TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI

TRƯỜNG KINH TẾ

LÊ NGỌC QUYỀN

KHÓA LUẬN TỐT NGHIỆP

ĐẠI HỌC NGÀNH PHÂN TÍCH DỮ LIỆU KINH DOANH

TÊN ĐỀ TÀI: ỨNG DỤNG HỌC MÁY VÀ HỌC SÂU ĐỂ DỰ BÁO GIÁ CỔ PHIẾU CỦA TỔNG CÔNG TY VIGLACERA

HÀ NỘI - 2025

TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI

TRƯỜNG KINH TẾ

KHÓA LUẬN TỐT NGHIỆP

ĐẠI HỌC NGÀNH PHÂN TÍCH DỮ LIỆU KINH DOANH

TÊN ĐỀ TÀI: ỨNG DỤNG HỌC MÁY VÀ HỌC SÂU ĐỂ DỰ BÁO GIÁ CỔ PHIẾU CỦA TỔNG CÔNG TY VIGLACERA

Họ và tên sinh viên: Lê Ngọc Quyền

Mã sinh viên: 2021600459

Lớp: DHDLKD01 – K16

Giáo viên hướng dẫn: TS. Dương Thị Hoàn

HÀ NỘI – 2025

LỜI CẢM ƠN

Khóa luận tốt nghiệp với đề tài "Ứng dụng học máy và học sâu để dự báo giá cổ phiếu VGC" là kết quả của quá trình học tập, nghiên cứu nghiêm túc và được định hướng rõ ràng dưới sự hướng dẫn tận tình của giảng viên và sự hỗ trợ từ nhà trường. Đề tài tập trung vào việc thu thập, xử lý và phân tích dữ liệu tài chính thực tế, kết hợp các phương pháp thống kê, học máy và kỹ thuật dự báo hiện đại để đánh giá xu hướng và dự đoán giá cổ phiếu trong ngắn hạn.Quá trình thực hiện khóa luận đã giúp em củng cố kỹ năng chuyên môn, rèn luyện tư duy logic và tiếp cận sát hơn với các yêu cầu thực tế trong lĩnh vực phân tích dữ liệu tài chính. Đồng thời, đây cũng là cơ hội để em tự đánh giá năng lực bản thân, phát hiện những thiếu sót cần bổ sung nhằm hoàn thiện hành trang cho chặng đường nghề nghiệp sau này.

Em xin gửi lời cảm ơn chân thành tới các thầy cô giáo Trường Đại học Công nghiệp Hà Nội, đặc biệt là Khoa Kinh Doanh Số – nơi đã trang bị cho em nền tảng kiến thức vững chắc và môi trường học tập chuyên nghiệp. Em đặc biệt biết ơn cô Dương Thị Hoàn – giảng viên hướng dẫn – người đã luôn tận tình chỉ dẫn, định hướng và đồng hành cùng em trong suốt quá trình thực hiện đề tài này.

Dù đã nỗ lực để hoàn thiện khóa luận một cách chỉn chu và khoa học, em vẫn không tránh khỏi những thiếu sót nhất định. Em rất mong nhận được sự góp ý từ quý thầy cô để nâng cao hơn nữa chất lượng nghiên cứu và khả năng học thuật của bản thân.

Em xin chân thành cảm ơn!

MỤC LỤC

DANH MỤC HÌNH

DANH MỤC BẢNG

DANH MỤC TỪ VIẾT TẮT

STT	Chữ viết tắt	Nghĩa viết tắt
1	RMSE	Căn bậc hai sai số bình phương trung bình
2	MAE	Sai số tuyệt đối trung bình
3	MAPE	Sai số phần trăm tuyệt đối trung bình
4	CPI	Chỉ số giá tiêu dùng
5	R&D	Nghiên cứu và phát triển

LỜI CẢM ƠN

Đề tài hướng đến việc thu thập, xử lý và trực quan hóa dữ liệu tài chính, từ đó áp dụng các phương pháp thống kê, học máy và kỹ thuật dự báo hiện đại nhằm đánh giá xu hướng giá cổ phiếu VGC trong quá khứ và dự đoán biến động trong tương lai gần. Thông qua công cụ lập trình Python và các thư viện chuyên biệt như Pandas, Matplotlib, Scikit-learn và TensorFlow, đề tài là minh chứng cho khả năng tích hợp kiến thức giữa phân tích dữ liệu, lập trình và kinh tế tài chính.

Quá trình thực hiện khóa luận đã giúp em củng cố kỹ năng chuyên môn, rèn luyện tư duy logic và tiếp cận sát hơn với các yêu cầu thực tế trong lĩnh vực phân tích dữ liệu tài chính. Đồng thời, đây cũng là cơ hội để em tự đánh giá năng lực bản thân, phát hiện những thiếu sót cần bổ sung nhằm hoàn thiện hành trang cho chặng đường nghề nghiệp sau này.

Em xin gửi lời cảm ơn chân thành tới các thầy cô giáo Trường Đại học Công nghiệp Hà Nội, đặc biệt là Khoa Kinh Doanh Số – nơi đã trang bị cho em nền tảng kiến thức vững chắc và môi trường học tập chuyên nghiệp. Em đặc biệt biết ơn cô Dương Thị Hoàn – giảng viên hướng dẫn – người đã luôn tận tình chỉ dẫn, định hướng và đồng hành cùng em trong suốt quá trình thực hiện đề tài này.

Dù đã nỗ lực để hoàn thiện khóa luận một cách chỉn chu và khoa học, em vẫn không tránh khỏi những thiếu sót nhất định. Em rất mong nhận được sự góp ý từ quý thầy cô để nâng cao hơn nữa chất lượng nghiên cứu và khả năng học thuật của bản thân.

Em xin chân thành cảm ơn!

PHẦN MỞ ĐẦU

❖ Tính cấp thiết của đề tài

Trong quá trình tìm hiểu và nghiên cứu về thị trường chứng khoán, nhận thấy rằng dự báo giá cổ phiếu luôn là một bài toán quan trọng, vừa mang tính thử thách, vừa có giá trị thực tiễn cao. Việc xây dựng các mô hình dự báo chính xác không chỉ giúp nhà đầu tư đưa ra quyết định kịp thời mà còn là cơ sở để phát triển các công cụ hỗ trợ giao dịch thông minh.

Cổ phiếu của Tổng Công ty Viglacera - CTCP (mã VGC) được lựa chọn làm đối tượng nghiên cứu vì đây là một mã cổ phiếu có thanh khoản ổn định, dữ liệu lịch sử phong phú và đại diện cho nhóm ngành xây dựng – bất động sản. Những đặc điểm này giúp việc xây dựng và kiểm thử mô hình dự báo trở nên khả thi và có tính ứng dụng thực tế.

Hiện nay, các phương pháp truyền thống như mô hình hồi quy hay phân tích kỹ thuật tuy vẫn được sử dụng phổ biến nhưng còn hạn chế khi xử lý chuỗi dữ liệu có tính phi tuyến và biến động phức tạp như giá cổ phiếu. Trong khi đó, các mô hình hiện đại như Random Forest và LSTM đã chứng minh được khả năng học và phát hiện quy luật trong dữ liệu lịch sử, đặc biệt phù hợp với các bài toán chuỗi thời gian.

Với mong muốn được tiếp cận, áp dụng và so sánh hiệu quả giữa hai phương pháp này trong bối cảnh thực tế tại Việt Nam, quyết định lựa chọn đề tài: "Ứng dụng học máy và học sâu để dự báo giá cổ phiếu của Tổng Công ty Viglacera" nhằm kiểm nghiệm tính hiệu quả và khả năng ứng dụng của các mô hình dự báo hiện đại trong bối cảnh thực tiễn.

❖ Tổng quan nghiên cứu

Trong bối cảnh thị trường chứng khoán Việt Nam ngày càng phát triển, bài toán dự báo giá cổ phiếu đang nhận được nhiều sự quan tâm từ các nhà nghiên cứu và nhà đầu tư. Việc xây dựng các mô hình dự báo chính xác không chỉ hỗ trợ nhà đầu tư đưa ra quyết định hiệu quả mà còn góp phần vào việc phát triển các hệ thống hỗ trợ ra quyết định trong quản trị rủi ro tài chính. Đặc biệt, với những doanh nghiệp lớn như Tổng Công ty Viglacera - CTCP (VGC), việc nghiên cứu và dự báo giá cổ phiếu không chỉ có ý nghĩa học thuật mà còn mang tính thực tiễn cao trong lĩnh vực đầu tư chứng khoán.

Trong các nghiên cứu trước đây, nhiều phương pháp dự báo truyền thống đã được áp dụng như Hồi quy tuyến tính, ARIMA, hay GARCH nhằm mô hình hóa xu hướng và độ biến động của giá cổ phiếu. Tuy nhiên, các phương pháp thống kê cổ điển này thường gặp khó khăn trong việc xử lý những mối quan hệ phi tuyến, phức tạp và tác động đan xen từ nhiều yếu tố kinh tế vĩ mô, vi mô đến thị trường tài chính.

Để khắc phục những hạn chế này, các mô hình học máy (Machine Learning) và học sâu (Deep Learning) đã dần được đưa vào ứng dụng trong nghiên cứu tài chính. Trong đó, Random Forest (Breiman, 2001) nổi bật với khả năng xử lý phi tuyến tốt, kháng nhiễu, hạn chế overfitting và đặc biệt hiệu quả khi làm việc với dữ liệu dạng bảng và nhiều chỉ báo kỹ thuật. Nhiều nghiên cứu gần đây như Patel et al. (2015), Chong et al. (2017) đã chứng minh Random Forest có thể dự báo giá cổ phiếu với độ chính xác vượt trội so với các mô hình truyền thống.

Song song với đó, học sâu — đặc biệt là Long Short-Term Memory (LSTM) do Hochreiter & Schmidhuber (1997) đề xuất — đã chứng minh năng lực vượt trội khi xử lý các chuỗi dữ liệu tài chính có phụ thuộc dài hạn. Nhờ cơ chế bộ nhớ và các cổng kiểm soát thông tin, LSTM có thể học được các mẫu ẩn phức tạp và các chu kỳ trong dữ liệu giá cổ phiếu. Fischer & Krauss (2018) cùng Nelson et al. (2017) đều ghi nhận hiệu quả cao của LSTM trong dự báo chỉ số chứng khoán ở thị trường Mỹ và Brazil.

Đối với thị trường Việt Nam, số lượng nghiên cứu ứng dụng đồng thời hai mô hình học máy và học sâu trên cùng một bộ dữ liệu cổ phiếu cụ thể, như trường hợp cổ phiếu VGC của Viglacera, vẫn còn hạn chế. Điều này mở ra khoảng trống nghiên cứu cần thiết nhằm đánh giá, kiểm chứng và so sánh hiệu quả giữa hai hướng tiếp cận. Việc thực hiện song song hai mô hình Random Forest và LSTM trong dự báo giá cổ phiếu Viglacera sẽ góp phần bổ sung bằng chứng thực nghiệm, làm rõ điểm mạnh, hạn chế của từng phương pháp trong bối cảnh dữ liệu tài chính thực tiễn tại Việt Nam.

