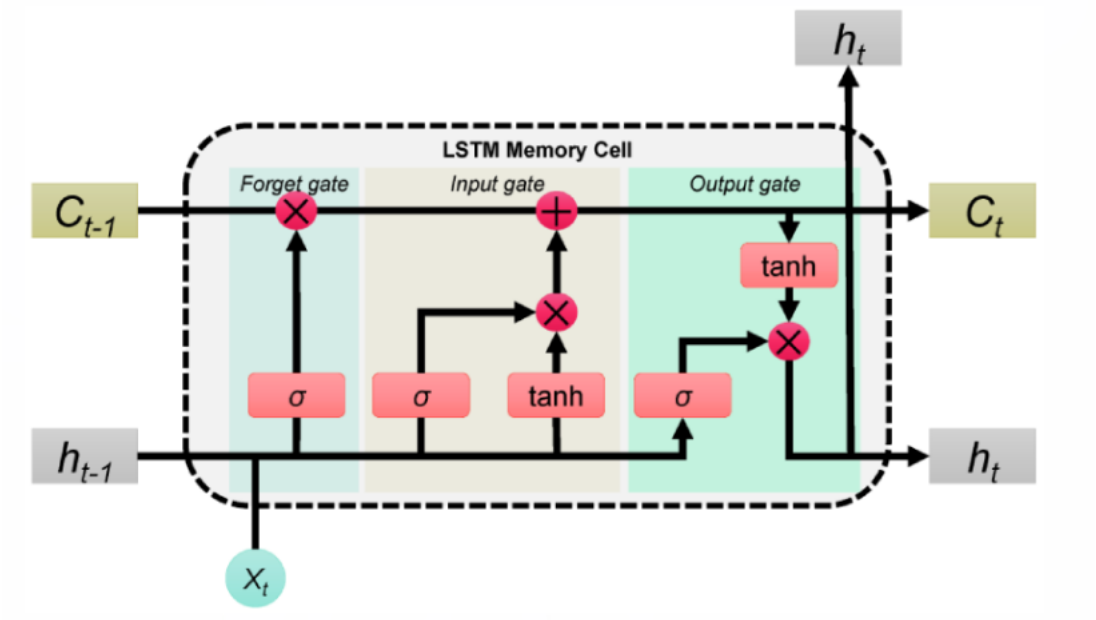
Ví dụ 2.3. Ta có chuỗi ký tự: "Tôi thích học nhưng hôm nay tôi bị ốm". Khi dùng LSTM, mô hình có thể bỏ qua thông tin "Tôi thích học" và chỉ tập trung vào vế phía sau "tôi bị ốm" nhằm dự đoán hành động tiếp theo như là nghỉ ngơi, uống thuốc, khám bệnh... *Note: Vẽ hình minh hoạ*

a) Cấu tạo của mô hình Long Short - Term Memory

Cấu tạo của mô hình Long Short - Term Memory được thể hiện như trong hình (2.4) như sau:

- Cell State: Kí hiệu c, thanh trạng thái bộ nhớ giúp lưu trữ thông tin quan trọng trong suốt thời gian dài kể cả thông tin của các timesteps ban đầu.
  => Khắc phục được nhược điểm Short - Term Memory.
- Hidden State: Kí hiệu h (có chức năng tương tự h trong mô hình RNN), đây là đầu ra chính tại mỗi bước thời gian được sử dụng trong các bước tính toán tiếp theo hoặc cho lớp đầu ra cuối cùng.

- Ba cổng chính: Forget Gate, Input Gate, Output Gate.

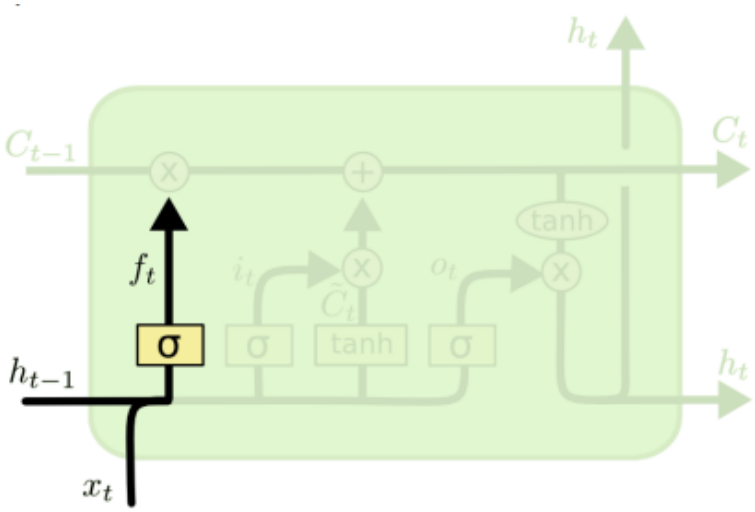


Hình 2.4: Mô hình Long Short - Term Memory

b) Thuật toán của mô hình Long Short - Term Memory

Bước 1: Input là các giá trị x_t, h_{t-1}.

Bước 2: Giá trị x_t, h_{t-1} sẽ được xử lý tại Forget Gate.



