**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**PHÂN HIỆU TẠI THÀNH PHỐ HỒ CHÍ MINH**

**BỘ MÔN CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO ĐỒ ÁN TỐT NGHIỆP**

**ĐỀ TÀI: NGHIÊN CỨU VÀ ỨNG DỤNG**

**MÔ HÌNH ARIMA VÀ MÔ HÌNH LSTM VÀO DỰ ĐOÁN THỊ TRƯỜNG CHỨNG KHOÁN**

Giảng viên hướng dẫn: ThS. Trần Thị Dung

Sinh viên thực hiện: Trần Viết Học

Lớp: CÔNG NGHỆ THÔNG TIN

Khóa: 60

TP. Hồ Chí Minh, năm 2024

**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**PHÂN HIỆU TẠI THÀNH PHỐ HỒ CHÍ MINH**

**BỘ MÔN CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO ĐỒ ÁN TỐT NGHIỆP**

**ĐỀ TÀI: NGHIÊN CỨU VÀ ỨNG DỤNG**

**MÔ HÌNH ARIMA VÀ MÔ HÌNH LSTM VÀO DỰ ĐOÁN THỊ TRƯỜNG CHỨNG KHOÁN**

Giảng viên hướng dẫn: ThS. Trần Thị Dung

Sinh viên thực hiện: Trần Viết Học

Lớp: CÔNG NGHỆ THÔNG TIN

Khóa: 60

TP. Hồ Chí Minh, năm 2024

TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢICỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

PHÂN HIỆU TẠI THÀNH PHỐ HỒ CHÍ MINHĐộc lập – Tự do – Hạnh phúc

**NHIỆM VỤ THIẾT KẾ TỐT NGHIỆP**

BỘ MÔN**: CÔNG NGHỆ THÔNG TIN**

**-------\*\*\*-------**

**Mã sinh viên: 6051071048 Họ và tên: Trần Viết Học**

**Khóa: 60 Lớp: Công nghệ thông tin**

**1.** **Tên đề tài:**

Nghiên cứu dự đoán thị trường chứng khoán.

**2.** **Mục tiêu:**

Tìm hiểu về ngôn ngữ Python và nghiên cứu một số thuật toán máy học về phân tích và dự đoán kết quả như ARIMA, RNN, Long – short term memory. Từ đó ứng dụng vào phân tích và đưa ra các dự đoán về giá dựa trên dataset về chứng khoán.

**3.** **Nội dung thực hiện:**

* Tìm hiểu ngôn ngữ Python và các thư viện cần sử dụng
* Tìm hiểu sơ bộ về Machine Learning
* Nghiên cứu thuật toán máy học: mô hình ARIMA, mạng RNN và mạng Long – short term memory.
* Nghiên cứu bài toán phân tích và dự đoán về giá chứng khoán
* Áp dụng kiến thức: ứng dụng ngôn ngữ Python và 2 thuật toán vào phân tích và đưa ra dự đoán về giá chứng khoán

**4.** **Công nghệ, công cụ và ngôn ngữ lập trình**

* Công cụ sử dụng: Visual Studio Code, Colab
* Ngôn ngữ: Python

**5. Các kết quả chính dự kiến**

* Hiểu và sử dụng được ngôn ngữ lập trình Python
* Hiểu được các thuật toán máy học cần sử dụng
* Cài đặt được môi trường sử dụng ngôn ngữ
* Áp dụng được kiến thức và cho ra kết quả

**6. Kế hoạch đang thực hiện**

* Tuần 1-2 và 3: Tìm và chọn đề tài
* Tuần 4: Đưa ra lựa chọn về đề tài
* Tuần 5-6: Tìm hiểu ngôn ngữ Python, thư viện cần sử dụng, đọc sách về ứng dụng AI vào phân tích thị trường chứng khoán.
* Tuần 7 đến 11: Nghiên cứu các thuật toán máy học và áp dụng kiến thức vào bài toán.
* Tuần 12: Viết báo cáo và làm slide
* Tuần 13: Nộp báo cáo và chờ duyệt

**7. Giảng viên và cán bộ hướng dẫn**

Họ tên: Ths. Trần Thị Dung

Đơn vị công tác: Trường Đại học Giao thông Vận tải Phân hiệu tại TP. Hồ Chí Minh

|  |  |
| --- | --- |
| Điện thoại: 0388389579 | Email: ttdung@utc2.edu.vn |

|  |  |
| --- | --- |
| **Ngày … tháng … năm 2023**  **Trưởng BM Công nghệ thông tin** | **Đã giao nhiệm vụ TKTN**  **Giảng viên hướng dẫn** |
|  | **Trần Thị Dung** |

**LỜI CẢM ƠN**

Đầu tiên em xin gửi lời cảm ơn và bày tỏ lòng biết ơn chân thành đến cô – Th.s Trần Thị Dung, người đã hướng dẫn, cung cấp kiến thức, nguồn tài liệu và tận tình hưỡng dẫn chỉ bảo cho em trong thời gian thực hiện đồ án tốt nghiệp của mình.

Em cũng xin gửi lời cảm ơn đến các thầy giáo, cô giáo khoa Công nghệ Thông tin nói riêng và các thầy giáo, cô giáo của Đại học Giao Thông Vận Tải nói chung đã tận tình giảng dạy, truyền đạt kiến thức cũng như những kinh nghiệm quý báu cho em cùng toàn thể các sinh viên khóa 60 trong suốt quá trình học tập tại trường.

Em xin bày tỏ lòng biết ơn sâu sắc đến Quý thầy cô vì đã tận tâm hướng dẫn và hỗ trợ em trong quá trình hoàn thành đề tài. Mặc dù em đã cố gắng hết sức để đạt được kết quả tốt nhất, nhưng chắc chắn rằng sẽ khó tránh khỏi những thiếu sót. Em mong nhận được những sự đánh giá, góp ý chân thành từ Quý thầy cô để em có thể rút ra những bài học, kinh nghiệm quý báu. Sự hỗ trợ và góp ý của Quý thầy cô là nguồn động viên quan trọng giúp em phát triển và hoàn thiện bản thân. Em sẽ lắng nghe và đón nhận mọi ý kiến xây dựng để nâng cao chất lượng công trình nghiên cứu của mình.

Cuối cùng em xin gửi lời chúc sức khỏe và thành công hơn đến với tất cả các quý thầy cô, bạn bè. Em xin chân thành cảm ơn!

Hồ Chí Minh, ngày tháng 12 năm 2023

Sinh viên

**Trần Viết Học**

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

|  |
| --- |
| ***Tp. Hồ Chí Minh, ngày ….… tháng ….… năm ….…***  **Giảng viên hướng dẫn**  **Trần Thị Dung** |

**MỤC LỤC**

[DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT 1](#_Toc155979652)

[DANH SÁCH CÁC HÌNH VẼ VÀ ĐỒ THỊ 2](#_Toc155979653)

[MỞ ĐẦU 4](#_Toc155979654)

[1. Lý do chọn đề tài 4](#_Toc155979655)

[2. Mục tiêu và mục tiêu của đồ án 5](#_Toc155979656)

[3. Bố cục đồ án 5](#_Toc155979657)

[Chương 1. CƠ SỞ LÝ THUYẾT 6](#_Toc155979658)

[1.1. Chứng khoán và thị trường chứng khoán 6](#_Toc155979659)

[1.1.1. Chứng khoán 6](#_Toc155979660)

[1.1.2. Thị trường chứng khoán 6](#_Toc155979661)

[1.2. Mối liên hệ giữa học máy và thị trường chứng khoán 6](#_Toc155979662)

[1.3. Ngôn ngữ Python 7](#_Toc155979663)

[1.3.1. Giới thiệu 7](#_Toc155979664)

[1.3.2. Một số thư viện của ngôn ngữ Python 7](#_Toc155979665)

[1.4. Sơ lược về Machine Learning 8](#_Toc155979666)

[1.4.1. Giới thiệu 8](#_Toc155979667)

[1.4.2. Phân loại học học máy 8](#_Toc155979668)

[1.5. Dữ liệu chuỗi thời gian 9](#_Toc155979669)

[1.6. Mạng RNN 11](#_Toc155979670)

[1.6.1. Giới thiệu 11](#_Toc155979671)

[1.6.2. Kiến trúc 11](#_Toc155979672)

[1.7. Mạng LSTM 12](#_Toc155979673)

[1.7.1. Giới thiệu 12](#_Toc155979674)

[1.7.2. Kiến trúc 13](#_Toc155979675)

[1.8. Mô hình ARIMA 16](#_Toc155979676)

[1.8.1. Giới thiệu 16](#_Toc155979677)

[1.8.2. Kiến trúc 16](#_Toc155979678)

[1.9. Phương pháp đánh giá mô hình 18](#_Toc155979679)

[Chương 2. BÀI TOÁN DỰ ĐOÁN GIÁ CHỨNG KHOÁN 20](#_Toc155979680)

[2.1. Chuẩn bị và phân tích dữ liệu 20](#_Toc155979681)

[2.2. Xây dựng mô hình 21](#_Toc155979682)

[2.3. Dữ liệu thực nghiệm 22](#_Toc155979683)

[2.4. Môi trường thực nghiệm 22](#_Toc155979684)

[2.5. Xây dựng thực nghiệm 23](#_Toc155979685)

[2.5.1. Cấu trúc mô hình 24](#_Toc155979686)

[2.5.2. Thực nghiệm 1: Cổ phiếu Apple – AAPL. 26](#_Toc155979687)

[2.5.3. Thực nghiệm 2: Cổ phiếu Microsoft– MSFT 32](#_Toc155979688)

[2.5.4. Thực nghiệm 3: Cổ biếu Tesla – TSLA 38](#_Toc155979689)

[2.5.5. Kết luận 43](#_Toc155979690)

[Chương 3. Xây dựng hệ thống dự báo chứng khoán 45](#_Toc155979691)

[3.1. Đặt vấn đề 45](#_Toc155979692)

[3.2. Phân tích yêu cầu hệ thống 45](#_Toc155979693)

[3.3. Sơ đồ Use-Case 45](#_Toc155979694)

[3.3.1. Sơ đồ use-case xem thông tin chứng khoán 45](#_Toc155979695)

[3.3.2. Sơ đồ use-case xem thông tin dự đoán giá chứng khoán 46](#_Toc155979696)

[3.4. Kết quả 47](#_Toc155979697)

[3.4.1. Màn hình trang chủ 47](#_Toc155979698)

[3.4.2. Màn hình dự đoán chứng khoán 48](#_Toc155979699)

[Kết luận 49](#_Toc155979700)

[1. Kết quả đạt được 49](#_Toc155979701)

[2. Hạn chế 49](#_Toc155979702)

[3. Hướng phát triển 50](#_Toc155979703)

[THAM KHẢO 51](#_Toc155979704)

# DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Từ viết tắt | Thuật ngữ | Ý nghĩa |
| 1 | RNN | Recurrent Neural Network | Mạng nơ-ron hồi quy |
| 2 | LSTM | Long short term memory | Bộ nhớ dài ngắn hạn |
| 3 | ARIMA | Autoregressive integrated moving average | Đường trung bình trượt tích hợp tự động hồi quy |

# DANH SÁCH CÁC HÌNH VẼ VÀ ĐỒ THỊ

[Hình 1.1 Ví dụ về Time series 9](#_Toc155979078)

[Hình 1.2 Mạng RNN 11](#_Toc155979079)

[Hình 1.3 Mạng LSTM 13](#_Toc155979080)

[Hình 1.4 Giải thích các ký hiệu trong hình mạng LSTM 13](#_Toc155979081)

[Hình 1.5 Một state của mạng LSTM 14](#_Toc155979082)

[Hình 1.6Cổng forget 14](#_Toc155979083)

[Hình 1.7 Cổng Input 15](#_Toc155979084)

[Hình 1.8 Cổng Output 15](#_Toc155979085)

[Hình 1.9 Giá trị state C 15](#_Toc155979086)

[Hình 1.10 Ct của LSTM 16](#_Toc155979087)

[Hình 2.1 Ví dụ một số mẫu thử nghiệm 20](#_Toc155979088)

[Hình 2.2 Mô hình đề xuất 21](#_Toc155979089)

[Hình 2.3 Dữ liệu thực nghiệm 23](#_Toc155979090)

[Hình 2.4 Kết quả ARIMA thực nghiệm 1 phần 1. 26](#_Toc155979091)

[Hình 2.5 Kết quả RNN thực nghiệm 1 phần 1 27](#_Toc155979092)

[Hình 2.6 Kết quả LSTM thực nghiệm 1 phần 1 28](#_Toc155979093)

[Hình 2.7 Kết quả ARIMA thực nghiệm 1 phần 2 29](#_Toc155979094)

[Hình 2.8 Kết quả RNN thực nghiệm 1 phần 2 30](#_Toc155979095)

[Hình 2.9 Kết quả mô hình LSTM thực nghiệm 1 phần 2 31](#_Toc155979096)

[Hình 2.10 Kết quả ARIMA thực nghiệm 2 phần 1 32](#_Toc155979097)

[Hình 2.11 Kết quả RNN thực nghiệm 2 phần 1 33](#_Toc155979098)

[Hình 2.12 Kết quả LSTM thực nghiệm 2 phần 1 34](#_Toc155979099)

[Hình 2.13 Kết quả mô hình ARIMA thực nghiệm 2 phần 2 35](#_Toc155979100)

[Hình 2.14 Kết quả mô hình RNN thực nghiệm 2 phần 2 36](#_Toc155979101)

[Hình 2.15 Kết quả LSTM thực nghiệm 2 phần 2 37](#_Toc155979102)

[Hình 2.16 Kết quả ARIMA thực nghiệm 3 phần 1 38](#_Toc155979103)

[Hình 2.17 Kết quả RNN thực nghiệm 3 phần 1 39](#_Toc155979104)

[Hình 2.18 Kết quả LSTM thực nghiệm 3 phần 1 40](#_Toc155979105)

[Hình 2.19 Kết quả ARIMA thực nghiệm 3 phần 2 41](#_Toc155979106)

[Hình 2.20 Kết quả RNN thực nghiệm 3 phần 2 42](#_Toc155979107)

[Hình 2.21 Kết quả LSTM thực nghiệm 3 phần 2 43](#_Toc155979108)

[Hình 3.1 Sơ đồ use case tổng quát 45](#_Toc155979109)

[Hình 3.2 Sơ đồ use xem giá dự đoán 46](#_Toc155979110)

[Hình 4.1Màn hình trang chủ 47](#_Toc155979111)

[Hình 4.2 Màn hình trang chủ với biểu đồ 47](#_Toc155979112)

[Hình 4.3Màn hình dự đoán bằng ARIMA 48](#_Toc155979113)

[Hình 4.4 Màn hình dự đoán bằng LSTM 48](#_Toc155979114)

# MỞ ĐẦU

## Lý do chọn đề tài

Trong kỉ nguyên của mạng internet hiện nay, chúng ta đã có khả năng tiếp cận rất nhiều thông tin khác nhau và việc nảy sinh ra ý định kiếm từ từ môi trường này cũng là điều hiển nhiên. Đầu tư chứng khoán cũng là một trong những hình thức phổ biến đáp ứng được mục đích đầu tư tài chính thông qua công nghệ hiện nay.

Trên thị trường chứng khoán ngày nay, có nhiều loại chứng khoán khác nhau để các nhà đầu tư lựa chọn. Khi tham gia đầu tư, mong muốn chính của họ là tối ưu hóa lợi nhuận theo thời gian và hiểu rõ hơn về thời điểm nào là lúc thích hợp để thêm vốn hoặc rút vốn. Để đạt được điều này, nhà đầu tư cần dự đoán chính xác về sự biến động trên thị trường chứng khoán. Quyết định thông minh về cách phân bổ vốn đầu tư là chìa khóa quan trọng.

Dự báo sự biến động trên thị trường chứng khoán là một chủ đề quan trọng trong lĩnh vực tài chính. Việc dự báo hiệu quả giúp nhà đầu tư xây dựng chiến lược đầu tư tối ưu và giảm thiểu rủi ro. Tuy nhiên, với sự phức tạp của các yếu tố ảnh hưởng như tăng trưởng kinh tế, tình hình chính trị, và thông tin truyền thông, việc dự báo trở nên khó khăn.

Trong lĩnh vực đầu tư chứng khoán, quyết định đúng đắn về thời điểm mua bán là một thách thức lớn, đòi hỏi người đầu tư cần có thông tin đầy đủ để dự đoán sự biến động của giá thị trường. Thông tin này đóng vai trò quan trọng, và việc phân tích lượng lớn thông tin này trở nên quan trọng để hiểu rõ xu hướng và biến động của thị trường chứng khoán.

Tính đến hiện nay, máy học (Machine Learning) đã trở thành một công cụ mạnh mẽ được sử dụng để hỗ trợ quản lý và đầu tư hiệu quả. Tuy nhiên, việc đưa vào mô hình các yếu tố phản ánh thực tế đòi hỏi sự hiểu biết sâu rộng về lĩnh vực chứng khoán.

Đồng thời, trong đề tài của mình, em đã nghiên cứu và ứng dụng mô hình ARIMA cùng với mô hình học sâu LSTM để dự đoán giá cổ phiếu dựa trên giá đóng cửa của cổ phiếu trong các ngày trước đó. Đây là một bước quan trọng nhằm xác định chiến lược đầu tư thông minh và giảm thiểu rủi ro đối với những người đầu tư.

## Mục tiêu và mục tiêu của đồ án

Tìm hiểu về ngôn ngữ Python và nghiên cứu thuật toán máy học về phân tích và dự đoán kết quả như ARIMA, Long – short term memory., RNN. Từ đó ứng dụng vào phân tích và đưa ra các dự đoán về giá dựa trên dataset về cổ phiếu được lấy từ trang finance.yahoo.com

## Bố cục đồ án

Bố cục của đồ án được chia làm 4 phần và bao gồm những nội dung sau:

* Chương 1: Cơ sở lý thuyết: Tìm hiểu thuật toán học máy Logistic Regression và mô hình mạng LSTM. Các khái niệm liên quan đến đề tài nghiên cứu.
* Chương 2: Bài toán dự đoán giá chứng khoán: Gồm phân tích dữ liệu, đưa ra mô hình phù hợp và phương pháp đánh giá mô hình, xây dựng cài đặt mô hình, huấn luyện mô hình, thực hiện thử nghiệm dự đoán.
* Chương 3: Phân tích thiết kế hệ thống: Phân tích những yêu cầu của hệ thống, thiết kế hệ thống đáp ứng yêu cầu cầu của hệ thống.
* Chương 4: Xây dựng chương trình: Xây dựng chương trình đã phân tích và áp dụng thuật toán đã nghiên cứu
* Kết luận: Tổng kết lại quá trình nghiên cứu và thực nghiệm, hệ thống đã xây dựng, những kết quả đạt được.

# CƠ SỞ LÝ THUYẾT

## Chứng khoán và thị trường chứng khoán

### Chứng khoán

Chứng khoán là bằng chứng xác nhận quyền và lợi ích hợp pháp của người sở hữu đối với tài sản hoặc phần vốn của tổ chức phát hành. Chứng khoán được thể hiện dưới hình thức chứng chỉ, bút toán ghi sổ hoặc dữ liệu điện tử.

Chứng khoán bao gồm: Cổ phiếu, trái phiếu, chứng chỉ quỹ đầu tư, chứng khoán phái sinh.

### Thị trường chứng khoán

Thị trường chứng khoán là một bộ phận quan trọng của thị trường vốn, hoạt động của nó nhằm huy động những nguồn vốn tiết kiệm nhỏ trong xã hội tập trung thành nguồn vốn lớn tài trợ dài hạn cho các doanh nghiệp, các tổ chức kinh tế và Nhà nước để phát triển sản xuất, tăng trưởng kinh tế hay cho các dự án đầu tư.

Thị trường chứng khoán là nơi diễn ra các hoạt động giao dịch mua bán các loại chứng khoán. Việc mua bán này trước tiên được tiến hành ở thị trường sơ cấp khi người mua mua được chứng khoán lần đầu từ những người phát hành và sau đó ở thị trường thứ cấp khi có sự mua đi bán lại các chứng khoán đã được phát hành ở thị trường sơ cấp. Do vậy, thị trường chứng khoán là nơi các chứng khoán được phát hành và trao đổi.

## Mối liên hệ giữa học máy và thị trường chứng khoán

Những biến động trong thị trường chứng khoán luôn có nhiều nguyên nhân phức tạp và khác nhau. Tuy nhiên, điều này không có nghĩa việc dự đoán xu hướng của thị trường này là việc không thể. Trên thực tế, học máy đã được ứng dụng vào dự đoán bằng việc phân tích và tận dụng tối đa lượng dữ liệu lịch sử kết hợp cùng với kiến thức về tài chính của người xây dựng mô hình dự đoán.

Học máy là mô hình AI được sử dụng rộng rãi nhất trong lĩnh vực tài chính, dựa trên một công trình nghiên cứu hồi năm 1943 của McCullogh và Pitts. Về nguyên tắc, một hệ thống học máy bao gồm: nguồn dữ liệu, mô hình, thuật toán tối ưu, hệ thống đánh giá và kiểm thử.