❖ **Mục tiêu nghiên cứu**

Trong bối cảnh thị trường chứng khoán biến động mạnh mẽ, việc dự báo giá cổ phiếu ngày càng trở nên quan trọng, đặc biệt là đối với các nhà đầu tư cá nhân và tổ chức cần ra quyết định trong thời gian ngắn. Dự báo giá cổ phiếu không chỉ giúp nhà đầu tư nhận diện xu hướng thị trường, mà còn đóng vai trò như một công cụ định hướng trong quản lý danh mục đầu tư, tối ưu hóa lợi nhuận và giảm thiểu rủi ro. Do đó, bài

toán này ngày càng thu hút nhiều sự quan tâm từ giới nghiên cứu, đặc biệt là khi các mô hình trí tuệ nhân tạo ngày càng phát triển mạnh mẽ.

Từ thực tiễn đó, xác định mục tiêu chính của khóa luận là xây dựng và đánh giá hiệu quả của hai mô hình hiện đại – Random Forest và LSTM – trong việc dự báo giá cổ phiếu của Tổng Công ty Viglacera – CTCP (mã VGC). Đây là hai phương pháp đại diện cho hai hướng tiếp cận khác nhau trong học máy: mô hình Random Forest phù hợp với dữ liệu dạng bảng và khai thác tốt các đặc trưng được thiết kế thủ công, trong khi đó LSTM lại nổi bật trong xử lý chuỗi thời gian với khả năng ghi nhớ dài hạn và tự học mối quan hệ giữa các bước dữ liệu liên tiếp.

Trên cơ sở đó, mục tiêu cụ thể của đề tài được triển khai theo các định hướng sau:

Thứ nhất, tiến hành thu thập và xử lý dữ liệu lịch sử giao dịch cổ phiếu VGC, bao gồm các thông tin về giá đóng cửa, khối lượng giao dịch và một số đặc trưng kỹ thuật thường dùng trong phân tích thị trường chứng khoán. Bước này đóng vai trò tiền đề để xây dựng bộ dữ liệu đầu vào có chất lượng, phục vụ cho việc huấn luyện mô hình.

Thứ hai, tiến hành xây dựng hai mô hình dự báo riêng biệt: mô hình Random Forest được huấn luyện trên dữ liệu dạng bảng gồm các đặc trưng trích xuất thủ công, còn mô hình LSTM được thiết kế để tiếp nhận chuỗi giá theo thời gian, tận dụng khả năng ghi nhớ để học các xu hướng biến động phức tạp. Việc lựa chọn hai mô hình này nhằm mục tiêu so sánh hai phương pháp dự báo theo hướng tách biệt: học máy truyền thống và học sâu tuần tự.

Thứ ba, thực hiện đánh giá hiệu quả của từng mô hình thông qua các chỉ số sai số phổ biến như MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), MSE (Mean Squared Error) và R^2 (hệ số xác định). Đồng thời, việc trực quan hóa kết quả bằng biểu đồ so sánh giữa giá thực tế và giá dự báo cũng được sử dụng để hỗ trợ đánh giá một cách trực quan và toàn diện.

Tiếp theo, dựa trên kết quả thực nghiệm, phân tích sự khác biệt trong hiệu suất giữa hai mô hình, từ đó xác định mô hình nào có khả năng dự báo tốt hơn đối với dữ liệu giá cổ phiếu của Viglacera. Việc xác định mô hình tối ưu không chỉ phục vụ mục tiêu học thuật mà còn định hướng cho việc xây dựng hệ thống hỗ trợ ra quyết định đầu tư, góp phần vào việc ứng dụng các mô hình dự báo hiện đại trong môi trường đầu tư thực tiễn.

Cuối cùng, từ quá trình nghiên cứu và thực nghiệm, đề xuất một số định hướng phát triển mô hình trong tương lai, bao gồm mở rộng phạm vi áp dụng sang các cổ phiếu khác, kết hợp thêm dữ liệu phi cấu trúc như tin tức hoặc cảm xúc thị trường, và tích hợp mô hình vào các hệ thống cảnh báo tự động hỗ trợ nhà đầu tư.

❖ **Đối tượng và phạm vi nghiên cứu**

Trong một nghiên cứu thực nghiệm, việc xác định rõ đối tượng và phạm vi nghiên cứu là điều cần thiết nhằm đảm bảo tính tập trung, độ tin cậy của kết quả và phù hợp với mục tiêu đề tài. Trên cơ sở đó, phần này trình bày cụ thể về đối tượng cũng như giới hạn phạm vi triển khai nghiên cứu.

Đối tượng nghiên cứu của đề tài là giá cổ phiếu của Tổng Công ty Viglacera – CTCP (mã chứng khoán VGC), được niêm yết và giao dịch trên Sở Giao dịch Chứng khoán TP. Hồ Chí Minh (HOSE). Đây là một cổ phiếu có tính thanh khoản tốt, dữ liệu giao dịch đầy đủ và ổn định trong nhiều năm, qua đó tạo điều kiện thuận lợi cho việc thu thập và xử lý chuỗi thời gian phục vụ cho việc huấn luyện các mô hình dự báo.

Song song với đó, đối tượng kỹ thuật trong nghiên cứu là hai mô hình học máy và học sâu gồm Random Forest Regression và LSTM. Việc lựa chọn hai mô hình này không chỉ nhằm phục vụ mục tiêu so sánh hiệu quả dự báo, mà còn giúp kiểm chứng khả năng thích ứng của các phương pháp hiện đại trong xử lý dữ liệu tài chính thực tế tại Việt Nam. Về phạm vi nghiên cứu, đề tài giới hạn ở các nội dung sau:

- Thứ nhất, nghiên cứu chỉ tập trung vào một mã cổ phiếu duy nhất là VGC, thay vì phân tích danh mục đa cổ phiếu. Việc giới hạn phạm vi này giúp đảm bảo độ sâu phân tích và giảm thiểu biến động ngoại sinh không kiểm soát được.
- Thứ hai, dữ liệu được thu thập trong khoảng thời gian từ ngày 01/03/2022 đến ngày 01/03/2025, nhằm đảm bảo tính liên tục, cập nhật và đủ dài để mô hình học được các xu hướng ngắn, trung và dài hạn trong giá cổ phiếu.
- Thứ ba, nghiên cứu chỉ sử dụng dữ liệu định lượng, bao gồm giá đóng cửa, khối lượng giao dịch và một số chỉ báo kỹ thuật cơ bản. Các yếu tố định tính như tin tức, chính sách vĩ mô hoặc phân tích cảm xúc thị trường chưa được đưa vào phạm vi của đề tài.

Như vậy, việc xác định rõ ràng đối tượng và phạm vi nghiên cứu không chỉ giúp định hướng rõ ràng quá trình thực hiện đề tài, mà còn tạo cơ sở để đánh giá kết quả một cách khách quan, nhất quán và có thể mở rộng trong các nghiên cứu tiếp theo.

❖ **Phương pháp nghiên cứu**

Để đạt được mục tiêu đã đề ra, bài khoá luận sử dụng kết hợp các phương pháp nghiên cứu định lượng và mô hình hóa dữ liệu nhằm đảm bảo tính khách quan, chính xác và có thể lặp lại trong thực nghiệm. Việc lựa chọn và triển khai các phương pháp này được xây dựng theo chuỗi logic từ thu thập dữ liệu, xử lý dữ liệu, xây dựng mô hình, đánh giá kết quả và rút ra kết luận. Cụ thể, các phương pháp nghiên cứu được áp dụng bao gồm:

Thứ nhất, phương pháp thu thập và xử lý dữ liệu: Dữ liệu lịch sử giá cổ phiếu VGC được thu thập từ các nguồn uy tín như VNDirect, Yahoo Finance hoặc Cafef. Trong đó, em sử dụng các biến cơ bản như giá đóng cửa, khối lượng giao dịch và các chỉ báo kỹ thuật. Sau khi thu thập, dữ liệu được làm sạch, chuẩn hóa và tổ chức lại theo định dạng phù hợp với từng mô hình: dữ liệu bảng cho Random Forest và dữ liệu chuỗi cho LSTM.

Thứ hai, phương pháp mô hình hóa: Hai mô hình dự báo được sử dụng là Random Forest Regression và LSTM. Random Forest Regression là mô hình học máy tổng hợp nhiều cây quyết định, có khả năng xử lý dữ liệu phi tuyến và tương tác giữa các đặc trưng. LSTM là một kiến trúc mạng nơ-ron hồi tiếp đặc biệt, được thiết kế để ghi nhớ các mối liên hệ trong chuỗi thời gian dài, rất phù hợp với bài toán dự báo chuỗi giá cổ phiếu. Mỗi mô hình được triển khai độc lập với các bộ tham số riêng, trên cùng một tập dữ liệu huấn luyện và kiểm thử.

Thứ ba, phương pháp đánh giá mô hình: Hiệu quả của mô hình được đánh giá dựa trên các chỉ số sai số phổ biến như:

MSE (Mean Squared Error): sai số bình phương trung bình

RMSE (Root Mean Squared Error): căn bậc hai của MSE

MAE (Mean Absolute Error): sai số tuyệt đối trung bình

R^2 (hệ số xác định): đo mức độ phù hợp giữa dữ liệu dự báo và dữ liệu thực tế

Ngoài ra, việc trực quan hóa kết quả dự báo thông qua biểu đồ cũng được sử dụng nhằm hỗ trợ việc đánh giá trực quan và phát hiện những điểm bất thường.

Thứ tư, phương pháp phân tích so sánh: Sau khi thu được kết quả từ hai mô hình, em tiến hành so sánh các chỉ số đánh giá, phân tích mức độ sai số, độ ổn định và khả năng phản ứng của mô hình với các biến động giá. Từ đó, em rút ra nhận định về mô hình phù hợp hơn với bài toán dự báo giá cổ phiếu VGC.

❖ **Câu hỏi nghiên cứu**

Để đạt được mục tiêu tổng quát là ứng dụng các phương pháp học máy và học sâu trong dự báo giá cổ phiếu của Tổng công ty Viglacera – CTCP (VGC), đề tài tập trung giải quyết các câu hỏi nghiên cứu sau:

- Mô hình học máy Random Forest có thể dự báo giá cổ phiếu VGC với độ chính xác như thế nào khi chỉ sử dụng dữ liệu giá lịch sử?
- Mô hình học sâu LSTM có hiệu quả hơn mô hình Random Forest trong việc dự báo giá cổ phiếu VGC không? Nếu có, mức độ cải thiện ra sao?
- Các chỉ số đánh giá mô hình như MAE, RMSE, MAPE và R^2 phản ánh sự khác biệt như thế nào giữa hai phương pháp dự báo?
- Dữ liệu giá cổ phiếu VGC trong giai đoạn 2022–2025 có thể được khai thác và trực quan hóa ra sao để phục vụ hiệu quả cho quá trình xây dựng mô hình?