Hình 2.5: Forget Gate

Trước tiên, LSTM sẽ xác định giữ lại hay loại bỏ phần thông tin nào trong trạng thái ô hiện tại c_t. Điều này giúp LSTM tập trung vào những thông tin cần thiết cho

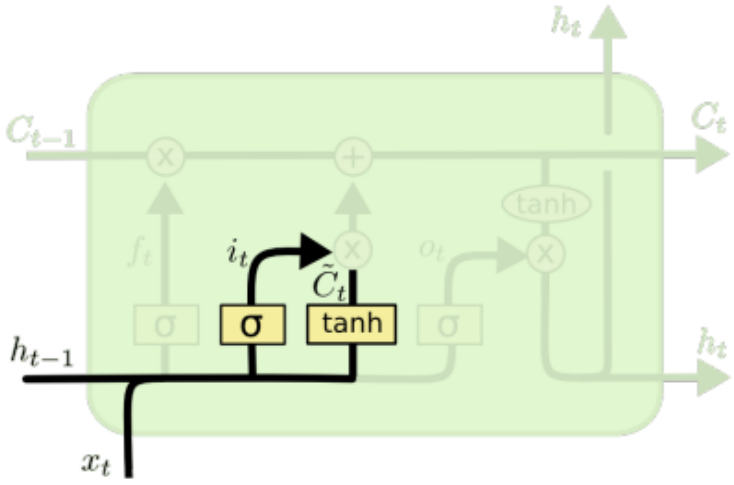
việc dự đoán tại thời điểm hiện tại và tương lai, đồng thời tránh lưu trữ thông tin không cần thiết, làm giảm hiệu quả mô hình. Tại đây sẽ sử dụng hàm σ (sigmoid) nhận đầu vào h_{t-1} và x_t để trả về các giá trị đầu ra nằm trong khoảng [0,1]. Khi giá trị gần tiến tới 0 nghĩa là thông tin sẽ bị loại bỏ còn giá trị gần bằng 1 tức là sẽ giữ lại hoàn toàn thông tin. Ta có công thức tính như sau:

$$f_t = \sigma(U_f * x_t + W_f * h_{t-1} + b_f) \tag{2.11}$$

Trong đó:

- f_t: Đầu ra của forget gate tại thời điểm t.
- σ: Hàm kích hoạt sigmoid.
- U_f: Trọng số liên quan đến đầu vào x_t.
- x_t: Đầu vào ở bước hiện tại.
- W_f: Trọng số của forget gate giúp quyết định hành động ở bước hiện tại.
- h_{t-1}: Trạng thái ẩn từ thời điểm trước đó $t-1$
- b_f: Hệ số bias của cổng quên.

Bước 2: Tại Input Gate.



Hình 2.6: Input Gate

Tại Input Gate sẽ quyết định thông tin nào cần được thêm vào trạng thái bộ nhớ c. Cổng này quyết định mức độ quan trọng của dữ liệu mới, đồng thời tạo ra một trạng thái bộ nhớ tiềm năng để kết hợp với thông tin cũ. Điều này giúp mô hình

liên tục cập nhật thông tin hữu ích.

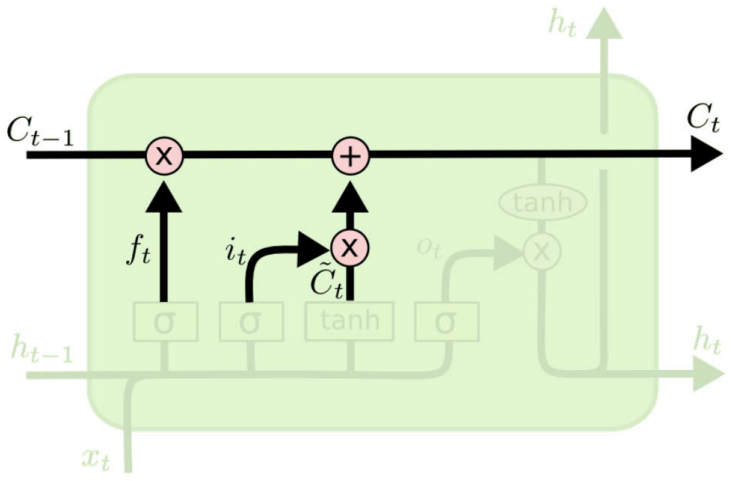
Đầu tiên, tại input gate sẽ sử dụng hàm σ (sigmoid) lần thứ hai để xác định mức độ thông tin mới sẽ được cập nhật. Một hàm *tanh* được dùng để tạo ra một vecto $\tilde{c}_t$, đại diện cho các giá trị trạng thái mới tiềm năng. Đầu ra của hàm sigmoid và hàm tanh được nhân với nhau để tạo thành thông tin cần được thêm vào cell state. Ta có công thức tính như sau:

$$i_t = \sigma(U_i * x_t + W_i * h_{t-1} + b_i) \tag{2.12}$$

$$\tilde{c}_t = \tanh(U_c * x_t + W_c * h_{t-1} + b_c) \tag{2.13}$$

Trong đó:

- i_t: Đầu ra của input gate tại thời điểm t.
- U_i: Trọng số giữa đầu vào x_t và input gate.
- W_i: Trọng số giữa trạng thái ẩn h_{t-1} và input gate.
- $\tilde{c}_t$: Giá trị được điều chỉnh bởi i_t trước khi thêm vào c_t.
- *tanh*: Hàm kích hoạt nén giá trị vào khoảng [-1,1].
- U_c: Trọng số giữa đầu vào x_t và $\tilde{c}_t$.
- W_c: Trọng số giữa trạng thái ấn h_{t-1} và $\tilde{c}_t$.
- b_i, b_c: Hệ số điều chỉnh của i_t, $\tilde{c}_t$.