## Ngôn ngữ Python

### Giới thiệu

Python là một ngôn ngữ lập trình bậc cao hướng đối tượng được Guido van Rossum cùng các cộng sự tạo ra năm 1991, dành cho mục đích lập trình đa năng. Python được thiết kế với ưu điểm mạnh là câu lệnh ngắn gọn, dễ nhớ, dễ hiểu. Cấu trúc chương trình của Python rõ ràng, dễ đọc và viết hơn rất nhiều so với những ngôn ngữ lập trình khác. Đây là ngôn ngữ lập trình thông dịch, có thể chạy trên nhiều hệ điều hành khác nhau. Python là ngôn ngữ mã nguồn mở và có cộng đồng người dùng lớn.

Một số đặc điểm của ngôn ngữ Python:

* Đơn giản, dễ học
* Miễn phí, mã nguồn mở
* Khả chuyển
* Khả năng mở rộng và khả năng nhúng

### Một số thư viện của ngôn ngữ Python

Numpy: Tên "Numpy" là viết tắt của "Numerical Python". Nó là thư viện thường được sử dụng. Một thư viện học máy phổ biến hỗ trợ các ma trận và dữ liệu đa chiều. Bao gồm các hàm toán học được xây dựng sẵn để dễ dàng tính toán.

Pandas: Pandas là một thư viện quan trọng cho các nhà khoa học dữ liệu. Đây là một thư viện máy học mã nguồn mở cung cấp các cấu trúc dữ liệu cấp cao linh hoạt và nhiều công cụ phân tích.

Matplotlib: Thư viện này chịu trách nhiệm vẽ dữ liệu số. Và đó là lý do tại sao nó được sử dụng trong phân tích dữ liệu.

Sklearn (hay scikit-learn): Nó là một thư viện Python nổi tiếng để làm việc với dữ liệu phức tạp. Sklearn là một thư viện mã nguồn mở hỗ trợ học máy

Keras: Keras cung cấp thư viện tiện ích numpy, cung cấp các hàm để thực hiện các hành động trên mảng numpy. Sử dụng phương thức to\_categorical (), một mảng numpy (hoặc) một vectơ có các số nguyên đại diện cho các danh mục khác nhau, có thể được chuyển đổi thành một mảng numpy (hoặc) một ma trận có các giá trị nhị phân và có các cột bằng số danh mục trong dữ liệu.

Seaborn: Seaborn là một thư viện trực quan tuyệt vời để vẽ đồ họa thống kê bằng Python. Nó cung cấp các kiểu và bảng màu mặc định đẹp mắt để làm cho các ô thống kê trở nên hấp dẫn hơn.

## Sơ lược về Machine Learning

### Giới thiệu

Học máy là một lĩnh vực của trí tuệ nhân tạo. Nó là một lĩnh vực nghiên cứu ở giao điểm của thống kê, trí tuệ nhân tạo và khoa học máy tính và còn được gọi là phân tích dự đoán hoặc học thống kê.

Việc nghiên cứu áp dụng các phương pháp học máy trong những năm gần đây đã trở nên phổ biến trong cuộc sống hàng ngày. Từ tự động gợi ý về những bộ phim nên xem, quảng cáo sản phẩm người dùng muốn mua, đề xuất các sản phẩm liên quan đến sản phẩm người dùng đã mua, hoặc muốn mua để người dùng đưa ra những so sánh rồi lựa chọn, chẩn đoán y khoa, phát hiện thẻ tín dụng giả, phân tích thị trường chứng khoán, phân loại các chuỗi DNA, nhận dạng tiếng nói và chữ viết, dịch tự động, chơi trò chơi và cử động rô-bốt (robot locomotion)…

Học máy rất gần với suy diễn thống kê (statistical inference) tuy có khác nhau về thuật ngữ. Một nhánh của học máy là học sâu (Deep Learning) phát triển rất mạnh mẽ gần đây và có những kết quả vượt trội so với các phương pháp học máy khác.

### Phân loại học học máy

Các thuật toán học máy được phân loại theo kết quả mong muốn của thuật toán. Các loại thuật toán thường dùng bao gồm:

• Học có giám sát - trong đó, thuật toán tạo ra một hàm ánh xạ dữ liệu vào tới kết quả mong muốn.

• Học không giám sát - mô hình hóa một tập dữ liệu, không có sẵn các ví dụ đã được gắn nhãn.

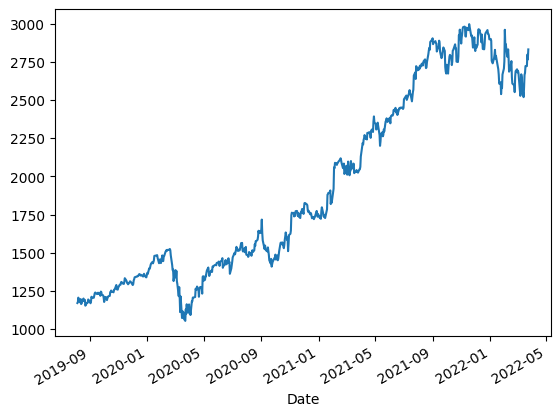
• Học nửa giám sát - kết hợp các ví dụ có gắn nhãn và không gắn nhãn để sinh một hàm hoặc một bộ phân loại thích hợp.

• Học tăng cường - trong đó, thuật toán học một chính sách hành động tùy theo các quan sát về thế giới. Mỗi hành động đều có tác động tới môi trường, và môi trường cung cấp thông tin phản hồi để hướng dẫn cho thuật toán của quá trình học.

## Dữ liệu chuỗi thời gian

Dữ liệu về chứng khoán được ghi nhận theo ngày và kéo dài liên tục cho tới khi chúng bị loại bỏ khỏi sàn giao dịch. Do vậy mà chúng được xếp vào dạng dữ liệu là Time series data.

Time series data: là một chuỗi các điểm dữ liệu, thường bao gồm các phép đo liên tiếp được thực hiện từ cùng một nguồn trong một khoảng thời gian. Phân tích chuỗi thời gian có mục đích nhận dạng và tập hợp lại các yếu tố, những biến đổi theo thời gian mà nó có ảnh hưởng đến giá trị của biến quan sát.



Hình . Ví dụ về Time series

Trong Time-series Data, có hai loại chính.

* Chuỗi thời gian thông thường (regular time series), loại thông thường được gọi là số liệu.
* Chuỗi thời gian bất thường (events) là những sự kiện.

Ứng dụng: Time-series data được ứng dụng rất rộng rãi trong các lĩnh vực:

* IoT
* DevOps
* Phân tích thời gian thực
* Dự báo kinh tế
* Tính toán doanh số bán hàng
* Phân tích lãi
* Phân tích thị trường
* Kiểm soát quy trình và chất lượng
* Phân tích điều tra

Ưu điểm của chuỗi thời gian là nó có thể lưu trữ được trạng thái của một trường dữ liệu theo thời gian. Dữ liệu chuỗi thời gian có tính ứng dụng rất cao và được áp dụng trong rất nhiều lĩnh vực khác nhau như: thống kê, kinh tế lượng, toán tài chính, dự báo thời tiết, dự đoán động đất, điện não đồ, kỹ thuật điều khiển, thiên văn, kỹ thuật truyền thông, xử lý tín hiệu.

Dữ liệu chuỗi thời gian có những tính chất đặc trưng riêng như:

* Tính xu hướng: Tính xu hướng là yếu tố thể hiện xu hướng thay đổi của dữ liệu theo thời gian. Đây là đặc trưng thường thấy của rất nhiều dữ liệu chuỗi thời gian.
* Tính chu kỳ: Là qui luật có tính chất lặp lại của dữ liệu theo thời gian. Ví dụ, dữ liệu về số lượng xe cộ qua lại trên một con đường có thể có tính chu kỳ hàng ngày, hàng tuần hoặc hàng năm.
* Tính mùa vụ: Là qui luật có tính chất lặp lại của dữ liệu theo thời gian theo mùa vụ. Ví dụ, dữ liệu về sản lượng lúa mì hàng năm có thể có tính mùa vụ.
* Tính ngẫu nhiên: Là qui luật không có tính chất lặp lại của dữ liệu theo thời gian. Ví dụ, dữ liệu về giá cổ phiếu của một công ty có thể có tính ngẫu nhiên.
* Tính dư thừa: Là qui luật có tính chất lặp lại của dữ liệu theo thời gian nhưng không có tính chu kỳ. Ví dụ, dữ liệu về số lượng khách hàng mua hàng trên một trang web có thể có tính dư thừa.
* Tính tương quan: Là qui luật mô tả mối quan hệ giữa hai chuỗi thời gian. Ví dụ, dữ liệu về giá cổ phiếu của hai công ty có thể có tính tương quan.

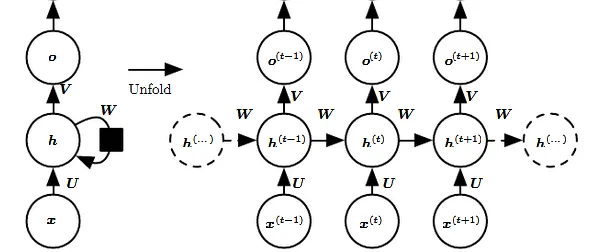
## Mạng RNN

### Giới thiệu

Mạng thần kinh hồi quy (RNN) là một loại mạng thần kinh nhân tạo có khả năng xử lý dữ liệu chuỗi thời gian. RNN được thiết kế để giải quyết các bài toán dạng chuỗi như hoàn thành câu, dịch máy, phân loại văn bản, và nhiều ứng dụng khác. Mạng RNN có khả năng lưu trữ thông tin từ các bước tính toán trước để dựa vào đó đưa ra dự đoán chính xác nhất cho bước dự đoán hiện tại. Mạng RNN bao gồm 3 phần chính: Input layer, Hidden layer và Output layer. Các input sẽ được kết hợp với hidden layer bằng hàm f\_W để tính toán ra hidden layer hiện tại và output sẽ được tính ra từ hidden layer đó.

### Kiến trúc

Kiến trúc RNN bao gồm các đơn vị lặp (recurrent units) kết hợp với trọng số có thể được chia sẻ qua nhiều bước thời gian. Điều này cho phép RNN giữ lại thông tin về quá khứ để áp dụng cho các dữ liệu mới khi chúng được đưa vào.



Hình . Mạng RNN

Phía bên trái của sơ đồ trên hiển thị ký hiệu của RNN và ở phía bên phải là RNN đang được mở (hoặc mở ra) thành một mạng đầy đủ.

* Đầu vào: x(t) được lấy làm đầu vào của mạng tại bước thời gian t
* Trạng thái ẩn : h(t) đại diện cho trạng thái ẩn tại thời điểm t và đóng vai trò là “bộ nhớ” của mạng. h(t) được tính dựa trên đầu vào hiện tại và trạng thái ẩn của bước thời gian trước đó: Hàm được lấy là một phép biến đổi phi tuyến tính như tanh , ReLU. **h(t)** = f(U **x(t)** + W **h(t**−**1)**).f
* Trọng số : RNN có đầu vào cho các kết nối ẩn được tham số hóa bằng ma trận trọng số U, các kết nối lặp lại ẩn đến ẩn được tham số hóa bởi ma trận trọng số W và các kết nối ẩn đến đầu ra được tham số hóa bởi ma trận trọng số V và tất cả các trọng số này ( U , V , W) được chia sẻ theo thời gian.
* Đầu ra : o(t) minh họa đầu ra của mạng. Trong hình chỉ đặt một mũi tên sau o(t) cũng thường chịu tính phi tuyến tính, đặc biệt khi mạng chứa các lớp tiếp theo ở phía dưới.