Những câu hỏi trên không chỉ giúp định hướng nội dung nghiên cứu mà còn là cơ sở để đánh giá mức độ hoàn thiện và giá trị ứng dụng của các mô hình được triển khai trong đề tài.

❖ **Cấu trúc khoá luận**

Khóa luận được chia thành bốn chương, nhằm thể hiện rõ mạch nghiên cứu từ lý thuyết đến thực nghiệm và kết luận, cụ thể như sau:

Chương 1 – Cơ sở lý thuyết về các mô hình dự báo cổ phiếu**:** Trình bày tổng quan về cổ phiếu, các yếu tố ảnh hưởng đến giá cổ phiếu, các phương pháp dự báo truyền thống cũng như các mô hình hiện đại như Random Forest và LSTM. Đồng thời, chương này cũng làm rõ các chỉ tiêu đánh giá hiệu quả dự báo.

Chương 2 – Thực trạng Tổng Công ty Viglacera cứu: Giới thiệu khái quát về doanh nghiệp Viglacera, tình hình cổ phiếu VGC trên thị trường chứng khoán

Chương 3 – Phương pháp và kết quả nghiên cứu: Trình bày quy trình xây dựng mô hình dự báo bằng Random Forest và LSTM. Mỗi mô hình được triển khai riêng biệt, sau đó đánh giá và so sánh hiệu quả dự báo dựa trên các chỉ số sai số. Kết quả trực quan cũng được minh họa bằng biểu đồ để hỗ trợ phân tích.

Chương 4 – Kết luận và đề xuất giải pháp phát triển: Tổng kết lại các kết quả chính đạt được trong nghiên cứu, nêu rõ mô hình dự báo hiệu quả hơn, đánh giá mức độ phù hợp, phân tích sai số, và trình bày một số đề xuất phát triển mô hình trong tương lai. Đồng thời, chương này cũng đưa ra kiến nghị về khả năng ứng dụng trong thực tế và định hướng mở rộng nghiên cứu tiếp theo.

CHƯƠNG 1. CƠ SỞ LÝ THUYẾT CÁC MÔ HÌNH DỰ BÁO CỔ PHIẾU

1.1. Tổng quan về cổ phiếu

1.1.1. Các vấn đề chung của cổ phiếu

❖ Khái niệm

Cổ phiếu là một loại chứng khoán xác nhận quyền sở hữu của người nắm giữ đối với một phần vốn điều lệ của công ty cổ phần. Khi sở hữu cổ phiếu, nhà đầu tư trở thành cổ đông và đồng thời là chủ sở hữu hợp pháp của doanh nghiệp với tư cách tương ứng với tỷ lệ cổ phần nắm giữ. Theo quy định pháp luật hiện hành, cổ đông có các quyền cơ bản như quyền biểu quyết tại Đại hội đồng cổ đông, quyền nhận cổ tức, quyền chuyển nhượng cổ phần và quyền được phân chia tài sản khi doanh nghiệp giải thể hoặc phá sản.

Trên thị trường chứng khoán, cổ phiếu đóng vai trò là công cụ quan trọng để doanh nghiệp huy động vốn từ công chúng. Thay vì vay vốn từ các tổ chức tín dụng, doanh nghiệp có thể phát hành cổ phiếu để thu hút dòng vốn đầu tư trực tiếp, đồng thời tạo ra tính thanh khoản cho nhà đầu tư thông qua hoạt động giao dịch trên thị trường thứ cấp.

Giá cổ phiếu được hiểu là mức giá mà tại đó cổ phiếu được mua bán trên thị trường chứng khoán tại một thời điểm cụ thể. Đây là một đại lượng biến thiên theo thời gian và thường xuyên chịu ảnh hưởng bởi nhiều yếu tố khác nhau, cả về mặt vi mô lẫn vĩ mô. Giá cổ phiếu phản ánh kỳ vọng của thị trường về giá trị doanh nghiệp trong tương lai, do đó có thể dao động lên hoặc xuống một cách linh hoạt dựa trên thông tin tài chính, tình hình kinh tế, tâm lý thị trường, hoặc các sự kiện bất thường.

Trong nghiên cứu này, khái niệm giá cổ phiếu được hiểu và sử dụng theo nghĩa thị trường, tức là giá đóng cửa hàng ngày của cổ phiếu VGC được ghi nhận trên sàn giao dịch chứng khoán. Đây là loại dữ liệu phổ biến và mang tính định lượng rõ ràng, thường được sử dụng trong các mô hình dự báo chuỗi thời gian. Việc lựa chọn giá đóng cửa làm đại diện cho giá cổ phiếu là hợp lý, vì đây là mức giá thể hiện sự đồng thuận giữa cung và cầu sau một phiên giao dịch, đồng thời có tính ổn định và ít nhiễu hơn so với các mức giá trong phiên như giá mở cửa, giá cao nhất hay giá thấp nhất.

Tóm lại, cổ phiếu vừa là chứng nhận quyền sở hữu vốn, vừa là tài sản tài chính có thể giao dịch, còn giá cổ phiếu chính là thước đo phản ánh giá trị kỳ vọng của doanh

nghiệp trong mắt thị trường. Việc phân tích và dự báo giá cổ phiếu, do đó, không chỉ mang tính học thuật mà còn có ý nghĩa quan trọng trong hoạt động đầu tư và ra quyết định tài chính.

❖ Phân loại cổ phiếu

Trong hệ thống thị trường chứng khoán, cổ phiếu được phân loại theo nhiều tiêu chí khác nhau nhằm phục vụ cho mục đích quản lý, giao dịch và phân tích đầu tư. Việc phân loại giúp nhận diện rõ quyền lợi, nghĩa vụ, cũng như mức độ rủi ro gắn liền với từng loại cổ phiếu, từ đó hỗ trợ nhà đầu tư đưa ra lựa chọn phù hợp với chiến lược tài chính của mình.

Theo quyền lợi và nghĩa vụ của cổ đông, cổ phiếu thường được chia thành hai nhóm chính là cổ phiếu phổ thông và cổ phiếu ưu đãi. Cổ phiếu phổ thông là loại hình cổ phiếu phổ biến nhất trên thị trường, cho phép người sở hữu tham gia biểu quyết trong Đại hội đồng cổ đông, hưởng cổ tức theo kết quả kinh doanh và có quyền nhận phần tài sản còn lại nếu doanh nghiệp giải thể. Ngược lại, cổ phiếu ưu đãi mang lại một số quyền lợi tài chính cao hơn, như cổ tức cố định hoặc được ưu tiên thanh toán trước trong trường hợp công ty thanh lý tài sản. Tuy nhiên, cổ đông sở hữu cổ phiếu ưu đãi thường bị giới hạn hoặc không có quyền biểu quyết trong các hoạt động quản trị doanh nghiệp.

Bên cạnh đó, nếu xét theo hình thức sở hữu, cổ phiếu có thể được phân loại thành cổ phiếu ghi danh và cổ phiếu vô danh. Cổ phiếu ghi danh là loại có tên của người sở hữu trên chứng chỉ hoặc hệ thống đăng ký điện tử của công ty, và mọi giao dịch chuyển nhượng phải được thông báo cho công ty phát hành. Trong khi đó, cổ phiếu vô danh không ghi tên chủ sở hữu và có thể chuyển nhượng một cách tự do, nhưng hiện nay hình thức này gần như không còn phổ biến tại các thị trường chứng khoán hiện đại do không đảm bảo tính minh bạch và tuân thủ pháp lý.

Ngoài ra, trong thực tế đầu tư, cổ phiếu còn có thể được phân loại theo một số tiêu chí khác như ngành nghề hoạt động của doanh nghiệp, quy mô vốn hóa thị trường (cổ phiếu vốn hóa lớn, vừa và nhỏ), mức độ tăng trưởng (cổ phiếu tăng trưởng, cổ phiếu giá trị), hay đặc điểm thanh khoản. Những cách phân loại này có thể thay đổi tùy theo cách tiếp cận và mục tiêu phân tích cụ thể của từng nhà đầu tư, tổ chức tài chính hoặc đơn vị nghiên cứu.

Việc hiểu rõ các loại cổ phiếu không chỉ giúp nhà đầu tư xác định được quyền lợi khi nắm giữ, mà còn hỗ trợ quá trình ra quyết định trong bối cảnh thị trường có nhiều

biến động và yêu cầu phân tích ngày càng chuyên sâu.

❖ Đặc điểm của cổ phiếu

Cổ phiếu, với tư cách là một loại chứng khoán phổ biến nhất trên thị trường tài chính, sở hữu nhiều đặc điểm riêng biệt phản ánh bản chất kép vừa là công cụ đầu tư, vừa là bằng chứng pháp lý về quyền sở hữu doanh nghiệp. Các đặc điểm nổi bật của cổ phiếu có thể được trình bày như sau:

Thứ nhất, cổ phiếu thể hiện quyền sở hữu. Người nắm giữ cổ phiếu là cổ đông và có quyền sở hữu một phần tương ứng trong doanh nghiệp. Tỷ lệ quyền sở hữu này thường được xác định dựa trên tỷ lệ giữa số lượng cổ phiếu sở hữu so với tổng số cổ phiếu đang lưu hành. Quyền sở hữu này mang tính chất lâu dài và chỉ chấm dứt khi cổ phiếu được chuyển nhượng cho người khác.

Thứ hai, cổ phiếu có khả năng sinh lời. Cổ đông có thể thu được lợi nhuận từ cổ phiếu thông qua hai hình thức chính: cổ tức (khoản lợi nhuận được chia từ kết quả kinh doanh của doanh nghiệp) và chênh lệch giá (khoản chênh lệch giữa giá mua và giá bán cổ phiếu). Khả năng sinh lời này phụ thuộc vào tình hình tài chính của công ty, điều kiện thị trường và thời điểm đầu tư.

Thứ ba, cổ phiếu mang tính rủi ro cao. Do giá cổ phiếu biến động thường xuyên, nhà đầu tư có thể chịu thua lỗ nếu giá thị trường giảm so với giá mua. Rủi ro này có thể đến từ cả yếu tố nội tại doanh nghiệp như kết quả kinh doanh suy giảm, lạm phát chi phí, thay đổi quản trị; lẫn các yếu tố bên ngoài như khủng hoảng kinh tế, thay đổi chính sách vĩ mô, hoặc biến động thị trường toàn cầu.

Thứ tư, cổ phiếu không có thời hạn. Khác với trái phiếu có kỳ hạn đáo hạn rõ ràng, cổ phiếu tồn tại cho đến khi doanh nghiệp bị giải thể, phá sản hoặc thực hiện mua lại cổ phần. Tính chất không kỳ hạn này làm cho cổ phiếu trở thành một hình thức đầu tư dài hạn phù hợp với nhà đầu tư có chiến lược tích lũy vốn hoặc gia tăng quyền kiểm soát doanh nghiệp.