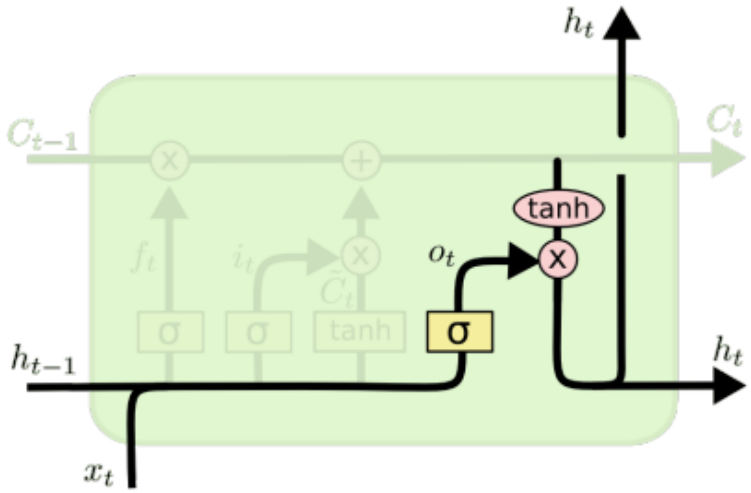


Hình 2.7: Input Gate

Sau đó, trạng thái thanh c_t được làm mới dựa trên thông tin từ forget input và input gate. Tại đây sẽ loại bỏ thông tin cũ bằng cách nhân trạng thái cũ c_{t-1} với đầu ra của forget input. Đồng thời, cập nhật thông tin mới với $\tilde{c}_t$ nhân với đầu ra của input gate. Cuối cùng, cộng tổng cả hai lại vào trạng thái đã được làm sạch. Công thức được viết như sau:

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{2.14}$$

Bước 3: Tại Output Gate



Hình 2.8: Output Gate

Output Gate điều chỉnh lượng thông tin được xuất ra ngoài y_t và truyền đến trạng thái tiếp theo h_t. Sử dụng hàm σ (sigmoid) để quyết định thông tin nào sẽ được đưa ra. Cell state c_t được đưa qua hàm *tanh* và nhân với đầu ra của sigmoid để tạo ra trạng thái ẩn h_t. Ta có công thức như sau:

$$o_t = \sigma(U_o x_t + W_o h_{t-1} + b_o) \tag{2.15}$$

$$h_t = o_t \tanh(c_t) \tag{2.16}$$

Trong đó:

- o_t: Đầu ra của output gate tại thời điểm t.
- U_o: Trọng số giữa đầu vào x_t và output gate.
- W_o: Trọng số giữa trạng thái ẩn h_t và output gate.

c) Mối quan hệ giữa short term memory và long term memory trong LSTM

Short-Term Memory (h_t) là trạng thái ẩn đại diện cho thông tin ngắn hạn và được tính dựa trên trạng thái Long-Term Memory (c_t). Điều này xảy ra vì $h_t = o_t \tanh(c_t)$ chứa thông tin tích lũy dài hạn đã được điều chỉnh bởi các cổng quên và đầu vào. Cổng quên giúp loại bỏ thông tin không còn hữu ích trong khi cổng đầu vào thêm thông tin mới x_t, đảm bảo c_t lưu giữ được những yếu tố quan trọng nhất qua thời gian. Sự kết hợp này tạo ra sự cân bằng giữa c_t mang ngữ cảnh dài hạn và ít bị ảnh hưởng bởi nhiễu, trong khi h_t phản ứng nhanh chóng với dữ liệu gần đây, giúp mô hình xử lý cả thông tin tổng quát và chi tiết. Điều này làm LSTM hiệu quả trong các chuỗi dữ liệu phức tạp, nơi cả quá khứ xa và gần đều quan trọng.

2.3.4 So sánh mô hình RNN và LSTM

Sự khác nhau giữa mô hình RNN và mô hình LSTM được thể hiện trong bảng (2.1) như sau:

Tiêu chí	Mô hình RNN	Mô hình LSTM
Cấu tạo	Đơn giản, mỗi bước thời gian chỉ có một trạng thái ẩn duy nhất	Phức tạp hơn với ba cổng (cổng quên, cổng đầu vào, cổng đầu ra) để điều khiển thông tin
Khả năng duy trì thông tin	Hạn chế trong việc duy trì thông tin qua nhiều bước thời gian	Có khả năng duy trì thông tin dài hạn, giúp xử lý chuỗi thời gian dài hiệu quả
Vấn đề gặp phải	Gặp vấn đề vanishing gradient và exploding gradient, khiến việc huấn luyện khó khăn trên chuỗi dài	Giải quyết được vấn đề vanishing gradient, giúp duy trì thông tin qua chuỗi dài
Thời gian huấn luyện	Huấn luyện nhanh hơn do cấu trúc đơn giản	Huấn luyện lâu hơn và tốn nhiều tài nguyên tính toán hơn
Hiệu suất	Hiệu suất giảm khi xử lý chuỗi dài hoặc có sự phụ thuộc dài hạn	Hiệu suất cao hơn trong việc xử lý các chuỗi dài và dữ liệu có mối quan hệ lâu dài
Ứng dụng	Hiệu suất cao hơn trong việc xử lý các chuỗi dài và dữ liệu có mối quan hệ lâu dài	Thích hợp cho chuỗi dài và bài toán cần giữ thông tin lâu dài

Bảng 2.1: So sánh mô hình RNN và LSTM

Chương 3

Ứng dụng mô hình chuỗi thời gian để dự báo giá chứng khoán

3.1 Phát biểu bài toán

Đồ án này tiến hành nghiên cứu và dự báo giá chứng khoán của Công ty Cổ phần Viễn thông FPT với cổ phiếu nổi bật của ngành này là FOX. Bài toán sẽ sử dụng các mô hình học sâu, cụ thể là các mô hình mạng nơ-ron hồi quy bao gồm thuật toán RNN và LSTM. Các mô hình này sẽ được đánh giá và so sánh bằng cách sử dụng các chỉ số đánh giá để đưa ra nhận xét và xác định mô hình tốt nhất. Mục tiêu chính của bài toán là sử dụng dữ liệu lịch sử về giá cổ phiếu và khối lượng giao dịch để dự đoán giá tương lai, từ đó hỗ trợ các nhà đầu tư đưa ra quyết định hợp lý và tối ưu hóa hiệu quả đầu tư.