Và một trong những vấn đề của mô hình RNN gặp vấn đề vanishing gradient, khi gradient (đạo hàm) truyền ngược từ lớp cuối cùng đến các lớp đầu tiên bị giảm dần và gây khó khăn trong việc cập nhật trọng số. Để khắc phục vấn đề này, các phiên bản cải tiến của RNN như LSTM (Long Short-Term Memory) và GRU (Gated Recurrent Unit) đã được phát triển.

## Mạng LSTM

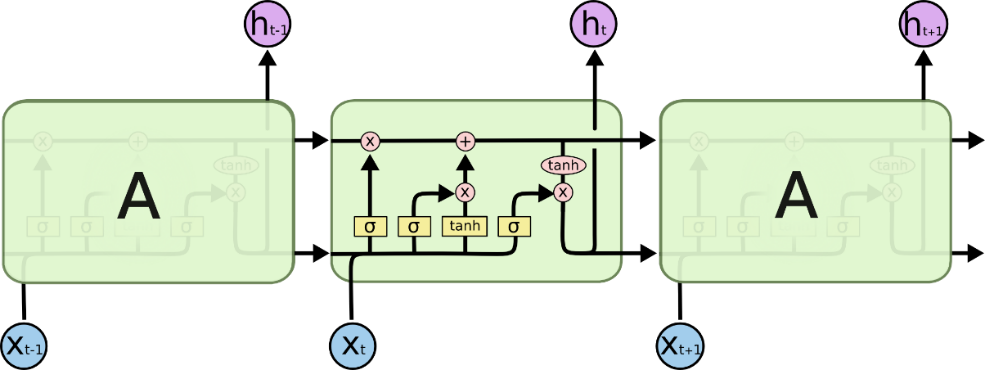
### Giới thiệu

Mạng LSTM là một loại mạng nơ-ron tuần hoàn. Trong RNN, đầu ra từ bước cuối cùng được cung cấp dưới dạng đầu vào trong bước hiện tại. LSTM được thiết kế bởi Hochreiter & Schmidhuber. LSTM giải quyết vấn đề phụ thuộc dài hạn của RNN. Theo mặc định, LSTM có thể giữ lại thông tin trong một khoảng thời gian dài. LSTM được sử dụng để xử lý, dự đoán và phân loại trên cơ sở dữ liệu chuỗi thời gian.

LSTM được thiết kế để tránh được vấn đề phụ thuộc xa (long-term dependency). Việc nhớ thông tin trong suốt thời gian dài là đặc tính mặc định của LSTM và không cần phải huấn luyện mô hình để có thể nhớ được.

### Kiến trúc

LSTM cũng có kiến trúc dạng chuỗi như mạng RNN, nhưng các mô-đun bên trong có cấu trúc khác với mạng RNN chuẩn. Thay vì chỉ có một lớp mạng nơ-ron, thì mạng LSTM có tới 4 lớp tương tác với nhau một cách rất đặc biệt.



Hình . Mạng LSTM

Các ký hiệu ở hình trên được giải thích như sau:

A black arrow pointing to the left

Description automatically generated

Hình . Giải thích các ký hiệu trong hình mạng LSTM

- Ô màu vàng được sử dụng để học trong mạng nơ-ron

- Hình tròn màu hồng biểu diễn các phép toán như phép cộng vector

- Đường hợp nhau kí hiệu việc kết hợp

- Đường rẽ nhánh kí hiệu việc nội dung được sao chép và di chuyển đến các nơi khác nhau

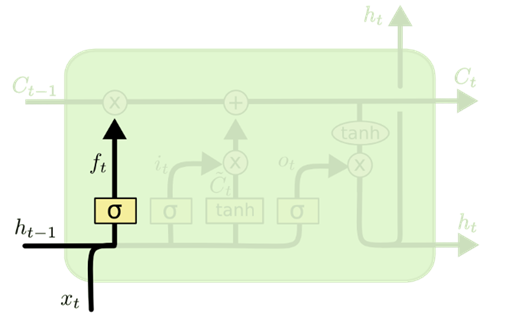
Cụ thể ở mỗi state thứ t của LSTM:



Hình . Một state của mạng LSTM

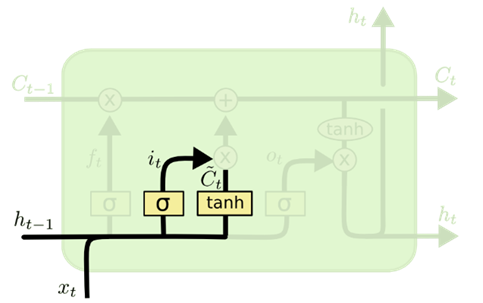
* Output: ct, ht ta gọi c là cell state, h là hidden state
* Input: ct-1, ht-1, xt. Trong đó, xt đóng vai trò là input ở state thứ t của model. ct-1, ht-1 là output của state trước đó. h ở đây có vai trò khá giống với a ở RNN, còn c là điểm mới của LSTM. (sigmoid), tanh là các activation functions. Tanh có dạng: tanh x = (). Phép nhân là element-wise multiplication, phép cộng là cộng các ma trận.
* ft, it, ot ứng với forget gate, input gate và output gate

Forget gate: (). Cổng này quyết định lượng thông tin từ state trước bị bỏ đi là bao nhiêu.



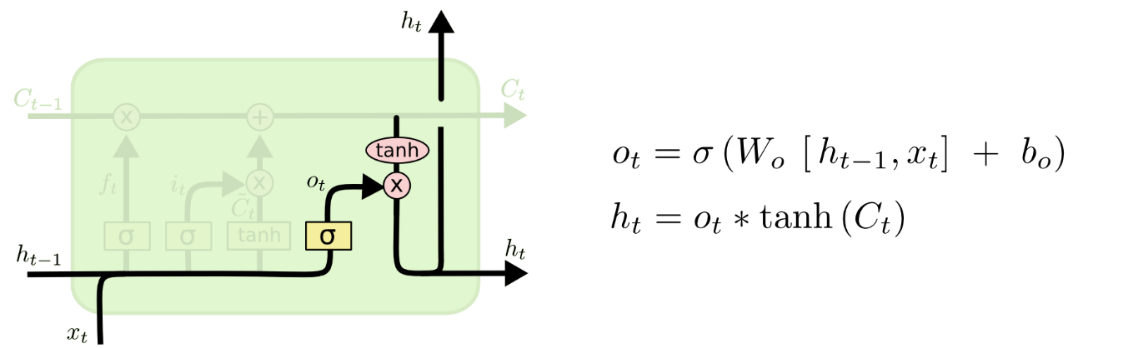
Hình .Cổng forget

Input gate: (). Cổng này quyết định lượng thông tin đầu vào ảnh hưởng đến state mới là bao nhiêu.



Hình . Cổng Input

Output gate: (). Cổng này điều chỉnh lượng thông tin có thể ra ngoài yt và lượng thông tin tới state tiếp theo.

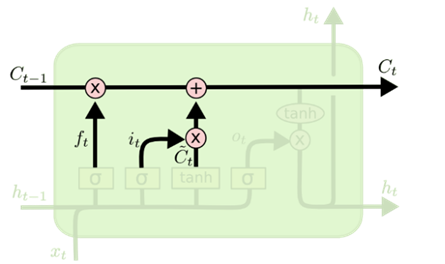


Hình . Cổng Output

Nhận xét 0 < ft, it, ot < 1 (giá trị của hàm sigmoid nằm trong khoảng [0;1]), bf, bi, bo là hệ số bias, W, U giống với RNN

(), ), giống tính at trong RNN.

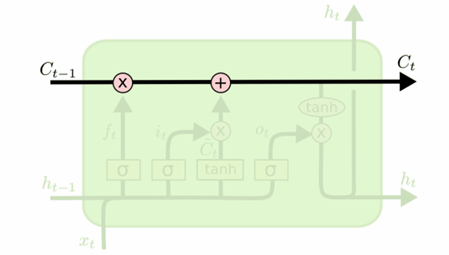
()



Hình . Giá trị state C

(), ngoài ra ht cũng được dùng để tính ra output yt cho state t.

ht, khá giống với RNN, nên model có short term memory. Trong khi đó ct giống như một băng chuyền ở mô hình RNN, thông tin nào cần quan trọng và dùng ở sau sẽ đc gửi vào và dùng khi cần => có thể mang đi xa => long term memory



Hình . Ct của LSTM

Tổng kết:

• LSTM giải quyết được phần nào vanishing gradient so với RNN.

• Quá trình huấn luyện của RNN chậm và LSTM chậm hơn RNN

• Tuy nhiên do được cải tiến hơn RNN, nên LSTM vẫn được sử dụng phổ biến

## Mô hình ARIMA

### Giới thiệu

ARIMA là viết tắt của Trung bình trượt tích hợp tự động hồi quy và đại diện cho nền tảng trong dự báo chuỗi thời gian. Đây là một phương pháp thống kê đã trở nên phổ biến rộng rãi nhờ tính hiệu quả của nó trong việc xử lý các cấu trúc thời gian tiêu chuẩn khác nhau có trong dữ liệu chuỗi thời gian. Model tổng quát của ARIMA được viết như sau:

Trong đó là giá trị sai phân bậc d và là các chuỗi nhiễu trắng.

### Kiến trúc

Bản chất của ARIMA:

* AR (Tự hồi quy): Đây là thành phần tự hồi quy bao gồm tập hợp các độ trễ của biến hiện tại. Độ trễ bậc p chính là giá trị lùi về quá khứ p bước thời gian của chuỗi. Độ trễ dài hoặc ngắn trong quá trình AR phụ thuộc vào tham số trễ p. Cụ thể AR(p) của chuỗi xt được biểu diễn như bên dưới:

* I (Tích hợp): Để đạt được một chuỗi thời gian dừng, một chuỗi không thể hiện xu hướng hoặc tính thời vụ, sai phân sẽ được áp dụng. Nó thường liên quan đến việc trừ một quan sát khỏi quan sát trước đó.

Sai phân bậc 1:

Sai phân bậc 2:

Trong đó là gia trị sai phân bậc d và là các chuỗi nhiễu trắng

* MA (Trung bình trượt): Thành phần này tập trung vào mối quan hệ giữa một quan sát và sai số dư từ mô hình trung bình trượt dựa trên các quan sát bị trễ.