Thứ năm, cổ phiếu có tính thanh khoản. Trên thị trường chứng khoán tập trung, nhà đầu tư có thể dễ dàng mua bán cổ phiếu trong phiên giao dịch. Tuy nhiên, mức độ thanh khoản của từng mã cổ phiếu là khác nhau, phụ thuộc vào khối lượng giao dịch, mức độ quan tâm của thị trường, và vị thế doanh nghiệp phát hành. Các cổ phiếu có vốn hóa lớn và thuộc nhóm ngành ổn định thường có thanh khoản cao hơn các cổ phiếu vốn hóa nhỏ.

Thứ sáu, cổ phiếu có tính dễ phân chia. Cổ phiếu có thể được chia nhỏ về mệnh giá hoặc số lượng, tạo điều kiện cho nhiều nhà đầu tư cùng tham gia vào thị trường với các quy mô vốn khác nhau. Tính chất này giúp tăng tính đại chúng trong sở hữu vốn, đồng thời góp phần thúc đẩy tính minh bạch và hiệu quả trong hoạt động doanh nghiệp.

Thứ bảy, cổ phiếu mang tính bất định trong thu nhập. Mức cổ tức được chia hằng năm không cố định, phụ thuộc vào kết quả kinh doanh và chính sách tài chính của doanh nghiệp. Khác với trái phiếu có lãi suất được ấn định trước, cổ phiếu không cam kết mức sinh lời cụ thể, do đó người nắm giữ phải chấp nhận rủi ro đi kèm với cơ hội sinh lợi cao hơn.

Thứ tám, cổ phiếu chịu ảnh hưởng lớn từ tâm lý thị trường. Những biến động về thông tin, kỳ vọng, tin đồn hoặc hiệu ứng đám đông có thể khiến giá cổ phiếu dao động mạnh trong thời gian ngắn mà không phản ánh đúng giá trị thực của doanh nghiệp. Điều này đòi hỏi nhà đầu tư phải có năng lực phân tích, kiểm soát cảm xúc và kỹ năng ra quyết định hợp lý.

Bên cạnh những đặc điểm nêu trên, cổ phiếu còn thể hiện tính pháp lý cao. Việc phát hành, giao dịch và sở hữu cổ phiếu đều chịu sự điều chỉnh của hệ thống pháp luật và các quy định thị trường chứng khoán. Tính pháp lý này góp phần bảo vệ quyền lợi của nhà đầu tư và đảm bảo trật tự trong hoạt động thị trường.

Với các đặc điểm đa chiều và phức hợp, cổ phiếu là một công cụ đầu tư vừa hấp dẫn vừa thách thức. Việc hiểu rõ bản chất và tính chất của cổ phiếu là tiền đề quan trọng để nghiên cứu các mô hình dự báo giá cổ phiếu một cách có hệ thống và phù hợp với đặc thù thị trường.

1.1.2. Các yếu tố ảnh hưởng đến sự thay đổi giá cổ phiếu

Giá cổ phiếu là kết quả tổng hợp của nhiều yếu tố tác động, trong đó có cả yếu tố thuộc về bản thân doanh nghiệp, yếu tố thị trường và các điều kiện kinh tế – xã hội bên ngoài. Việc nhận diện đầy đủ và phân loại các yếu tố này là cơ sở quan trọng cho việc xây dựng các mô hình dự báo giá cổ phiếu có tính hiệu quả và ứng dụng cao.

Trước hết, yếu tố nội tại doanh nghiệp đóng vai trò then chốt trong việc hình thành giá trị thực của cổ phiếu. Những yếu tố này bao gồm kết quả hoạt động sản xuất – kinh doanh, tốc độ tăng trưởng doanh thu và lợi nhuận, khả năng quản trị tài chính, cấu trúc vốn, mức độ minh bạch thông tin, cũng như chất lượng quản trị của ban điều hành. Những doanh nghiệp có nền tảng tài chính vững chắc và hiệu quả kinh doanh ổn định

thường có cổ phiếu được thị trường đánh giá cao và giá cổ phiếu có xu hướng tăng trưởng bền vững.

Tiếp theo, yếu tố kinh tế vĩ mô ảnh hưởng đến toàn bộ thị trường và hành vi nhà đầu tư. Các biến số như tốc độ tăng trưởng GDP, tỷ lệ lạm phát, lãi suất điều hành, tỷ giá hối đoái, chính sách tài khóa và tiền tệ của Chính phủ đều có tác động gián tiếp nhưng sâu sắc đến giá cổ phiếu. Ví dụ, lãi suất tăng thường làm chi phí vốn cao hơn, ảnh hưởng đến lợi nhuận của doanh nghiệp và sức mua cổ phiếu của nhà đầu tư.

Bên cạnh đó, yếu tố thị trường bao gồm tình trạng cung – cầu cổ phiếu, mức độ biến động chung của chỉ số thị trường, xu hướng đầu tư của khối ngoại, các hiệu ứng lan truyền trong tâm lý đám đông và thông tin từ các tổ chức phân tích. Đây là nhóm yếu tố có tác động mạnh đến giá cổ phiếu trong ngắn hạn, nhưng lại khó định lượng và kiểm soát.

Cuối cùng, yếu tố kỹ thuật có vai trò hỗ trợ nhà đầu tư trong việc phân tích xu hướng và điểm mua – bán hợp lý thông qua các chỉ báo được xây dựng từ dữ liệu lịch sử như đường trung bình động, chỉ số sức mạnh tương đối, dao động ngẫu nhiên, mô hình nến Nhật… Các yếu tố này tuy không phản ánh giá trị nội tại của doanh nghiệp nhưng lại có thể ảnh hưởng đến kỳ vọng và hành vi giao dịch của nhà đầu tư.

Tổng hợp các nhóm yếu tố nêu trên cho thấy, giá cổ phiếu là một đại lượng phản ánh sự tương tác phức tạp giữa dữ liệu định lượng, tâm lý thị trường và các yếu tố vĩ mô và vi mô. Do đó, mô hình dự báo giá cổ phiếu cần phải linh hoạt, có khả năng học và phản ứng với nhiều dạng thông tin để cho kết quả chính xác và ổn định.

1.1.3. Lí thuyết dự báo giá cổ phiếu

Việc dự báo giá cổ phiếu có vai trò rất quan trọng trong hoạt động đầu tư và phân tích thị trường tài chính. Trên thực tế, dự báo giá cổ phiếu không chỉ nhằm xác định mức giá tương lai của một mã chứng khoán, mà còn phục vụ mục tiêu ra quyết định hiệu quả, quản trị rủi ro và tối ưu hóa danh mục đầu tư.

Đối với nhà đầu tư cá nhân, dự báo giá cổ phiếu giúp xác định thời điểm mua vào hoặc bán ra hợp lý, từ đó tối đa hóa lợi nhuận hoặc hạn chế thiệt hại khi thị trường biến động bất lợi. Trong bối cảnh thị trường chứa đựng nhiều thông tin không hoàn hảo và chịu ảnh hưởng bởi tâm lý đám đông, việc có được một công cụ hỗ trợ ra quyết định dựa trên dữ liệu là rất cần thiết.

Đối với các tổ chức tài chính như công ty chứng khoán, quỹ đầu tư, hoặc ngân

hàng, dự báo giá cổ phiếu là một phần không thể thiếu trong quá trình xây dựng chiến lược đầu tư, phát triển các sản phẩm tài chính phái sinh hoặc đưa ra các khuyến nghị đầu tư cho khách hàng. Việc áp dụng mô hình dự báo cũng giúp nâng cao chất lượng dịch vụ phân tích và hỗ trợ quản lý danh mục đầu tư một cách hiệu quả.

Về mặt học thuật và nghiên cứu, dự báo giá cổ phiếu là bài toán đặc trưng của phân tích chuỗi thời gian phi tuyến và biến động cao. Đây cũng là một trong những lĩnh vực ứng dụng phổ biến và hiệu quả của các mô hình học máy và học sâu, đặc biệt là khi dữ liệu ngày càng phong phú và phức tạp hơn. Do đó, việc nghiên cứu và cải tiến các mô hình dự báo không chỉ mang lại ý nghĩa thực tiễn, mà còn góp phần vào sự phát triển của lĩnh vực phân tích tài chính định lượng.

Nhìn chung, vai trò của việc dự báo giá cổ phiếu ngày càng trở nên quan trọng trong bối cảnh thị trường tài chính hiện đại. Với sự hỗ trợ của công nghệ, đặc biệt là các mô hình trí tuệ nhân tạo, việc xây dựng các công cụ dự báo chính xác và linh hoạt sẽ góp phần nâng cao hiệu quả đầu tư, hỗ trợ hoạch định chiến lược và quản trị rủi ro một cách chủ động.

1.1.4. Các chỉ báo phân tích kĩ thuật trong dự báo giá cổ phiếu

❖ SMA

Trong phân tích dữ liệu tài chính, đặc biệt là chuỗi thời gian giá cổ phiếu, việc trực quan hóa nhằm khám phá xu hướng và đặc điểm biến động là bước quan trọng trước khi triển khai bất kỳ mô hình dự báo nào. Một trong những công cụ trực quan hóa đơn giản nhưng hữu hiệu được sử dụng phổ biến là đường trung bình động đơn giản (Simple Moving Average – SMA).

SMA là chỉ báo kỹ thuật được tính bằng cách lấy trung bình cộng giá đóng cửa trong một số phiên gần nhất. Chỉ báo này có vai trò làm mượt chuỗi thời gian, từ đó giúp nhà phân tích dễ dàng quan sát xu hướng tổng thể của giá. Không giống như các chỉ số phức tạp khác, SMA không đòi hỏi giả định thống kê hay tham số học, và do đó rất phù hợp để sử dụng trong giai đoạn khám phá dữ liệu ban đầu.

Trong đề tài này, SMA được sử dụng với mục tiêu trực quan hóa, không phải là biến đầu vào cho mô hình. Hai mức SMA được lựa chọn là SMA_5 và SMA_10, đại diện cho trung bình động trong 5 và 10 phiên gần nhất, thường được giới đầu tư cá nhân và tổ chức sử dụng để đánh giá xu hướng ngắn hạn.

Việc kết hợp SMA với biểu đồ giá đóng cửa thực tế mang lại nhiều lợi ích trong

việc hiểu rõ hơn về hành vi:

- Xác định xu hướng chính: SMA giúp làm mượt dữ liệu, từ đó người phân tích dễ dàng phân biệt được các giai đoạn thị trường đang có xu hướng tăng, giảm hay đi ngang.
- Phát hiện điểm đảo chiều tiềm năng: Khi đường giá cắt lên hoặc cắt xuống đường SMA, đó có thể là tín hiệu cho sự chuyển đổi xu hướng. Điều này giúp nhận diện những thời điểm bất ổn hoặc cơ hội đầu tư.
- Đánh giá độ biến động giá: Khi đường giá dao động mạnh quanh SMA, điều đó thể hiện thị trường đang biến động mạnh và không ổn định. Ngược lại, nếu đường giá bám sát hoặc nằm trên SMA với độ lệch nhỏ, đó có thể là dấu hiệu của một xu hướng rõ ràng và ổn định.