3.2 Thu thập dữ liệu

vnstock là một thư viện python mã nguồn mở, được phát triển để giúp người dùng dễ dàng tải về và phân tích dữ liệu từ thị trường chứng khoán Việt Nam một cách nhanh chóng và miễn phí. Đây là nơi cung cấp một loạt các API cho phép truy cập và khai thác dữ liệu từ các sàn giao dịch chứng khoán tại Việt Nam bao gồm thông tin về giá cổ phiếu theo thời gian, dữ liệu tài chính của các công ty cũng như các chỉ số thị trường. Thư viện này lấy dữ liệu từ các nguồn đáng tin cậy như API của SSI (Công ty Chứng khoán Sài Gòn), TCBS (Chứng khoán Techcombank)... đảm bảo cung cấp thông tin chính xác và luôn được cập nhật về thị trường.

3.3 Mô tả dữ liệu

Bộ dữ liệu là thông tin về giá và khối lượng giao dịch hàng ngày của Công ty cổ phần Viễn thông FPT (Mã cổ phiếu: FOX) được lấy trong khoảng thời gian từ 20/11/2019 đến 20/11/2024. Em sử dụng thư viện *vnstock* để lấy dữ liệu chứng khoán của FPT từ sàn giao dịch HOSE.

Bộ dữ liệu chứa các thông tin như:

- time: Ngày giao dịch.
- open: Giá mở cửa.
- high: Giá cao nhất.
- low: Giá thấp nhất.
- close: Giá đóng cửa.
- volume: Khối lượng giao dịch.

	time	open	high	low	close	volume	ticker
0	2019-11-21	24060	24180	23680	23730	3303950	FPT
1	2019-11-22	23840	24180	23510	23730	1625820	FPT
2	2019-11-25	23770	24150	23640	23980	1087990	FPT
3	2019-11-26	24150	24230	24060	24180	968690	FPT
4	2019-11-27	24230	24310	24110	24150	812280	FPT

Hình 3.1: Dữ liệu giá và khối lượng giao dịch của cổ phiếu FPT

Các bước thực hiện chung cho các mô hình mạng nơ-ron bao gồm:

- Tiền xử lý dữ liệu: Chuyển dữ liệu thành dạng đa chiều và phù hợp với kiến trúc mạng của từng mô hình; chia dữ liệu thành các tập huấn luyện, xác thực và kiểm tra, chuẩn hoá dữ liệu.
- Xây dựng và huấn luyện mô hình: Xây dựng và huấn luyện các mô hình trên dữ liệu huấn luyện. Mỗi mô hình sẽ có quá trình huấn luyện riêng.
- Tính toán các chỉ số đánh giá: Tính toán các chỉ số đánh giá dựa trên kết quả dự báo từ mô hình và tập kiểm tra để đánh giá hiệu quả của mô hình.

3.4 Tiền xử lý dữ liệu

Loại bỏ các dữ liệu dư thừa và nhiễu: Loại bỏ các bản ghi có khối lượng giao dịch là 0 và các bản ghi trùng lặp.

Chia tập dữ liệu: Dữ liệu trong 5 năm được chia thành các tập huấn luyện, xác thực và kiểm tra. Tập huấn luyện dài 3 năm, từ ngày 20/11/2019 đến 20/11/2022. Tập xác thực dài 1 năm, từ ngày 20/11/2022 - 20/11/2023. Tập kiểm tra dài 1 năm, từ 20/11/2023 - 20/11/2024.

Chuẩn hoá dữ liệu: Chuẩn hoá dữ liệu bằng phương pháp Min-Max. Mục tiêu của phương pháp này là giá trị của các biến được biến đổi sao cho thuộc vào một khoảng giá trị từ [0;1]. Phương pháp này giúp thống nhất phạm vi của dữ liệu, đảm bảo tính đồng nhất và giảm thiểu ảnh hưởng của các biến có giá trị lớn đến quá trình xử lý và phân tích dữ liệu.

3.5 Xây dựng và huấn luyện mô hình

3.5.1 Dự báo bằng mô hình RNN đa biến

Ta tính toán mức độ tương quan giữa các biến open, high, low, close và volume với biến close. Kết quả tính toán:

- Open: 0.992533
- High: 0.997056
- Low: 0.996219
- Close: 1.000000
- Volume: 0.465851

Từ kết quả trên, mô hình RNN đa biến được đào tạo dựa trên chuỗi giá cao nhất (tương quan cao nhất với biến open), chuỗi khối lượng giao dịch (tương quan thấp nhất với biến open) và chuỗi giá đóng cửa của bộ dữ liệu. Việc tính toán và lựa chọn các biến có tương quan cao và thấp với biến close đóng vai trò quan trọng trong quá trình xây dựng mô hình dự báo chuỗi thời gian. Các biến có tương quan cao thường được sử dụng vì chúng có ảnh hưởng mạnh mẽ đến biến mục tiêu và cung cấp thông tin quan trọng để dự báo. Tuy nhiên, việc bổ sung các biến có tương

quan thấp có ý nghĩa để đảm bảo tính đa dạng và độc lập của thông tin, phát hiện các tương tác phức tạp, giảm thiểu dữ liệu nhiễu, và phát triển mô hình dự báo toàn diện hơn. Sự kết hợp cả hai loại biến này giúp mô hình hiểu rõ hơn về mối quan hệ giữa các biến và cải thiện khả năng dự báo chuỗi thời gian.