Vế (1) có nghĩa rằng kỳ vọng của chuỗi bằng 0 để đảm bảo chuỗi dừng không có sự thay đổi về trung bình theo thời gian. Vế (2) là phương sai của chuỗi không đổi. Do kì vọng và phương sai không đổi nên chúng ta gọi phân phối của nhiễu trắng là phân phối xác định (identical distribution) và được ký kiệu là ~ WN(0, 2). Nhiễu trắng là một thành phần ngẫu nhiên thể hiện cho yếu tố không thể dự báo của model và không có tính quy luật. Quá trình trung bình trượt được biểu diễn theo nhiễu trắng như sau:

Mỗi thành phần này được chỉ định rõ ràng trong mô hình dưới dạng tham số. Ký hiệu tiêu chuẩn được sử dụng cho ARIMA(p,d,q) trong đó các tham số được thay thế bằng các giá trị số nguyên để nhanh chóng chỉ ra mô hình ARIMA cụ thể đang được sử dụng.

Các tham số của mô hình ARIMA được xác định như sau:

* p : Thứ tự độ trễ, biểu thị số lượng quan sát độ trễ được đưa vào mô hình.
* d : Mức độ sai phân, biểu thị số lần quan sát thô trải qua sai phân.
* q : Thứ tự của đường trung bình động, biểu thị kích thước của cửa sổ đường trung bình động.

## Phương pháp đánh giá mô hình

Sau khi đã xây dựng mô hình dự đoán, chúng ta cần các mô hình đánh giá xem mô hình của chúng ta đã xây dựng dự đoán đúng đến đâu:

* RMSE (Root Mean Squared Error) lỗi trung bình bình phương gốc là một phương pháp phổ biến để đánh giá độ chính xác của mô hình dự đoán. Nó là căn bậc hai của trung bình của bình phương của các sai số giữa giá trị dự đoán và giá trị thực tế. Nó cho biết các giá trị dự đoán khác nhau bao xa so với các giá trị quan sát được trong một mô hình. Công thức của RMSE như sau:

()

Trong đó:

* n là số lượng điểm dữ liệu
* yi là giá trị thực tế tại điểm dữ liệu thứ i
* là giá trị dự đoán tại điểm dữ liệu thứ i

RMSE là một cách hữu ích để xem mức độ phù hợp của một mô hình với tập dữ liệu. RMSE càng lớn thì chênh lệch giữa giá trị dự đoán và giá trị quan sát càng lớn, nghĩa là mô hình phù hợp với dữ liệu càng tệ. Ngược lại, RMSE càng nhỏ thì mô hình càng phù hợp với dữ liệu.

* MAPE (Mean Absolute Percentage Error) lỗi phần trăm tuyệt đối trung bình tính tỷ lệ phần trăm trung bình của sự sai lệch tuyệt đối giữa giá trị dự đoán và giá trị thực tế. Thường được sử dụng để dho độ chính xác dự đoán của mô hình.

Trong đó:

* Yi là giá trị thực tế
* là giá trị được dự đoán bởi mô hình
* n là số lượng quan sát

MAPE được sử dụng phổ biến vì nó dễ hiểu và dễ giải thích. Ví dụ: giá trị MAPE

là 12.8% có nghĩa là chênh lệch trung bình giữa giá trị dự đoán và giá trị thực tế là

12.8%.

Giá trị MAPE càng thấp thì mô hình càng có khả năng dự đoán các giá trị tốt

hơn. Ví dụ: một mô hình có MAPE là 5% sẽ chính xác hơn một mô hình có MAPE là

10%.

# BÀI TOÁN DỰ ĐOÁN GIÁ CHỨNG KHOÁN

## Chuẩn bị và phân tích dữ liệu

Dữ liệu được sử dụng trong bài toán là giá cổ phiếu của công ty Microsoft. Dữ liệu được lấy từ ngày 3/1/2000 đến ngày 31/12/2023 với 6037 mẫu, dữ liệu các ngày giao dịch không liên tục và bị gián đoạn vào các ngày nghỉ, cuối tuần.

Một phần của bộ dữ liệu:

A screenshot of a graph

Description automatically generated

Hình . Ví dụ một số mẫu thử nghiệm

Trong đó, ý nghĩa các cột như sau:

* Date: ngày giao dịch
* Open: giá mở cửa, cũng là giá đóng của của phiên giao dịch ngày liền trước
* High: giá cao nhất trong mộ phiên giao dịch hoặc trong một chu kỳ theo dõi biến động giá
* Low: giá thấp nhất trong một phiên giao dịch hoặc trong một chu kỳ theo dõi biến động giá
* Close: Giá đóng cửa là giá thực hiện tại lần khớp lệnh cuối cùng trong ngày giao dịch
* Volume: khối lượng giao dịch
* Dữ liệu được sử dụng là dạng dữ liệu time-series, nghĩa là dữ liệu thay đổi theo thời gian. Chu kỳ của dữ liệu là 1 ngày, ta xem mỗi ngày là 1 time-step.

## Xây dựng mô hình

Sau khi thu thập và phân tích dữ liệu trước khi tiến hành học máy. Dữ liệu lấy mẫu được chia thành 2 nhóm: tập dữ liệu huấn luyện được sử dụng để thiết lập các mô hình học máy, tập còn lại dùng để test sau khi đã huấn luyện xong.

Mô hình tổng quát được đề xuất để dự đoán:

A screenshot of a cell phone

Description automatically generated

Hình . Mô hình đề xuất

Với mô hình ARIMA:

* Chuẩn bị dữ liệu, chia dữ liệu
* Tìm kiếm các thông số p, d, q
* Xây dựng mô hình với các thông số vừa tìm được và thực hiện huấn luyện
* Trực quan hóa kết quả dự đoán
* Đánh giá mô hình

Với mạng RNN:

* Vì mỗi ngày là một time step, ta sẽ sử dụng 60 time steps làm input để đưa vào mạng train. Đầu ra là time step tiếp theo. Nghĩa là dùng giá 60 ngày để dự đoán giá của ngày kế tiếp)
* Scale dữ liệu đầu vào về khoảng [0;1]
* Xây dựng model với các layer input và output
* Thực hiện huấn luyện và test mô hình
* Trực quan hóa kết quả test
* Đánh giá mô hình

Với mạng LSTM: Tương tụ như mô hình RNN

* Vì mỗi ngày là một time step, ta sẽ sử dụng 60 time steps làm input để đưa vào mạng train. Đầu ra là time step tiếp theo. Nghĩa là dùng giá 60 ngày để dự đoán giá của ngày kế tiếp)
* Scale dữ liệu đầu vào về khoảng [0;1]
* Xây dựng model với các layer input và output
* Thực hiện huấn luyện và test mô hình
* Trực quan hóa kết quả test
* Đánh giá mô hình

## Dữ liệu thực nghiệm

Phân chia dữ liệu thành tập training và testing:

Với cả 3 mô hình thực hiện chia theo tỉ lệ 90% cho tập training và 10% cho tập testing

Dự liệu thực nghiệm được lấy chia theo thời gian, mỗi thực nghiệm được chia thành 2 phần, phần 1 với dữ liệu từ 01-01-2010 đến 31-12-2023 và phần 2 dữ liệu từ 01-01-2022 đến 31-12-2023. Với TSLA phần 1 bắt đầu từ 01-01-2014 đến 31-12-2023.

* Thực nghiệm 1 với cổ phiếu mã AAPL: Phần 1 với 3522 mẫu. Phần 2 501 mẫu
* Thực nghiệm 2 với cổ phiếu mã MSFT: Phần 1 với 3522 mẫu. Phần 2 501 mẫu.
* Thực nghiệm 3 với cổ phiếu mã TSLA: Phần 1 với 2516 mẫu. Phần 2 501 mẫu

## Môi trường thực nghiệm

* Processor: Intel(R) Core(TM) i5-9300H CPU @ 2.4GHz
* Memory RAM: 16GB
* System type: 64-bit operating system, x64-based processor
* Edition: Windows 10 Enterprise LTSC
* Các thử nghiệm được cài đặt và sử dụng ngôn ngữ Python trên môi trường Visual Code. Với các thư viện của Python như Numpy, Panda, Keras, Matplotlib, Seaborn, Sklearn, pmdarima.

## Xây dựng thực nghiệm

Đọc dữ liệu từ yfinace

import yfinance as yf

ticker = yf.Ticker("MSFT")

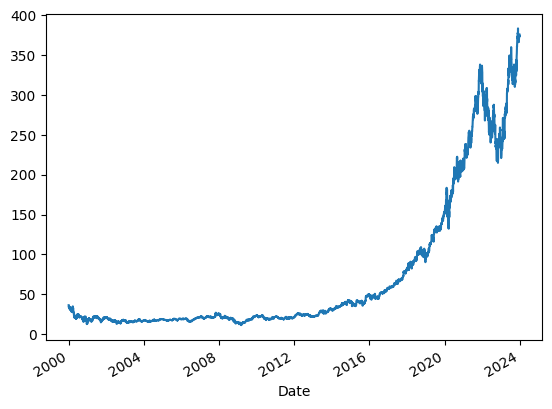
data = ticker.history(start="2000-01-01", end="2023-12-31")

A screenshot of a computer

Description automatically generated

Hình . Dữ liệu thực nghiệm

Mô hình hóa cột “Open”:



Hình . Biểu đồ cột Open

### Cấu trúc mô hình

Cấu trúc mô hình ARIMA

    # find parameters

        arima\_model = pm.auto\_arima(data, start\_p=0, start\_q=0,

                         test='adf',

                         max\_p=3, max\_q=3, m=1,

                         start\_P=0, seasonal=False,

                         d=None, D=0, trace=True,

                         error\_action='ignore',

                         suppress\_warnings=True,

                         stepwise=True)

        arima\_model.summary()

        model = sm.tsa.arima.ARIMA(data, order=arima\_model.order)

        model\_fit = model.fit()

Cấu trúc mô hình RNN

    model = keras.models.Sequential([

                keras.layers.SimpleRNN(20, return\_sequences=True, input\_shape=[None,1]),

                keras.layers.SimpleRNN(20, return\_sequences=True),

                keras.layers.SimpleRNN(20, return\_sequences=True),

                keras.layers.SimpleRNN(1)])

        model.add(Dense(1))

        model.compile(loss='mean\_squared\_error', optimizer='adam')

        model.fit(Xtrain, self.y\_train, epochs=1, batch\_size=1)

Cấu trúc mô hình LSTM

      # Build the LSTM model

        model = Sequential()

        model.add(LSTM(128, return\_sequences=True,

                       input\_shape= (Xtrain.shape[1], 1)))

        model.add(Dropout(0.2))

        model.add(LSTM(64, return\_sequences=False))

        model.add(Dropout(0.2))

        model.add(Dense(32))

        model.add(Dropout(0.2))

        model.add(Dense(1))

        # We are adding dropout to reduce overfitting

        # Compile the model

        model.compile(optimizer='adam', loss='mean\_squared\_error')

        # Train the model

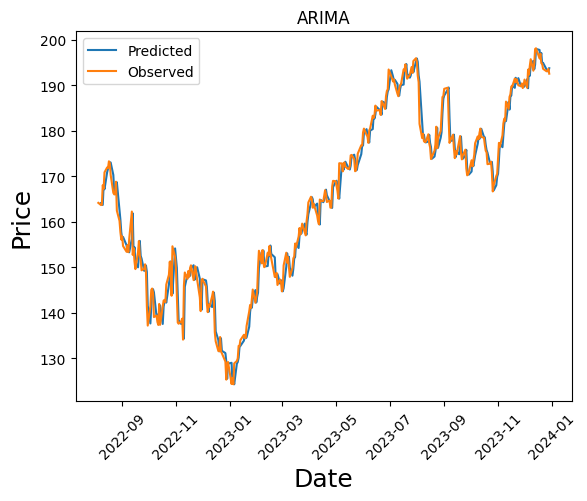
        model.fit(Xtrain, self.y\_train, batch\_size=32, epochs= 2)

### Thực nghiệm 1: Cổ phiếu Apple – AAPL.

**Phần 1:**

Thực hiện huấn luyện, dự đoán đánh giá lần lượt bằng các mô hình ARIMA, RNN, LSTM, ta thu được kết quả.