Trong ngữ cảnh của cổ phiếu VGC, việc trực quan hóa SMA hỗ trợ người nghiên cứu trong việc nhận biết các giai đoạn tăng trưởng, điều chỉnh, và tích lũy của cổ phiếu qua từng chu kỳ thời gian. Điều này có giá trị tham khảo thực tiễn quan trọng đối với cả phân tích đầu tư và xây dựng mô hình dự báo.

Mặc dù SMA không được sử dụng làm đầu vào cho các mô hình học máy trong đề tài này, nhưng vai trò của nó trong việc khám phá dữ liệu ban đầu là không thể thay thế. Đây là minh chứng rõ ràng cho việc kết hợp giữa trực quan hóa truyền thống và mô hình hóa hiện đại nhằm đạt được cái nhìn toàn diện và khoa học trong phân tích tài chính định lượng.

❖ EMA

Bên cạnh SMA, một công cụ trực quan hóa khác cũng được sử dụng phổ biến trong phân tích kỹ thuật là đường trung bình động hàm mũ (Exponential Moving Average – EMA). EMA là một biến thể của đường trung bình động, nhưng được cải tiến bằng cách gán trọng số lớn hơn cho các quan sát gần thời điểm hiện tại. Nhờ vậy, EMA có khả năng phản ánh nhanh nhạy hơn các thay đổi đột ngột của giá, đặc biệt trong môi trường thị trường có biến động mạnh.

Trong đề tài, EMA được sử dụng với mục đích trực quan hóa xu hướng và biến động ngắn hạn của giá cổ phiếu VGC, hỗ trợ nhà phân tích nhận diện sớm các tín hiệu tăng hoặc giảm. Không giống SMA – vốn làm mượt đều chuỗi dữ liệu – EMA nhạy hơn với các điểm ngoặt trong xu hướng giá và thường được ưu tiên sử dụng trong chiến lược phân tích động.

Ý nghĩa và ứng dụng trong đề tài: Phản ứng nhanh với giá hiện tại: So với SMA, EMA phản ứng nhạy hơn với giá mới. Điều này giúp phát hiện nhanh các thay đổi xu hướng – đặc biệt quan trọng trong thị trường có độ biến động cao như thị trường chứng khoán Việt Nam. Làm nổi bật động lượng (momentum): Khi EMA ngắn hạn (ví dụ: EMA_5) cắt lên EMA dài hạn (EMA_10), điều này thường được xem là tín hiệu tăng giá tiềm năng, ngược lại là tín hiệu giảm giá. Giảm độ trễ trong quan sát: EMA hạn chế độ trễ so với SMA, giúp nhà phân tích không bỏ lỡ các điểm đảo chiều quan trọng khi thị trường bắt đầu chuyển trạng thái. Hiển thị rõ ràng trên biểu đồ: Trên đồ thị trực quan hóa, đường EMA giúp phân biệt rõ ràng các vùng tăng trưởng ổn định và các pha điều chỉnh ngắn hạn, nhờ vào độ cong đặc trưng so với đường giá gốc.

Trong đề tài này, EMA được sử dụng như một công cụ trực quan bổ sung, nhằm làm nổi bật các vùng chuyển động của giá cổ phiếu VGC theo thời gian. Mặc dù không trực tiếp tham gia vào mô hình dự báo, nhưng EMA đóng vai trò quan trọng trong việc khám phá hành vi chuỗi thời gian và xây dựng cơ sở nhận thức định tính trước khi tiến hành mô hình hóa định lượng.

❖ RSI

Bên cạnh các chỉ báo làm mượt xu hướng giá như SMA và EMA, một trong những công cụ phân tích kỹ thuật hữu ích trong việc đánh giá động lượng thị trường là chỉ số sức mạnh tương đối (Relative Strength Index – RSI). RSI được phát triển bởi nhà phân tích kỹ thuật J. Welles Wilder với mục tiêu đo lường tốc độ và mức độ thay đổi của giá chứng khoán trong một khoảng thời gian xác định. Không giống như các chỉ báo đơn thuần theo dõi xu hướng, RSI cho phép nhận diện các giai đoạn mà thị trường có thể đã quá mua hoặc quá bán, từ đó đưa ra cảnh báo sớm về khả năng đảo chiều.

Trong bối cảnh phân tích cổ phiếu VGC, RSI được sử dụng như một công cụ trực quan hóa nhằm bổ sung thêm góc nhìn định tính cho quá trình khai phá dữ liệu. Cụ thể, chỉ báo này được tính toán dựa trên biến động giá tăng và giảm liên tục trong một chuỗi thời gian nhất định (thường là 14 phiên). Giá trị RSI dao động từ 0 đến 100, trong đó các ngưỡng 30 và 70 được sử dụng phổ biến để nhận biết các trạng thái cực đoan của thị trường. Khi RSI vượt trên 70, cổ phiếu có thể đang trong trạng thái bị mua quá mức, ngược lại, khi RSI rơi xuống dưới 30, điều đó có thể phản ánh trạng thái bán tháo hoặc bị định giá thấp bất thường.

Việc trực quan hóa RSI trên biểu đồ cùng với đường giá đóng cửa thực tế không

chỉ giúp làm nổi bật các giai đoạn biến động mạnh, mà còn cung cấp cơ sở để đánh giá mức độ ổn định hoặc nhiễu loạn trong hành vi giá. Đặc biệt, trong các giai đoạn thị trường thiếu xu hướng rõ ràng, RSI thường dao động quanh vùng trung lập (gần 50), phản ánh sự giằng co giữa bên mua và bên bán. Ngược lại, tại các vùng RSI đạt cực trị, mô hình giá thường xuất hiện các tín hiệu điều chỉnh hoặc phục hồi.

❖ MACD

Trong lĩnh vực phân tích kỹ thuật tài chính, chỉ báo MACD (Moving Average Convergence Divergence – hội tụ phân kỳ trung bình động) là một công cụ phổ biến và có độ tin cậy cao trong việc đánh giá xu hướng và động lượng của giá tài sản. Đây là chỉ báo được phát triển bởi Gerald Appel vào cuối những năm 1970 và đến nay vẫn được sử dụng rộng rãi trong giao dịch cổ phiếu, tiền tệ và các tài sản tài chính khác.

Một điểm nổi bật của MACD là khả năng kết hợp thông tin từ hai loại trung bình động với độ nhạy khác nhau, từ đó phản ánh sự thay đổi trong động lực thị trường theo thời gian. Khác với các chỉ báo chỉ dựa vào giá đóng cửa, MACD dựa trên mối quan hệ giữa hai đường trung bình động hàm mũ (EMA), giúp nhà phân tích dễ dàng phát hiện các tín hiệu đảo chiều xu hướng.

Về mặt cấu trúc, MACD bao gồm ba thành phần chính. Thứ nhất là đường MACD, được tính toán bằng hiệu số giữa EMA ngắn hạn (thường là 12 phiên) và EMA dài hạn (thường là 26 phiên). Đường này phản ánh tốc độ thay đổi của giá trong ngắn hạn so với trung hạn, từ đó chỉ ra xu hướng giá chủ đạo. Thứ hai là đường tín hiệu (Signal Line), thường là EMA 9 phiên của đường MACD. Đường tín hiệu giúp làm mượt chuỗi dữ liệu và là cơ sở để phát hiện các điểm giao cắt – yếu tố cốt lõi tạo nên tín hiệu mua hoặc bán. Cuối cùng, biểu đồ MACD còn có thể hiển thị histogram, là phần trực quan hóa độ chênh lệch giữa hai đường nói trên.

Về nguyên tắc hoạt động, MACD tạo ra tín hiệu giao dịch thông qua các giao điểm giữa đường MACD và đường tín hiệu. Khi đường MACD cắt lên trên đường tín hiệu, đó là dấu hiệu cho thấy thị trường đang tăng động lượng, từ đó hình thành tín hiệu mua. Ngược lại, khi MACD cắt xuống dưới đường tín hiệu, nhà đầu tư có thể cân nhắc đây là dấu hiệu cho thấy đà giảm đang mạnh lên – một tín hiệu bán. Bên cạnh đó, khi đường MACD nằm trên ngưỡng 0, thị trường được đánh giá là đang trong xu hướng tăng. Trái lại, khi MACD nằm dưới mức 0, điều này phản ánh tâm lý thị trường đang nghiêng về chiều giảm.

Tuy nhiên, cũng cần lưu ý rằng MACD là một chỉ báo thuộc nhóm "trễ" (lagging indicator), do được tính toán từ dữ liệu giá trong quá khứ. Vì lý do này, MACD đôi khi phản ứng chậm trước các cú sốc thị trường hoặc biến động đột ngột. Do đó, trong thực tế ứng dụng, MACD thường được kết hợp với các chỉ báo nhanh hơn như RSI (Relative Strength Index) hoặc các mô hình phân tích định lượng để gia tăng độ chính xác.

1.2. Tổng quan về học háy (Machine Learning)

1.2.1. Lí do chọn học máy

Học máy là một nhánh thuộc lĩnh vực trí tuệ nhân tạo, chuyên nghiên cứu và phát triển các thuật toán cho phép máy tính có thể học từ dữ liệu để cải thiện hiệu suất thực hiện một tác vụ cụ thể mà không cần được lập trình tường minh. Thay vì xác định trước các quy tắc xử lý bằng tay, học máy hướng đến việc xây dựng các mô hình có khả năng tự động rút ra quy luật từ dữ liệu đầu vào thông qua quá trình huấn luyện, từ đó đưa ra các dự đoán hoặc quyết định cho dữ liệu mới chưa từng xuất hiện.

Theo định nghĩa được Tom Mitchell (1997) đề xuất, một chương trình máy tính được coi là học nếu hiệu suất của nó trong việc thực hiện một tập hợp các nhiệm vụ cụ thể, được đo bằng một tiêu chí định lượng nhất định, được cải thiện theo kinh nghiệm. Định nghĩa này nhấn mạnh ba yếu tố quan trọng trong học máy: dữ liệu đầu vào, tác vụ cụ thể, và thước đo đánh giá hiệu suất.

Về bản chất, quá trình học máy là quá trình tối ưu hóa một hàm mục tiêu, thường là hàm mất mát (loss function), phản ánh mức độ sai lệch giữa kết quả dự đoán và giá trị thực tế. Thông qua quá trình huấn luyện, thuật toán điều chỉnh dần các tham số của mô hình để tối thiểu hóa sai số, từ đó nâng cao độ chính xác của dự đoán. Sau khi huấn luyện, mô hình có thể được sử dụng để dự đoán đầu ra trên dữ liệu chưa từng xuất hiện trước đó, điều này được gọi là khả năng tổng quát hóa (generalization).