Các siêu tham số của mô hình này như sau:

- Timesteps: 30
- Hidden layers (hl): 50 và 45
- Learning rate (lr): 0.001
- Batch size: 32
- Epochs: 200
- Trainable params: 7087

Mô hình RNN được xây dựng bằng việc khởi tạo một mô hình tuần tự. Quá trình xây dựng mô hình RNN như sau:

- Lớp RNN đầu tiên có số đơn vị trong lớp RNN này bằng số lượng đặc trưng của mỗi bước thời gian trong dữ liệu đầu vào, hàm kích hoạt ReLU và trả về toàn bộ chuỗi đầu ra cho mỗi bước thời gian.
- Tiếp theo, chúng ta thêm một lớp RNN ẩn vào mô hình. Lớp RNN ẩn này có 50 đơn vị, sử dụng hàm kích hoạt ReLU và trả về toàn bộ chuỗi đầu ra.
- Lớp RNN cuối cùng trong mạng có 45 đơn vị. Cả hai lớp này cũng sử dụng hàm kích hoạt ReLU và trả về toàn bộ chuỗi đầu ra.
- Cuối cùng, chúng ta thêm một lớp Dense với 1 đơn vị, được sử dụng để dự báo giá trị đầu ra.

Ta lựa chọn tối ưu bằng phương pháp tối ưu Adam, sử dụng tốc độ học 0.001 và hàm mất mát được đánh giá thông qua chỉ số MSE.

3.5.2 Dự báo bằng mô hình LSTM đơn biến

Mô hình LSTM đơn biến chỉ được đào tạo dựa trên chuỗi giá đóng cửa của bộ dữ liệu. Do đó, nó là một mô hình đơn biến.

Các siêu tham số đóng một vai trò quan trọng, ảnh hưởng trực tiếp đến hiệu quả của mô hình. Vì vậy, trước khi quá trình huấn luyện bắt đầu, các siêu tham số của mô hình sẽ được tối ưu hóa bằng phương pháp GridSearch. Phương pháp này tự động kiểm tra các bộ kết hợp khác nhau của siêu tham số, và xác định bộ siêu tham số tốt nhất cho mô hình.

Các siêu tham số của mô hình này như sau:

- Timesteps: 40.
- Hidden layers (hl): 40 và 35.
- Learning rate (lr): 0.001.
- Batch size: 64.
- Epochs: 250.
- Trainable params: 17408

Mô hình LSTM đơn biến được xây dựng bằng việc khởi tạo một mô hình tuần tự từ Keras. Mô hình này cho phép xây dựng mạng nơ-ron bằng cách xếp chồng các lớp (layers) tuần tự:

- Lớp LSTM đầu tiên có 1 đơn vị, đầu vào là một chuỗi thời gian có 40 bước thời gian và mỗi bước thời gian có 1 đặc trưng, hàm kích hoạt ReLU, trả về toàn bộ chuỗi đầu ra để các lớp LSTM tiếp theo có thể nhận đầu vào là chuỗi các bước thời gian.
- Tiếp theo, chúng ta thêm một lớp LSTM ẩn vào mô hình. Lớp LSTM ẩn này có 40 đơn vị, sử dụng hàm kích hoạt ReLU và trả về toàn bộ chuỗi đầu ra.
- Lớp LSTM cuối cùng trong mạng có 35 đơn vị và sử dụng hàm kích hoạt ReLU, nhưng không cần trả về toàn bộ chuỗi đầu ra vì đây là lớp cuối cùng trong kiến trúc.
- Cuối cùng, chúng ta thêm một lớp Dense với 1 đơn vị, được sử dụng để dự báo giá trị đầu ra.

Trong mô hình này ta sẽ lựa chọn tối ưu bằng phương pháp tối ưu Adam, sử dụng tốc độ học 0.001 và hàm mất mát được đánh giá thông qua chỉ số MSE.

3.5.3 Dự báo bằng mô hình LSTM đa biến

Tương tự mô hình RNN, mô hình LSTM đa biến cũng được đào tạo với ba chuỗi giá cao nhất, giá đóng cửa và khối lượng giao dịch.

Các siêu tham số của mô hình này như sau:

- Timesteps: 50.
- Hidden layers (hl): 40 và 35.
- Learning rate (lr): 0.001.
- Batch size: 64.
- Epochs: 250.
- Trainable params: 17800

Tương tự mô hình RNN, quá trình xây dựng mô hình LSTM đa biến như sau:

- Lớp LSTM đầu tiên có số đơn vị trong lớp LSTM này bằng số lượng đặc trưng của mỗi bước thời gian trong dữ liệu đầu vào, hàm kích hoạt ReLU và trả về toàn bộ chuỗi đầu ra cho mỗi bước thời gian.
- Tiếp theo, chúng ta thêm một lớp LSTM ẩn vào mô hình. Lớp LSTM ẩn này có 40 đơn vị, sử dụng hàm kích hoạt ReLU và trả về toàn bộ chuỗi đầu ra.
- Lớp LSTM cuối cùng trong mạng có 35 đơn vị. Cả hai lớp này cũng sử dụng hàm kích hoạt ReLU và trả về toàn bộ chuỗi đầu ra.
- Cuối cùng, chúng ta thêm một lớp Dense với 1 đơn vị, được sử dụng để dự báo giá trị đầu ra.