**ARIMA:**

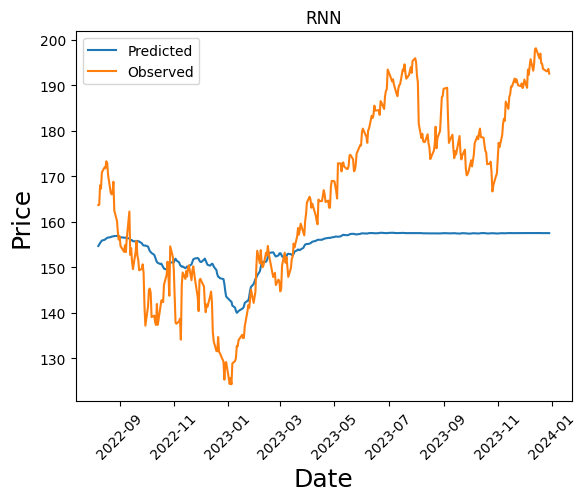
****

Hình . Kết quả ARIMA thực nghiệm 1 phần 1.

Các chỉ số:

* RMSE: 2.567319880920929
* MAPE: 0.01197033499466877

**RNN:**



Hình . Kết quả RNN thực nghiệm 1 phần 1

Các chỉ số:

* RMSE: 19.200762735415637
* MAPE: 0.08940961982514764

**LSTM:**

A graph with blue and orange lines

Description automatically generated

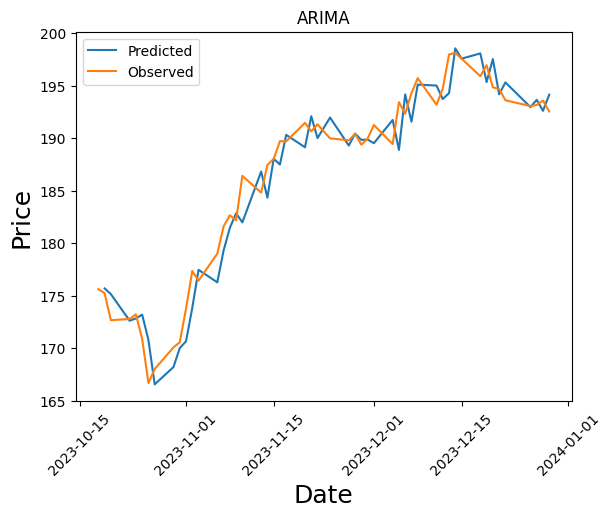
Hình . Kết quả LSTM thực nghiệm 1 phần 1

Các chỉ số:

* RMSE: 5.46069882162589
* MAPE: 0.027253900607681084

**Phần 2:**

**ARIMA:**

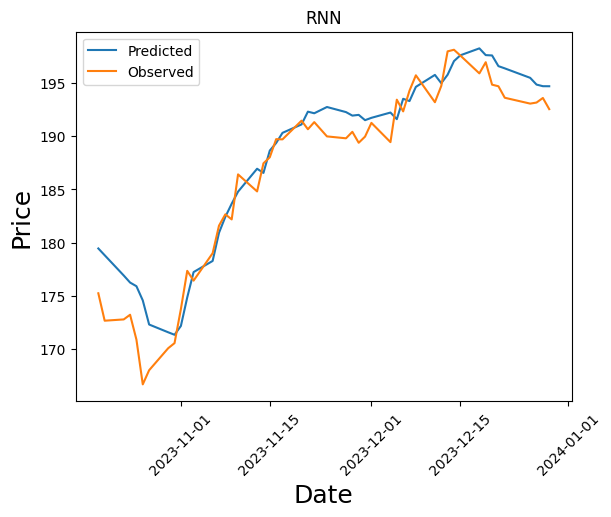


Hình . Kết quả ARIMA thực nghiệm 1 phần 2

Các chỉ số:

* RMSE: 2.0086015170413085
* MAPE: 0.008751356749571787

**RNN:**

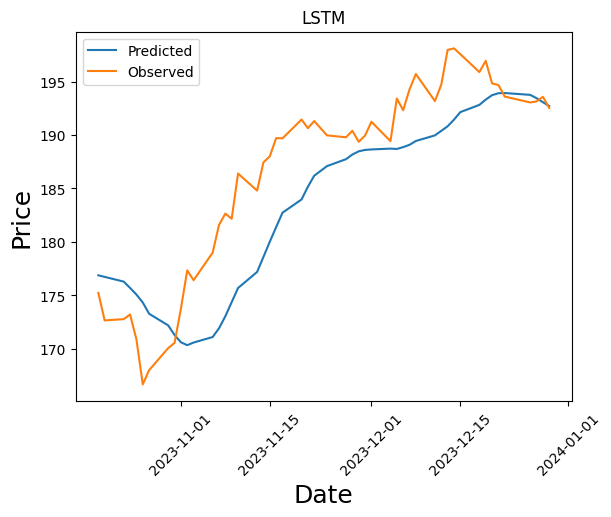


Hình . Kết quả RNN thực nghiệm 1 phần 2

Các chỉ số:

* RMSE: 2.4801619839912536
* MAPE: 0.010674020435027663

**LSTM:**



Hình . Kết quả mô hình LSTM thực nghiệm 1 phần 2

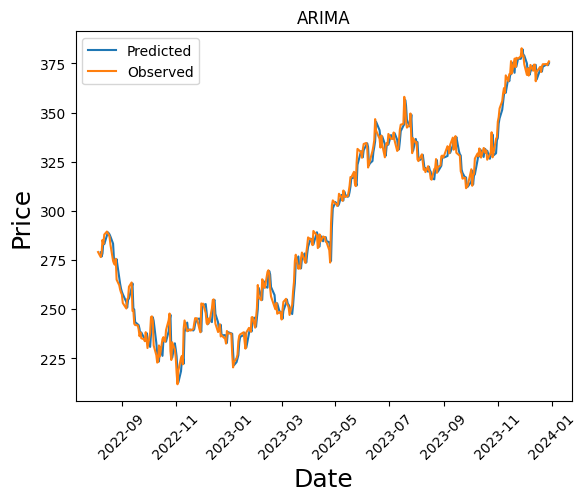
Các chỉ số:

* RMSE: 5.2918859687695
* MAPE: 0.023821904299138507

### Thực nghiệm 2: Cổ phiếu Microsoft– MSFT

Phần 1:

**ARIMA**

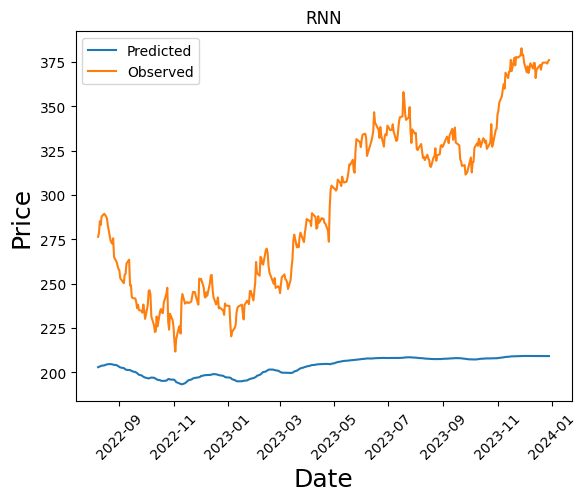


Hình . Kết quả ARIMA thực nghiệm 2 phần 1

Các chỉ số:

* RMSE: 4.95817616822027
* MAPE: 0.01337559120449731

**RNN**

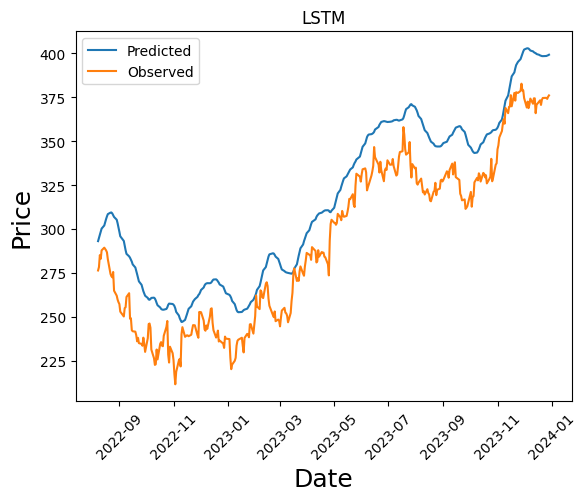
****

Hình . Kết quả RNN thực nghiệm 2 phần 1

Các chỉ số:

* RMSE: 99.42042648789118
* MAPE: 0.29061659249861754

**LSTM**

****

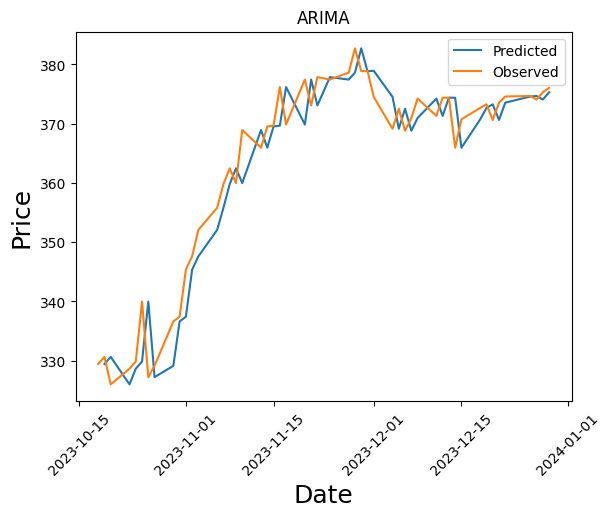
Hình . Kết quả LSTM thực nghiệm 2 phần 1