Dựa trên đặc điểm của dữ liệu đầu vào và phương thức huấn luyện, học máy được phân chia thành ba loại chính: học có giám sát, học không giám sát và học bán giám sát. Trong đó, học có giám sát là phương pháp phổ biến nhất, trong đó mỗi điểm dữ liệu đầu vào đều đi kèm với một nhãn đầu ra. Mục tiêu của mô hình là học được mối quan hệ giữa đầu vào và đầu ra để áp dụng cho các trường hợp mới. Ngược lại, học không giám sát sử dụng dữ liệu không có nhãn, thường được dùng trong các bài toán phân cụm hoặc giảm chiều dữ liệu. Học bán giám sát là sự kết hợp giữa hai phương pháp trên, tận dụng

dữ liệu có nhãn để hỗ trợ học từ phần dữ liệu chưa có nhãn.

Ngoài các phương pháp trên, một hướng tiếp cận khác trong học máy là học tăng cường, trong đó mô hình học thông qua tương tác với môi trường bằng cách thực hiện hành động và nhận phần thưởng hoặc hình phạt tương ứng. Phương pháp này đặc biệt phù hợp với các bài toán ra quyết định trong môi trường thay đổi liên tục.

Trong bối cảnh dữ liệu ngày càng trở nên lớn và phức tạp, học máy được xem là công cụ có tiềm năng lớn trong việc phát hiện mẫu ẩn, mô hình hóa các mối quan hệ phi tuyến và hỗ trợ dự báo trong nhiều lĩnh vực khác nhau. Trong tài chính và cụ thể là dự báo giá cổ phiếu, học máy cung cấp một hướng tiếp cận mới, linh hoạt và hiệu quả hơn so với các mô hình thống kê truyền thống, đặc biệt trong việc xử lý dữ liệu chuỗi thời gian có tính biến động và độ nhiễu cao.

Tuy nhiên, việc áp dụng học máy vào các bài toán thực tế cũng đặt ra nhiều thách thức. Hiệu quả của mô hình phụ thuộc chặt chẽ vào chất lượng dữ liệu đầu vào, khả năng lựa chọn đặc trưng phù hợp, và quá trình đánh giá mô hình sau khi huấn luyện. Ngoài ra, hiện tượng quá khớp (overfitting) và thiếu khả năng giải thích mô hình (interpretability) cũng là những vấn đề cần được xem xét nghiêm túc.

1.2.2. Học có giám sát

Học có giám sát là một trong những phương pháp cơ bản và phổ biến nhất trong học máy. Phương pháp này hoạt động dựa trên cơ sở sử dụng tập dữ liệu huấn luyện đã được gán nhãn đầy đủ, nghĩa là mỗi quan sát đầu vào đều đi kèm với một giá trị đầu ra (nhãn) tương ứng. Mục tiêu của mô hình học có giám sát là tìm ra một hàm ánh xạ tối ưu từ không gian đầu vào đến không gian đầu ra, sao cho có thể dự đoán chính xác nhãn của các quan sát mới chưa từng thấy trước đó.

Quá trình huấn luyện trong học có giám sát bao gồm việc đưa vào mô hình một tập hợp các cặp dữ liệu dạng (x, y), trong đó x là đầu vào (vector đặc trưng), còn y là nhãn tương ứng (biến mục tiêu). Mô hình sẽ học thông qua việc tối ưu hóa một hàm mất mát, nhằm giảm thiểu sai số giữa giá trị dự đoán và giá trị thực tế. Khi huấn luyện hoàn tất, mô hình có thể được sử dụng để suy đoán đầu ra đối với các dữ liệu đầu vào mới chưa gán nhãn.

Về hình thức đầu ra, học có giám sát có thể được chia thành hai loại bài toán chính: bài toán phân loại và bài toán hồi quy. Bài toán phân loại hướng đến việc gán nhãn cho đầu vào theo các lớp rời rạc, trong khi bài toán hồi quy dự đoán đầu ra là các giá trị liên

tục. Trong khuôn khổ đề tài này, dự báo giá cổ phiếu là một bài toán hồi quy, do giá cổ phiếu là một đại lượng liên tục biến thiên theo thời gian.

Một số thuật toán phổ biến trong học có giám sát bao gồm hồi quy tuyến tính, hồi quy logistic, cây quyết định, máy vector hỗ trợ (SVM), mạng nơ-ron nhân tạo, và các mô hình tổ hợp như Random Forest hoặc Gradient Boosting. Mỗi thuật toán đều có những ưu – nhược điểm riêng, tùy thuộc vào cấu trúc dữ liệu, độ nhiễu, mức độ phi tuyến và yêu cầu về khả năng giải thích.

Việc áp dụng học có giám sát trong dự báo giá cổ phiếu mang lại nhiều lợi ích do dữ liệu thị trường tài chính thường bao gồm các chuỗi dữ liệu lịch sử được gán nhãn rõ ràng (ví dụ: giá đóng cửa ngày hôm sau là đầu ra, còn giá và khối lượng các ngày trước là đầu vào). Nhờ vậy, có thể xây dựng các mô hình dự báo theo thời gian một cách rõ ràng và hệ thống. Tuy nhiên, trong thực tiễn, việc chọn lựa tập biến đầu vào phù hợp, đảm bảo tính đại diện của dữ liệu huấn luyện và tránh quá khớp là những vấn đề cần được đặc biệt chú trọng để đảm bảo khả năng tổng quát của mô hình.

1.2.3. Tổng quan về mô hình Random Forest Regressor

Random Forest Regression (RFR) là một biến thể của thuật toán Random Forest được thiết kế nhằm giải quyết các bài toán hồi quy thay vì phân loại. Mô hình này được xây dựng trên nền tảng của phương pháp cây hồi quy (regression tree) bằng cách kết hợp nhiều cây đơn lẻ (base learners) thành một tổ hợp mô hình (ensemble) để đưa ra dự đoán trung bình. Mỗi cây trong rừng được huấn luyện trên một tập con bootstrap ngẫu nhiên của dữ liệu gốc, đồng thời tại mỗi nút phân tách, chỉ một tập con ngẫu nhiên các biến đầu vào được chọn để xem xét chia nhánh (Breiman, 2001).

Khác với hồi quy tuyến tính bội truyền thống (Multiple Linear Regression – MLR), vốn đòi hỏi các giả định nghiêm ngặt như tuyến tính giữa biến độc lập và phụ thuộc, phân phối chuẩn của sai số, tính đồng nhất phương sai (homoskedasticity), và không tự tương quan, RFR không yêu cầu bất kỳ giả định phân phối cụ thể nào đối với dữ liệu. Điều này khiến mô hình trở thành một lựa chọn hấp dẫn khi dữ liệu có tính phi tuyến mạnh, chứa nhiều biến tương tác phức tạp, hoặc khi các mối quan hệ giữa biến là không rõ ràng hoặc không thể biểu diễn theo một công thức hàm cụ thể (Smith et al., 2013).

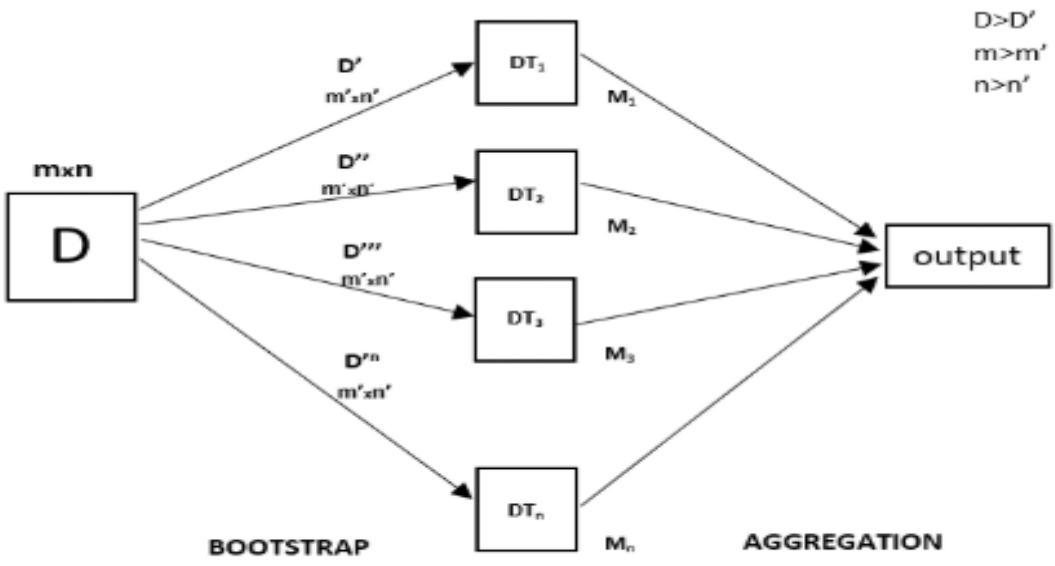
Một đặc điểm quan trọng của RFR là khả năng nội tại trong việc xử lý tương tác giữa các biến đầu vào mà không cần xây dựng trước các biến tương tác (interaction terms). Ngoài ra, RFR cũng hỗ trợ đánh giá mức độ quan trọng của từng biến đầu vào

(variable importance), từ đó hỗ trợ việc lựa chọn biến (feature selection) mà không cần các bước xử lý thủ công như trong MLR.

❖ Cơ chế và cấu trúc hoạt động

Trước hết, từ tập dữ liệu huấn luyện ban đầu ký hiệu là D, có kích thước m×n (trong đó m là số lượng quan sát và n là số lượng biến đầu vào), mô hình tiến hành tạo ra nhiều tập con dữ liệu bằng phương pháp lấy mẫu bootstrap có hoàn lại. Các tập con này được ký hiệu lần lượt là D′,D″,D‴,...,D^{Γ},với mỗi tập có kích thước nhỏ hơn

Hình 1.1. Mô hình hồi quy Rừng ngẫu nhiên hoạt động



Mô hình hồi quy rừng ngẫu nhiên hoạt động

(Nguồn: Geeksforgeeks.org)

Trước hết, từ tập dữ liệu huấn luyện ban đầu ký hiệu là D, có kích thước m×n (trong đó m là số lượng quan sát và n là số lượng biến đầu vào), mô hình tiến hành tạo ra nhiều tập con dữ liệu bằng phương pháp lấy mẫu bootstrap có hoàn lại. Các tập con này được ký hiệu lần lượt là D′,D″,D‴,...,D^{Γ},với mỗi tập có kích thước nhỏ hơn tập dữ liệu gốc. Việc sử dụng kỹ thuật bootstrap giúp mỗi cây hồi quy trong rừng được huấn luyện trên một phân phối dữ liệu khác nhau, góp phần tạo nên tính ngẫu nhiên và giảm thiểu hiện tượng overfitting.