Tương tự 2 mô hình trên, ta cũng sẽ lựa chọn tối ưu bằng phương pháp tối ưu Adam, sử dụng tốc độ học 0.001 và hàm mất mát được đánh giá thông qua chỉ số MSE.

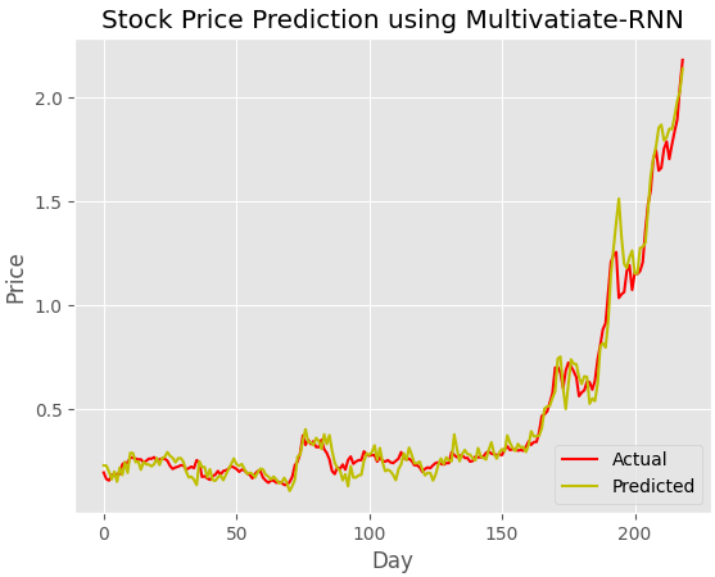
Chương 4

Phân tích kết quả và đánh giá mô hình

4.1 Kết quả của mô hình

4.1.1 Kết quả của mô hình RNN đa biến

Kết quả chạy mô hình RNN đa biến được mô tả như hình (4.1):



Hình 4.1: Dự đoán giá cổ phiếu FPT bằng mô hình RNN đa biến

Với dữ liệu kiểm tra là 750 ngày cuối cùng trong tập dữ liệu thì các chỉ số đánh giá cho mô hình RNN đa biến thu được kết quả như sau:

- Đối với mô hình RNN dự báo dựa trên ba biến giá đóng cửa, giá cao nhất và khối lượng giao dịch, ta có một số nhận xét như sau:
- Dự báo trên dữ liệu cổ phiếu FPT cho kết quả MSE thấp, cho thấy sai số bình phương trung bình giữa giá trị dự đoán và giá trị thực tế nhỏ.

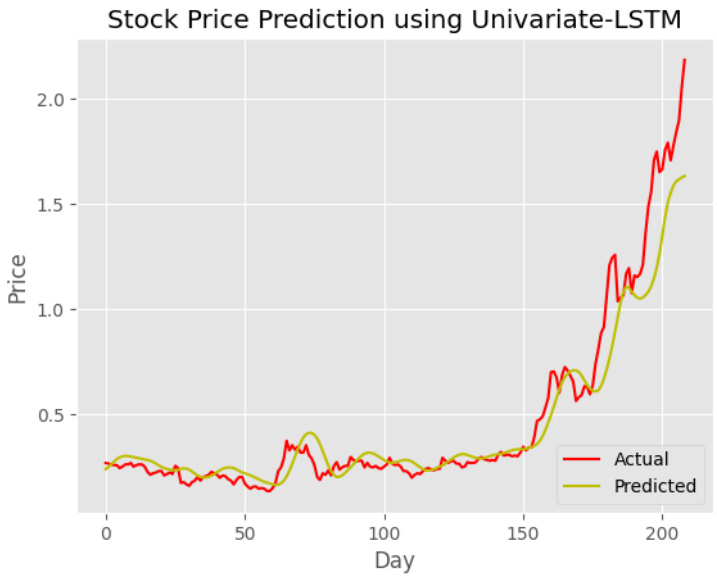
- Chỉ số RMSE cho thấy kết quả dự báo trên dữ liệu cổ phiếu FPT lệch khoảng 0.0665 đơn vị mỗi ngày so với giá trị thực tế.
- Với dữ liệu cổ phiếu FPT, sai số phần trăm trung bình giữa giá trị dự đoán và giá trị thực tế là khoảng 12.23% dựa trên chỉ số MAPE.
- Ở chỉ số R^2 có sự chênh lệch không quá lớn của bộ dữ liệu. Mô hình giải thích được 97.66% biến động của giá cổ phiếu FPT.