Các chỉ số:

* RMSE: 24.92031483673187
* MAPE: 0.08132100904220524

Phần 2:

**ARIMA**

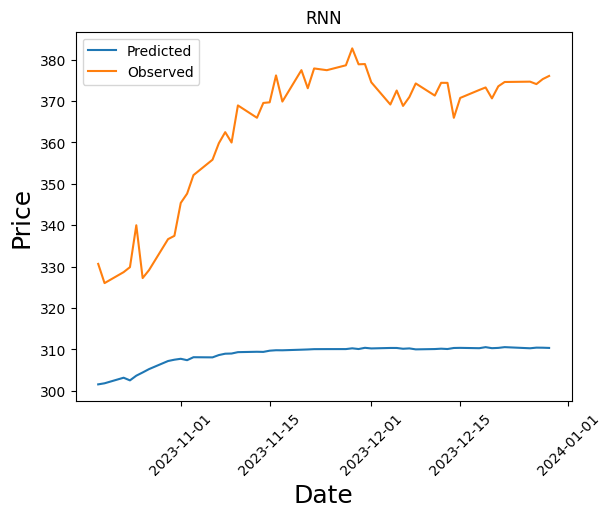
****

Hình . Kết quả mô hình ARIMA thực nghiệm 2 phần 2

Các chỉ số:

* RMSE: 4.500766502984545
* MAPE: 0.009863336098612339

**RNN**

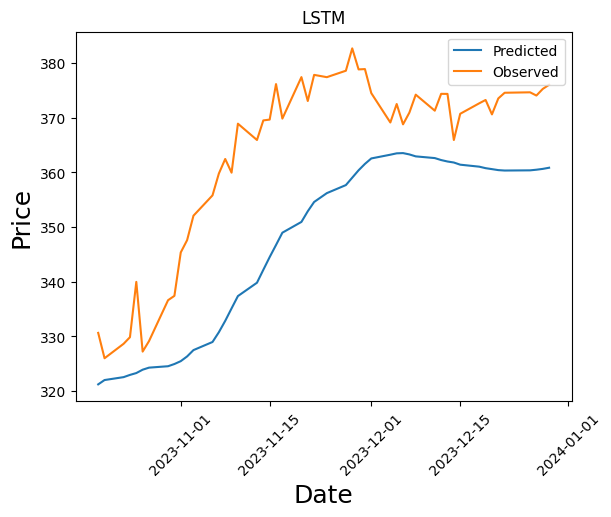
****

Hình . Kết quả mô hình RNN thực nghiệm 2 phần 2

Các chỉ số:

* RMSE: 56.40102506639573
* MAPE: 0.1484826725292435

**LSTM**



Hình . Kết quả LSTM thực nghiệm 2 phần 2

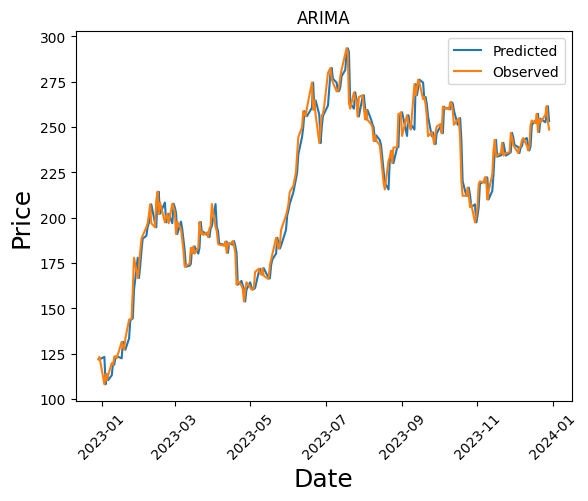
Các chỉ số:

* RMSE: 17.864610364174005
* MAPE: 0.04380147134372017

### Thực nghiệm 3: Cổ biếu Tesla – TSLA

Phần 1:

**ARIMA**

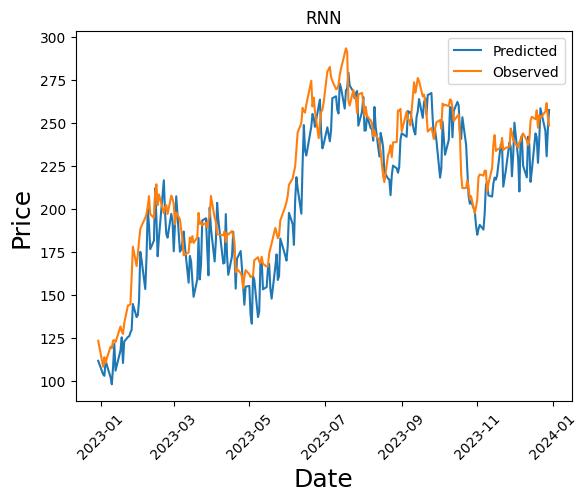
****

Hình . Kết quả ARIMA thực nghiệm 3 phần 1

Các chỉ số:

* RMSE: 7.1124416084077415
* MAPE: 0.02567074391818049

**RNN**

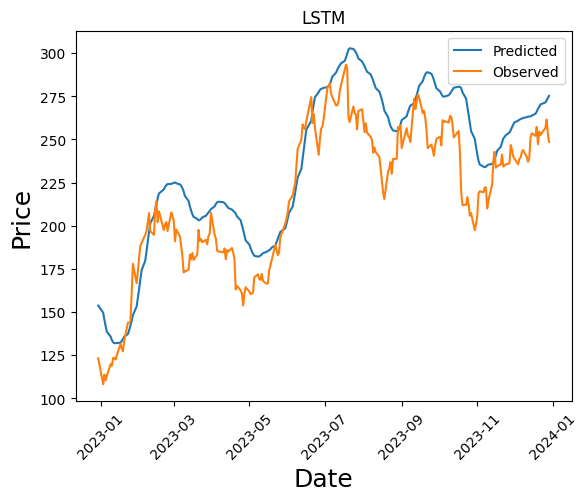
****

Hình . Kết quả RNN thực nghiệm 3 phần 1

Các chỉ số:

* RMSE: 18.629472163028087
* MAPE: 0.07345415221196877

**LSTM**

****

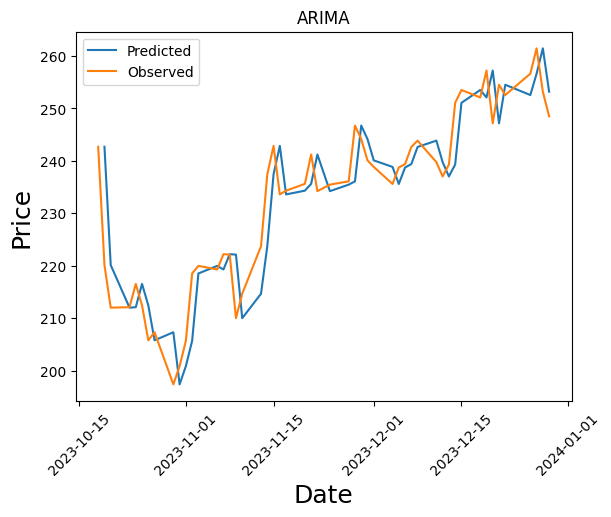
Hình . Kết quả LSTM thực nghiệm 3 phần 1

Các chỉ số:

* RMSE: 24.943493756578153
* MAPE: 0.1014670707366399

Phần 2:

**ARIMA**

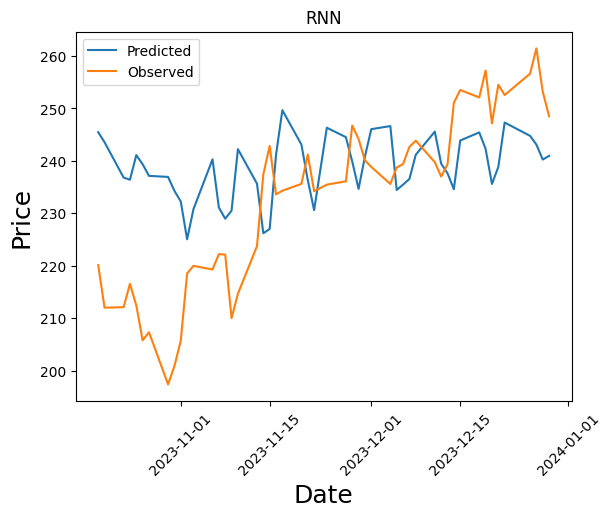
****

Hình . Kết quả ARIMA thực nghiệm 3 phần 2

Các chỉ số:

* RMSE: 6.885704663062172
* MAPE: 0.022761922599664104

**RNN**

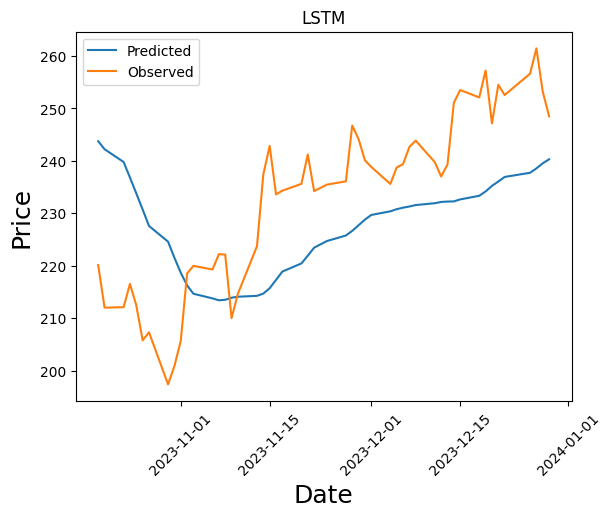
****

Hình . Kết quả RNN thực nghiệm 3 phần 2

Các chỉ số:

* RMSE: 16.761353414472836
* MAPE: 0.06138932910418322

**LSTM**



Hình . Kết quả LSTM thực nghiệm 3 phần 2

Các chỉ số:

* RMSE: 16.425729400700632
* MAPE: 0.06357902818752464

### Kết luận

Từ thực nghiệm cho thấy các mô hình đều hoạt động tốt. Đối với trường hợp ít dữ liệu 1 năm cho thấy mô hình ARIMA cho kết quả tốt hơn 2 mô hình còn lại, RNN tốt hơn LSTM. Trường hợp ít dữ liệu đầu vào hơn đối với cả 3 mã chứng khoán: mô hình ARIMA tốt hơn hẳn, sau đó đén RNN và cuối cùng là LSTM. Trường hợp dữ liệu nhiều ARIMA vẫn cho kết quả tốt, cùng với đó là LSTM đã cho thấy sự khác biệt độ chính xác, sai số cải thiện tốt hơn hẳn, RNN thì đã bị tụt xuống. Cho thấy việc nhớ lâu của LSTM hiệu quả trong trường hợp nhiều dữ liệu huấn luyện.