Sau khi tạo ra các tập bootstrap, từng tập dữ liệu con sẽ được sử dụng để huấn luyện một cây hồi quy riêng biệt, ký hiệu là $DT_1, DT_{2,\dots} DT_r$.Trong quá trình xây dựng mỗi cây, tại mỗi nút chia tách, mô hình không sử dụng toàn bộ các biến đầu vào, mà thay vào đó là một tập con ngẫu nhiên của các biến. Đây chính là điểm khác biệt cốt lõi

so với cây quyết định truyền thống và là yếu tố làm tăng tính độc lập giữa các cây. Nhờ đó, Random Forest có thể tận dụng tốt ưu điểm của tập thể mô hình (ensemble) để cải thiện khả năng tổng quát hóa.

Sau khi toàn bộ các cây hồi quy đã được huấn luyện xong, mô hình chuyển sang giai đoạn thứ hai là tổng hợp kết quả. Đối với mỗi điểm dữ liệu mới cần dự đoán, các cây trong rừng sẽ lần lượt cho ra một giá trị dự đoán riêng. Cuối cùng, mô hình sẽ lấy trung bình các dự đoán này để tạo ra kết quả đầu ra cuối cùng. Quá trình này giúp làm giảm phương sai của mô hình tổng thể, từ đó nâng cao độ chính xác và độ ổn định của dự báo.

1.2.4. Ứng dụng Random Forest Regressor trong dự giá cổ phiếu

Trong lĩnh vực tài chính, việc dự báo giá cổ phiếu là một bài toán phức tạp do chịu ảnh hưởng của nhiều yếu tố phi tuyến tính, nhiễu loạn và thường xuyên biến động theo thời gian. Đặc biệt, mối quan hệ giữa các biến kinh tế vĩ mô, chỉ số tài chính nội tại của doanh nghiệp, yếu tố tâm lý thị trường và giá cổ phiếu thường mang tính phi tuyến và ẩn chứa nhiều tương tác phức tạp. Chính vì vậy, các mô hình hồi quy tuyến tính truyền thống như Multiple Linear Regression (MLR) thường không thể hiện hiệu quả tốt khi phải xử lý dữ liệu có tính chất phi tuyến hoặc có nhiễu cao. Trong bối cảnh đó, Random Forest Regression (RFR) nổi lên như một công cụ hữu hiệu nhờ khả năng học mô hình phi tuyến, kháng nhiễu tốt và linh hoạt trong việc xử lý dữ liệu có cấu trúc đa chiều.

Mô hình Random Forest Regression, với cơ chế tổng hợp đầu ra từ nhiều cây hồi quy được xây dựng trên các tập dữ liệu con (bootstrap samples), có khả năng giảm thiểu hiện tượng overfitting vốn rất phổ biến trong các bài toán tài chính. Mỗi cây hồi quy trong rừng sẽ học được một khía cạnh khác nhau của dữ liệu, từ đó giúp cải thiện độ chính xác tổng thể của mô hình khi lấy trung bình kết quả đầu ra. Điều này đặc biệt hữu ích trong các trường hợp dữ liệu chứa nhiều nhiễu (noise) – một đặc điểm thường thấy trong dữ liệu giá cổ phiếu do tác động từ các yếu tố bên ngoài như tin tức, chính sách hoặc hiệu ứng thị trường.

Ngoài ra, điểm mạnh vượt trội của RFR so với các mô hình khác nằm ở khả năng xử lý dữ liệu có chiều cao mà không yêu cầu các giả định nghiêm ngặt về phân phối, tính tuyến tính hay quan hệ độc lập giữa các biến. Trong thực tiễn, dữ liệu tài chính thường bao gồm hàng chục hoặc hàng trăm chỉ số kỹ thuật, chỉ số tài chính, hoặc đặc trưng thời gian (time-lag features) – việc lựa chọn đặc trưng phù hợp cho mô hình là rất

quan trọng. RFR không chỉ thực hiện dự báo, mà còn cung cấp thước đo về mức độ quan trọng của từng biến đầu vào (feature importance), giúp nhà nghiên cứu xác định đâu là những yếu tố có ảnh hưởng lớn nhất đến giá cổ phiếu.

Nhiều nghiên cứu thực nghiệm đã chỉ ra rằng, trong các bài toán dự báo giá cổ phiếu, RFR thường cho kết quả tốt hơn so với các phương pháp tuyến tính như MLR hoặc hồi quy Ridge, đồng thời có tính ổn định hơn so với các mô hình nhạy cảm với tham số như Support Vector Regression (SVR) hoặc các mô hình mạng nơ-ron. RFR đặc biệt tỏ ra hiệu quả khi được áp dụng trên dữ liệu có khối lượng lớn, độ biến động cao, và mối quan hệ phức tạp giữa các yếu tố ảnh hưởng – những đặc điểm rất phổ biến trong thị trường chứng khoán.

1.2.5. Ưu nhược điểm của Random Forest Regression

❖ Ưu điểm của Random Forest Regression

Random Forest Regression (RFR) là một phương pháp học máy thuộc nhóm mô hình tổ hợp (ensemble learning), được xây dựng dựa trên việc kết hợp nhiều cây hồi quy (regression trees) độc lập để cải thiện độ chính xác dự đoán. Trong bối cảnh bài toán dự báo giá cổ phiếu – vốn có tính biến động cao, phi tuyến và nhạy cảm với nhiều yếu tố kinh tế vĩ mô, vi mô – mô hình RFR cho thấy nhiều ưu thế vượt trội so với các phương pháp hồi quy truyền thống.

➢ Khả năng mô hình hóa các mối quan hệ phi tuyến, phức tạp.

Không giống như hồi quy tuyến tính nhiều biến (Multiple Linear Regression – MLR) vốn yêu cầu giả định tuyến tính giữa biến đầu vào và đầu ra, RFR có khả năng linh hoạt học được cả các mối quan hệ phi tuyến mà không cần bất kỳ giả định thống kê nào về phân phối dữ liệu. Trong thực tế, thị trường tài chính thường chịu tác động của nhiều yếu tố tương tác phức tạp, phi tuyến (ví dụ: tâm lý nhà đầu tư, chính sách vĩ mô, thông tin nội bộ doanh nghiệp,...). Việc sử dụng RFR giúp mô hình hóa các mối quan hệ này một cách tự nhiên và chính xác hơn.

➢ Hạn chế hiện tượng overfitting.

Một điểm nổi bật khác là RFR có tính chất kháng overfitting cao. Nhờ cơ chế bootstrap aggregating (bagging), mỗi cây trong rừng được huấn luyện trên một tập con ngẫu nhiên của tập dữ liệu gốc, sau đó tổng hợp dự báo thông qua trung bình (với bài toán hồi quy). Việc tổng hợp này giúp làm mượt các dự đoán, giảm thiểu sự ảnh hưởng của nhiễu và tránh học thuộc đặc trưng cục bộ – một vấn đề phổ biến trong các mô hình.

- Đánh giá tầm quan trọng của biến đầu vào

RFR cung cấp khả năng đánh giá tầm quan trọng của từng biến độc lập đối với biến mục tiêu thông qua các chỉ số như Gini importance hay permutation importance. Đây là một tính năng đặc biệt hữu ích trong bài toán dự báo giá cổ phiếu, khi nhà nghiên cứu muốn xác định những yếu tố ảnh hưởng mạnh nhất đến biến động giá (ví dụ: lợi nhuận ròng, khối lượng giao dịch, tỷ lệ P/E, v.v.). Qua đó, mô hình không chỉ đóng vai trò dự báo mà còn hỗ trợ quá trình phân tích dữ liệu tài chính chuyên sâu.

- Hiệu quả trong xử lý dữ liệu có nhiều chiều và chứa nhiễu

Trong các bài toán dữ liệu thực, đặc biệt là trong tài chính, không hiếm gặp hiện tượng tập dữ liệu có rất nhiều biến đầu vào (high dimensionality), hoặc chứa nhiều yếu tố không liên quan, gây nhiễu. Random Forest, thông qua cơ chế lựa chọn ngẫu nhiên tập con biến ở mỗi node chia nhánh, đã chứng minh khả năng hoạt động hiệu quả ngay cả trong môi trường dữ liệu "nhiễu cao". Ngoài ra, mô hình không yêu cầu chuẩn hóa dữ liệu đầu vào, có thể làm việc trực tiếp với dữ liệu rời rạc (categorical) lẫn liên tục (continuous).

- Tính ổn định và khả năng mở rộng

RFR có khả năng hoạt động ổn định ngay cả khi có sự biến đổi nhẹ trong dữ liệu. Mô hình cũng rất dễ mở rộng bằng cách tăng số lượng cây hoặc áp dụng song song hóa (parallelization) – giúp rút ngắn thời gian huấn luyện mà không làm mất tính toàn vẹn của mô hình.

❖ Nhược điểm của Random Forest Regression

Bên cạnh các ưu điểm nêu trên, mô hình Random Forest Regression vẫn tồn tại một số hạn chế nhất định, đặc biệt khi xét đến các yêu cầu về diễn giải mô hình và chi phí tính toán:

- Khó diễn giải và thiếu tính minh bạch

Một trong những nhược điểm chính của RFR là tính "hộp đen" (black-box) của mô hình. Do kết quả dự đoán là sự tổng hợp của hàng trăm, thậm chí hàng nghìn cây con, việc phân tích ảnh hưởng cụ thể của từng biến đầu vào đến biến mục tiêu là khá phức tạp. Điều này gây khó khăn khi người dùng cần diễn giải mô hình cho các bên liên quan (ví dụ: nhà đầu tư, doanh nghiệp, cơ quan quản lý tài chính), hoặc khi cần thiết lập chính sách dựa trên kết quả phân tích.

- Chi phí tính toán lớn

Việc xây dựng và huấn luyện nhiều cây hồi quy trên các mẫu bootstrap yêu cầu tài nguyên tính toán lớn, đặc biệt khi số lượng quan sát hoặc biến đầu vào tăng cao. Điều này có thể gây khó khăn trong trường hợp triển khai mô hình trên môi trường hạn chế phần cứng, hoặc yêu cầu thời gian phản hồi nhanh (ví dụ: hệ thống giao dịch tự động thời gian thực).

- Kém hiệu quả khi dự báo giá trị ngoài phạm vi dữ liệu huấn luyện

Do đặc trưng sử dụng trung bình các kết quả từ tập huấn luyện, RFR hoạt động rất tốt trong phạm vi dữ liệu đã quan sát. Tuy nhiên, khi cần dự báo các giá trị ngoài phạm vi này (extrapolation), mô hình không thể suy luận được các xu hướng mới, dẫn đến kết quả kém chính xác. Đây là một điểm hạn chế rõ rệt so với các mô hình dựa trên giả định hàm (như hồi quy tuyến tính) có khả năng mở rộng kết quả.