Giá chứng khoán	**Biến sử dụng**	**MSE**	**RMSE**	**MAPE**	R^2
Cổ phiếu FPT	high, volume, close	0.0044	0.0665	0.1223	0.9766

Bảng 4.1: Kết quả chỉ số đánh giá của mô hình RNN đa biến

4.1.2 Kết quả của mô hình LSTM đơn biến

Kết quả chạy mô hình LSTM đơn biến được mô tả như hình (4.2):



Hình 4.2: Dự đoán giá cổ phiếu FPT bằng mô hình LSTM đơn biến

Với dữ liệu kiểm tra là 750 ngày cuối cùng trong tập dữ liệu thì các chỉ số đánh giá cho mô hình LSTM đơn biến thu được kết quả như sau:

- Đối với mô hình LSTM chỉ dự báo dựa trên biến giá đóng cửa, ta có một số nhận xét như sau:

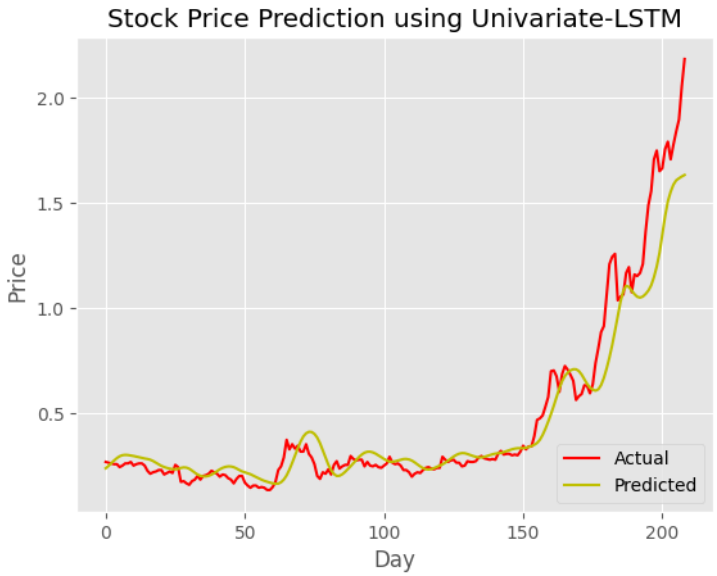
- Dự báo trên dữ liệu cổ phiếu FPT cho kết quả MSE thấp, cho thấy sai số bình phương trung bình giữa giá trị dự đoán và giá trị thực tế nhỏ.
- Chỉ số RMSE cho thấy kết quả dự báo trên dữ liệu cổ phiếu FPT lệch khoảng 0.1321 đơn vị mỗi ngày so với giá trị thực tế.
- Với dữ liệu cổ phiếu FPT, sai số phần trăm trung bình giữa giá trị dự đoán và giá trị thực tế là khoảng 16.79% dựa trên chỉ số MAPE.
- Ở chỉ số R2 có sự chênh lệch không quá lớn lớn của bộ dữ liệu. Mô hình giải thích được 91.06% biến động của giá cổ phiếu FPT.

Giá chứng khoán	**Biến sử dụng**	**MSE**	**RMSE**	**MAPE**	R^2
Cổ phiếu FPT	close	0.0174	0.1321	0.1679	0.9106

Bảng 4.2: Kết quả chỉ số đánh giá của mô hình LSTM đơn biến

4.1.3 Kết quả của mô hình LSTM đa biến

Kết quả chạy mô hình LSTM đa biến được mô tả như hình (4.3):



Hình 4.3: Dự đoán giá cổ phiếu FPT bằng mô hình LSTM đa biến

Với dữ liệu kiểm tra là 750 ngày cuối cùng trong tập dữ liệu thì các chỉ số đánh giá cho mô hình LSTM đa biến thu được kết quả như sau:

- Đối với mô hình LSTM dự báo dựa trên ba biến giá đóng cửa, giá cao nhất và khối lượng giao dịch, ta có một số nhận xét như sau:
- Dự báo trên dữ liệu cổ phiếu FPT cho kết quả MSE thấp, cho thấy sai số bình phương trung bình giữa giá trị dự đoán và giá trị thực tế nhỏ.
- Chỉ số RMSE cho thấy kết quả dự báo trên dữ liệu cổ phiếu FPT lệch khoảng 0.0939 đơn vị mỗi ngày so với giá trị thực tế.
- Với dữ liệu cổ phiếu FPT, sai số phần trăm trung bình giữa giá trị dự đoán và giá trị thực tế là khoảng 12.68% dựa trên chỉ số MAPE.
- Ở chỉ số R^2 có sự chênh lệch không quá lớn của bộ dữ liệu. Mô hình giải thích được 95.65% biến động của giá cổ phiếu FPT.

Giá chứng khoán	**Biến sử dụng**	**MSE**	**RMSE**	**MAPE**	R^2
Cổ phiếu FPT	high, volume, close	0.0088	0.0939	0.1268	0.9565

Bảng 4.3: Kết quả chỉ số đánh giá của mô hình LSTM đa biến

Nhìn chung, các chỉ số đánh giá của cả ba mô hình khi áp dụng trên bộ dữ liệu đều rất tích cực. Chỉ số MSE thấp cho thấy mô hình không có những sai lệch lớn giữa giá trị dự đoán và giá trị thực tế. Chỉ số RMSE cung cấp thông tin về mức độ lệch rất nhỏ của giá đóng cửa mỗi ngày mà mô hình dự đoán so với giá trị thực tế. Tương tự, chỉ số MAPE cho thấy tỷ lệ phần trăm lệch trung bình của dự đoán so với giá trị thực tế là khá nhỏ trên toàn bộ tập dữ liệu. Cuối cùng, giá trị R^2 cao chứng tỏ mô hình có khả năng giải thích phần lớn biến động của dữ liệu, thể hiện tính hiệu quả và độ tin cậy của các mô hình dự báo.