Nhìn chung, kết quả chứng minh ARIMA, LSTM là những mô hình tốt để dự đoán dữ liệu chuỗi thời gian. Mô hình ARIMA cho ra kết quả có sai số thấp hơn hai mô hình còn lại. Mô hình RNN và LSTM tuy sai số cao hơn nhưng ưu điểm là có thể huấn luyện và dự đoán trên bất kì chuỗi nào (nhờ khả năng xử lý và lưu trữ của mạng neuron). Mô hình ARIMA khi dự đoán cần chuỗi phải dừng điều này là yêu cầu khó khăn khi dự đoán một thứ khó đoán như giá chứng khoán. Mô hình LSTM-RNN có ưu điểm về khả năng huấn luyện dự báo với các chuỗi khác nhau nên khả năng đáp ứng dự báo với giá chứng khoán sẽ tốt hơn ARIMA.

Vì vậy để có thể dự đoán với giá chứng khoán trong thực tế, em sẽ chọn mô hình LSTM và ARIMA để dự báo giá chứng khoán. Đây là lựa chọn an toàn khi dựa trên tính chất nhớ lâu của mô hình và mô hình truyền thống hơn đã được chứng minh. Giúp cho kết quả có sự so sánh khách quan nhất có thể.

# Xây dựng hệ thống dự báo chứng khoán

## Đặt vấn đề

Trong thời đại hiện nay, việc có các khoản đầu tư khác là một cách kiếm tiền hết sức đúng đắn. Trong đó có đầu tư về tài chính, chứng khoán, cổ phiếu. Tuy nhiên việc kiếm tiền từ thị trường chứng khoán không phải chuyện dễ dàng. Vì vậy để có thể chắc chắn hơn trong việc sinh lời chúng ta cần có nhiều kiến thức và dự báo về thị trường này hơn.

Để giải quyết vấn đề này, em đã xây dựng một hệ thống cung cấp thông tin về thị trường chứng khoán và những chỉ số dự đoán tham khảo cho mọi người. Giúp con đường tìm kiếm lợi nhuận bớt phần khó khăn.

## Phân tích yêu cầu hệ thống

Yêu cầu chức năng:

* Xem giá chứng khoán của một mã cụ thể
* Xem thông tin giá cả dự đoán tham khảo.

## Sơ đồ Use-Case

### Sơ đồ use-case xem thông tin chứng khoán

A diagram of a company

Description automatically generated

Hình . Sơ đồ use case tổng quát

Mô tả chi tiết:

- Đối tượng sử dụng (actor): Người dùng.

- Use case xem thông tin về mã cổ phiếu.

- Các bước thực hiện:

+ Người dùng truy cập trang web chọn mã cổ phiếu muốn xem

### Sơ đồ use-case xem thông tin dự đoán giá chứng khoán

A diagram of a company

Description automatically generated

Hình . Sơ đồ use xem giá dự đoán

Mô tả chi tiết:

- Đối tượng sử dụng (actor): Người dùng.

- Use case xem thông tin về mã cổ phiếu dự đoán

- Các bước thực hiện:

+ Người dùng truy cập trang web chọn mã cổ phiếu muốn xem, chọn thời gian muốn dự đoán

## Kết quả

### Màn hình trang chủ

Đầu tiên người dùng truy cập trang web, sẽ nhận được giao diện như sau. Trên trang web hiển thị ngày bắt đầu và kết thúc của giá cổ phiếu được hiển thị. Bên dưới là chọn mã cổ phiếu. Sau đó là chọn thời gian dự đoán.

A screenshot of a computer

Description automatically generated

Hình .Màn hình trang chủ

Sau đó kéo xuống dưới để xem biểu đồ thông tin thể hiện giá.

A graph of a graph

Description automatically generated with medium confidence

Hình . Màn hình trang chủ với biểu đồ

### Màn hình dự đoán chứng khoán

Người dùng bấm vào tab ‘Forecasting’ để xem thông tin dự báo. Đầu tiên xem thông tin dự báo của mô hình ARIMA. Trang này hiển thị danh sách giá dự đoán trong tương lai và biểu đồ hiển thị về giá tương ứng với giá dự đoán

A screenshot of a computer

Description automatically generated

Hình .Màn hình dự đoán bằng ARIMA

Tiếp theo bấm vào bút ARIMA để hiển thị ra danh sách mô hình. Chọn mô hình LSTM để nhận thông tin dự báo của mô hình LSTM.

A graph with a line and a person pointing at it

Description automatically generated with medium confidence

Hình . Màn hình dự đoán bằng LSTM

# Kết luận

## Kết quả đạt được

Về bản thân:

- Nâng cao kỹ năng đọc - hiểu, tìm kiếm tài liệu.

- Nâng cao khả năng tự học, nghiên cứu và tìm cách giải quyết vấn đề.

- Học được thêm nhiều kiến thức mới và các ứng dụng.

- Biết cách trình bày báo cáo một cách chính xác, rõ ràng, khoa học.

- Kỹ năng sắp xếp, phân chia thời gian biểu hợp lý.

Về bài nghiên cứu:

- Trong bài nghiên cứu này, đã thực nghiệm được 3 mô hình với mỗi mô hình là 3 thực nghiệm. Kết quả cho thấy mô hình ARIMA cho kết quả tốt nhất trong bộ dữ liệu. Sau đó là mô hình LSTM cho kết quả với trường hợp dữ liệu đầu vào lớn.

- Kết quả của thực nghiệm mang tính chính xác tương đối, vì khoảng cách giữa giá trị thật và giá trị dự đoán còn khá xa.

- Kết quả của các mô hình được đánh giá dựa trên các phương pháp khác nhau đã làm nổi bật lên điểm mạnh của mô hình được đề xuất. Tuy nhiên, các kết quả thực nghiệm chưa thể ứng dụng vào đời sống mà chỉ để phục vụ nghiên cứu. Vì thực nghiệm mang tính chủ quan, trên thực tế thì giá của cổ phiếu còn bị ảnh hưởng bởi nhiều yếu tố khách quan khác. Đòi hỏi người xây dựng mô hình có kiến thức sâu rộng hơn về mảng tài chính để áp dụng công nghệ vào. Từ đó mới đưa ra những dự đoán đáng tin cậy.

## Hạn chế

- Kiến thức về mảng chứng khoán còn hạn hẹp nên chưa thể xây dựng được mô hình có những yếu tố ảnh hưởng tới kết quả dự đoán.

- Dữ liệu còn hạn chế chuỗi thời gian liên tục và chưa phong phú.

- Phạm vi sử dụng kết quả còn hạn chế, độ chính xác chưa cao

- Chưa đánh giá toàn diện được các phương pháp thử nghiệm.

## Hướng phát triển

- Đây là một hướng đi nhiều tiềm năng phát triển trong tương lai. Đòi hỏi người xây dựng mô hình dự đoán không chỉ có kiến thức về công nghệ thông tin, mà còn cần trau dồi thêm kiến thức về tài chính để phát triển và xây dựng mô hình một cách tốt nhất.

- Có thể thử nghiệm trên các mô hình học máy khác và so sánh để cho ra cái nhìn tổng quan hơn về việc dự đoán giá chứng khoán, đồng thời tìm ra mô hình tối ưu cho bài toán dự đoán giá chứng khoán.

- Thêm các yếu tố có thể ảnh hưởng đến dự đoán vào mô hình để kết quả dự đoán mang tính chính xác cao hơn. Như yếu tố cảm xúc của nhà đầu tư, mức độ ảnh hưởng của tin tức đến giá chứng khoán.

- Dự đoán giá chứng khoán là việc giúp giảm thiểu rủi do và gia tăng lợi nhuận. Có thể thử nghiệm để tạo ra mô hình dự đoán trên nhiều mã chứng khoán từ đó giúp tìm ra chiến lược đầu tư hợp lý.

# THAM KHẢO

[1] *Giáo trình Thị trường chứng khoán*, chủ biên: TS. Bạch Đức Hiển, năm 2009, nhà xuất bản Tài Chính, chương 1, 2.

[2*] Master ML Algorithms*, Jason Brownlee, © Copyright 2016 Jason Brownlee. All Rights Reserved. Chapter III.13.

[3] *Introduction to ML with Python*: A guide for Data Scientists, Andreas C. Müller & Sarah Guido, Printed in the United States of America, Published by O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472. Chapter 1.

[4] *Giáo trình Python cơ bản*, TS. Nguyễn Văn Hậu – TS. Nguyễn Duy Tân – ThS. Nguyễn Thị Hải Năng – ThS. Nguyễn Hoàng Hiệp, năm 2019, nhà xuất bản Đại học Quốc gia Hà Nội. Bài 1.

[5] Libraries in Python, article contributed by: parthmanchanda81, last update: 18 Oct, 2021. Published in GeeksforGeeks. Link: Libraries in Python - GeeksforGeeks

[6] Python Keras | keras.utils.to\_categorical(), article contributed by: manmayi, last update: 23 Jun, 2021. Published in GeeksforGeeks. Link: Python Keras | keras.utils.to\_categorical() - GeeksforGeeks

[7] Introduction to Seaborn – Python, article contributed by: 09amit, last update: 03 Jun, 2020. Published in GeeksforGeeks. Link: Introduction to Seaborn - Python - GeeksforGeeks

[8] Article “Implement Logistic Regression from scractch in Python”, by Casper Hansen. Published February 14, 2022. Link: Implementing logistic regression from scratch in Python - IBM Developer

[9] Long Short-term Memory. Link: <https://www.researchgate.net/publication/13853244_Long_Short-term_Memory>

[10] mape-python. Link: <https://www.statology.org/mape-python/>

[11] rmse-python.Link: <https://www.statology.org/rmse-python/>

[12] Time Series Forecasting with RNN. Link: <https://www.kaggle.com/code/hozler/time-series-forecasting-with-rnn>

[13] How to Make Out-of-Sample Forecasts with ARIMA in Python. Link: <https://machinelearningmastery.com/make-sample-forecasts-arima-python/>

[14] Microsoft Stock EDA-SMA-EWM-ARIMA-SARIMAX-LSTM. Link: <https://www.kaggle.com/code/aleynakrmz/microsoft-stock-eda-sma-ewm-arima-sarimax-lstm>

[15] Dự Đoán Giá Chứng Khoán SP500 Sử Dụng LSTM. Link: <https://www.phamduytung.com/blog/2018-11-10-stock-prediction_v1/>

[16] pmdarima: ARIMA estimators for Python. Link: <https://alkaline-ml.com/pmdarima/index.html>

[17] Mô hình ARIMA trong time series. Link: <https://phamdinhkhanh.github.io/2019/12/12/ARIMAmodel.html>