4.2 So sánh các mô hình

Ta sẽ tiến hành đánh giá hiệu suất của ba mô hình đã áp dụng trong bài toán dự báo giá chứng khoán ngành cổ phiếu FPT. Bằng việc đánh giá chi tiết và so sánh các mô hình này, ta có thể xác định mô hình nào mang lại kết quả dự báo tốt nhất. Để có thể đánh giá một cách minh bạch ta sẽ sẽ tiến hành so sánh ba mô hình thông qua các chỉ số: MSE, RMSE, MAPE, R^2. Kết quả được trình bày trong bảng (4.4):

Mô hình	Biến sử dụng	MSE	RMSE	MAPE	R^2
RNN đa biến	high, volume, close	0.0044	0.0665	0.1223	0.9766
LSTM đơn biến	close	0.0174	0.1321	0.1679	0.9106
LSTM đa biến	high, volume, close	0.0088	0.0939	0.1268	0.9565

Bảng 4.4: Chỉ số đánh giá kết quả dự báo giá chứng khoán cổ phiếu FPT

Từ bảng kết quả, ta nhận thấy mô hình LSTM đơn biến có hiệu suất kém hơn với hai mô hình đa biến. Mô hình RNN đã cho thấy kết quả tốt nhất trong cả bốn chỉ số đánh giá, với độ chính xác rất cao. Đối với mô hình LSTM đa biến, mặc dù vẫn có hiệu suất khá, nhưng chưa đạt được hiệu suất như mong đợi. Lý do cho điều này có thể là do bộ dữ liệu không có sự phụ thuộc lâu dài đủ để mô hình có thể học được hoặc cần thêm dữ liệu dài hơn để đào tạo.

Ngoài ra, từ số lượng tham số có thể đào tạo (hai mô hình LSTM đơn biến có khoảng 17000 tham số, trong khi đó mô hình RNN chỉ có 7087 tham số), ta có thể thấy mô hình RNN đã giảm độ phức tạp của mô hình, khiến cho việc phân tích đa biến hiệu quả hơn, chứng tỏ là một công cụ phù hợp để dự báo giá chứng khoán. Bên cạnh đó, chúng ta nhận thấy sự tương đồng rõ rệt trong kết quả dự báo và sự chênh lệch không đáng kể về các chỉ số đánh giá giữa ngành Viễn thông và cổ phiếu FPT. Điều này chứng tỏ rằng các mô hình dự báo hoạt động hiệu quả, sự nhất quán này khẳng định tính chính xác và độ tin cậy của các mô hình dự báo. Vì vậy tùy vào bối cảnh ứng dụng cụ thể và mục tiêu đặt ra mà ta sẽ tiến hành lựa chọn và xây dựng mô hình phù hợp nhất.

Kết luận

1. Kết quả đạt được của đồ án

Qua quá trình tìm hiểu, nghiên cứu và làm việc, đồ án tốt nghiệp với đề tài "Dự báo giá chứng khoán ngành Viễn thông dựa trên mô hình mạng nơ-ron hồi quy đa biến" đã thực hiện được những công việc sau:

- 1. Hiểu về chuỗi thời gian và bài toán dự báo chuỗi thời gian. Nghiên cứu và làm quen với các mô hình học sâu, đặc biệt là mô hình mạng nơ-ron hồi quy RNN và biến thể của nó LSTM.
- Thu thập dữ liệu chứng khoán từ thư viện vnstock trong Python.
- Xây dựng các mô hình mạng nơ-ron hồi quy như RNN, LSTM và ứng dụng giải bài toán thực tế - bài toán dự báo giá chứng khoán ngành và mã cổ phiếu tiềm năng đã xác định trước đó.

2. Hướng phát triển của đồ án trong tương lai

- Tiếp tục cải tiến và thử nghiệm thêm nhiều mô hình khác nhau, tối ưu hóa các tham số để nâng cao độ chính xác của dự báo.
- Tích hợp thêm các yếu tố kinh tế vĩ mô và sự kiện chính trị để tăng cường khả năng dự báo và đưa ra các dự đoán chính xác hơn trong các điều kiện thị trường phức tạp.

Một lần nữa, em muốn gửi lời cảm ơn đến giảng viên hướng dẫn em trong đồ án lần này là thầy **PGS.TS. Nguyễn Đình Hân**, cùng tất cả thầy cô Khoa Toán Tin đã dạy em rất nhiều kiến thức bổ ích từ các học phần khác nhau, góp phần giúp em hoàn thành đồ án. Đồ án không tránh khỏi những sai sót vụng về, một số thiết kế chưa phù hợp... Em rất mong nhận được sự đánh giá và góp ý từ thầy cô để bản thân có thêm kinh nghiệm hơn. Em xin chân thành cảm ơn ạ!

Tài liệu tham khảo

[1] VTV.vn. “Chứng khoán thế giới đi qua quý i "rực rỡ" nhất trong 5 năm.” Truy cập ngày 24 tháng 11 năm 2024. (2024), [Online]. Available: https://vtv.vn/kinh-te/chung-khoan-the-gioi-di-qua-quy-i-ruc-ro-nhat-trong-5-nam-20240401161722761.htm.

[2] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning internal representations by error propagation,” *Nature*, vol. 323, no. 6088, pp. 533–536, 1986. DOI: 10.1038/323533a0.

[3] R. J. Williams and D. Zipser, “A learning algorithm for continually running fully recurrent neural networks,” *Neural Computation*, vol. 1, no. 2, pp. 270–280, 1989. DOI: 10.1162/neco.1989.1.2.270.

[4] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735.

Phụ lục