- Giảm độ nhạy với giá trị cực trị

Vì bản chất trung bình kết quả đầu ra từ nhiều cây, RFR có xu hướng làm "mềm hóa" các giá trị dự đoán. Trong bài toán giá cổ phiếu – nơi biến động đột ngột hoặc các cú sốc thị trường thường có ý nghĩa chiến lược – việc đánh giá thấp giá trị cực trị có thể dẫn tới sai lệch đáng kể trong kết luận đầu tư.

❖ Tính phù hợp với đề tài nghiên cứu

Mặc dù tồn tại những hạn chế nêu trên, Random Forest Regression vẫn được lựa chọn là một trong hai mô hình chủ đạo trong đề tài này bởi các lý do sau:

Tính thực tiễn cao và không yêu cầu giả định khắt khe: Trong khi MLR yêu cầu các giả định nghiêm ngặt về tuyến tính, phân phối chuẩn, độc lập và đồng phương sai của phần dư – những điều kiện thường không thỏa mãn trong dữ liệu tài chính – thì RFR hoàn toàn không yêu cầu các điều kiện này. Điều này làm tăng tính ứng dụng mô hình trong các tập dữ liệu thực tế.

Khả năng xử lý quan hệ phi tuyến và biến tương tác: Với dữ liệu kinh tế tài chính vốn không tuân theo nguyên tắc tuyến tính rõ ràng, việc sử dụng mô hình phi tham số như RFR là hoàn toàn hợp lý. Hơn nữa, các biến tài chính thường có mối tương quan chéo phức tạp (ví dụ: EPS, ROE, thanh khoản...), và RFR có thể tự động phát hiện, tận dụng các tương tác này.

Khả năng đo lường tầm quan trọng của các chỉ tiêu tài chính: Đây là điểm đặc biệt phù hợp với mục tiêu phụ của đề tài, không chỉ dừng lại ở việc dự báo giá cổ phiếu mà

còn xác định được các yếu tố có ảnh hưởng mạnh mẽ nhất đến biến động giá, giúp nâng cao khả năng ra quyết định cho nhà đầu tư.

Tính ổn định và kháng nhiễu: Dữ liệu giá cổ phiếu thường có nhiễu (noise) lớn và bất ổn định theo thời gian. Việc sử dụng mô hình có khả năng tổng quát hóa tốt như RFR sẽ làm tăng độ chính xác của dự báo so với các mô hình đơn lẻ.

Từ các phân tích trên, có thể kết luận rằng Random Forest Regression là một mô hình hồi quy mạnh, linh hoạt và phù hợp với bài toán dự báo giá cổ phiếu trong đề tài này. Dù tồn tại một số hạn chế về diễn giải và chi phí tính toán, song lợi ích mà mô hình mang lại – đặc biệt trong khả năng xử lý dữ liệu thực tế phức tạp, nhiễu và phi tuyến – hoàn toàn xứng đáng để được sử dụng và đánh giá so sánh trong nghiên cứu học thuật và ứng dụng thực tiễn.

1.3. Tổng quan về Học Sâu (Deep Learning)

1.3.1. Lí do chọn học sâu

Học sâu (Deep Learning) là một nhánh tiên tiến của học máy (Machine Learning), được thiết kế để mô hình hóa các biểu diễn trừu tượng cao trong dữ liệu thông qua nhiều tầng xử lý phi tuyến. Khác với các mô hình học máy truyền thống – vốn thường phụ thuộc vào các đặc trưng được thiết kế thủ công – học sâu cho phép hệ thống học trực tiếp từ dữ liệu thô thông qua việc xây dựng các biểu diễn tầng bậc, trong đó các đặc trưng cấp cao được suy diễn từ các đặc trưng cấp thấp.

Ý tưởng cốt lõi của học sâu bắt nguồn từ mạng nơ-ron nhân tạo (Artificial Neural Networks), nhưng được mở rộng với nhiều tầng ẩn (hidden layers) và các kiến trúc phức tạp hơn như Mạng nơ-ron tích chập (CNN), Mạng nơ-ron hồi tiếp (RNN), Mạng niềm tin sâu (DBN) hay Mạng bộ nhớ dài ngắn hạn (LSTM). Mỗi tầng trong mạng học sâu hoạt động như một khối trích xuất đặc trưng phi tuyến, góp phần làm tăng khả năng biểu diễn và học của toàn hệ thống.

Khái niệm học sâu lần đầu được đưa vào nghiên cứu một cách bài bản nhờ công trình của Geoffrey Hinton vào năm 2006, trong đó ông đề xuất cơ chế tiền huấn luyện tầng (layer-wise pretraining) để khắc phục những khó khăn trong quá trình huấn luyện các mạng nơ-ron sâu. Kể từ đó, học sâu đã nhanh chóng trở thành trung tâm trong lĩnh vực trí tuệ nhân tạo hiện đại, với những thành công đột phá trong nhận dạng hình ảnh, xử lý ngôn ngữ và đặc biệt là trong các bài toán dự báo và phân loại với dữ liệu lớn.

Một trong những điểm nổi bật của học sâu là khả năng khai thác dữ liệu lớn (Big Data) để học được các đặc trưng phức tạp mà không cần đến sự can thiệp của con người trong quá trình trích chọn đặc trưng. Hơn nữa, các mô hình học sâu cũng thể hiện khả năng khái quát hóa mạnh mẽ trong các không gian phi tuyến cao chiều – điều mà các mô hình học nông (shallow learning) thường không thể làm tốt.

Với sự phát triển của phần cứng tính toán như GPU và các nền tảng phân tán, học sâu ngày càng chứng minh được vai trò cốt lõi trong việc giải quyết các bài toán học máy phức tạp trong nhiều lĩnh vực khác nhau – từ nhận diện khuôn mặt, nhận dạng tiếng nói, dịch máy cho đến dự báo tài chính và thị trường chứng khoán. Trong phần tiếp theo, khóa luận sẽ trình bày cụ thể về các kiến trúc mạng học sâu phổ biến cũng như ứng dụng của mô hình mạng LSTM trong dự báo giá cổ phiếu của Tổng công ty Viglacera - CTCP.

1.3.2. Tổng quan mạng LSTM (Long Short-Term Memory)

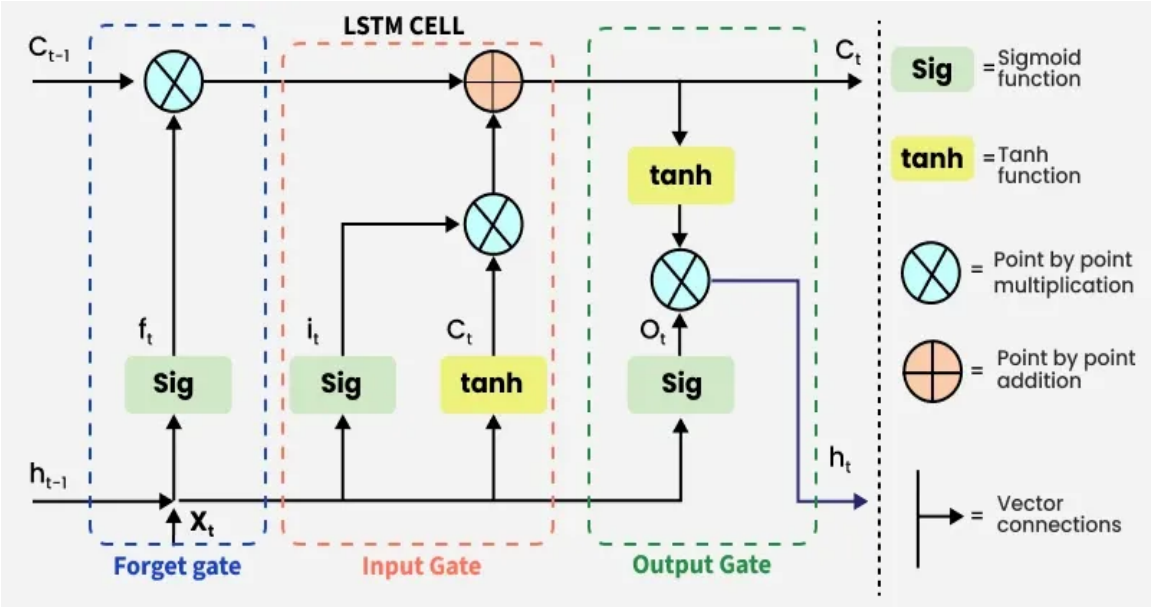
Long Short-Term Memory (LSTM) là một kiến trúc mạng nơ-ron hồi tiếp (Recurrent Neural Network - RNN) được phát triển nhằm khắc phục một trong những hạn chế lớn nhất của các mô hình RNN truyền thống: hiện tượng suy giảm đạo hàm (vanishing gradient) trong quá trình học các phụ thuộc dài hạn. LSTM được đề xuất bởi Hochreiter và Schmidhuber vào năm 1997 như một cải tiến quan trọng giúp mạng hồi tiếp có khả năng ghi nhớ và truyền tải thông tin trong chuỗi thời gian một cách ổn định và hiệu quả hơn so với RNN thông thường.

Khác với mạng truyền thẳng (feedforward neural networks) vốn chỉ xử lý các đầu vào có kích thước cố định, RNN – và đặc biệt là LSTM – có thể xử lý chuỗi dữ liệu có độ dài biến đổi, phù hợp cho các bài toán thời gian như dự báo chuỗi, xử lý ngôn ngữ tự nhiên, và phân tích dữ liệu tín hiệu sinh học. Trong các mô hình LSTM, khả năng ghi nhớ thông tin tại các thời điểm khác nhau trong chuỗi được kiểm soát bởi một loạt các cổng (gates), cho phép mô hình học được mối liên hệ cả ngắn hạn và dài hạn.

Khả năng học phụ thuộc thời gian của LSTM khiến nó trở thành một công cụ mạnh mẽ trong các tác vụ như dự đoán nhiệt độ bề mặt biển, giá cổ phiếu, dữ liệu cảm biến môi trường hoặc bất kỳ hệ thống nào có bản chất thời gian. Đặc biệt, trong các nghiên cứu gần đây như bài báo của Yang et al. (2018), LSTM còn được kết hợp với mạng tích chập (CNN) để tạo thành mô hình CFCC-LSTM, cho phép xử lý đồng thời thông tin không gian và thời gian trong chuỗi dữ liệu đa chiều như ảnh vệ tinh, nhận diện hình ảnh.

❖ Cấu trúc LSTM và cơ chế hoạt động

Hình 1.2. Cấu trúc mô hình LSTM



(nguồn: geeksforgeeks.org)

Long Short-Term Memory (LSTM) là một loại mạng nơ-ron hồi tiếp (Recurrent Neural Network - RNN) có cấu trúc đặc biệt nhằm giải quyết vấn đề về độ dài phụ thuộc trong chuỗi thời gian. Khác với RNN thông thường, LSTM có khả năng học và lưu giữ thông tin dài hạn nhờ vào một kiến trúc nội bộ phức tạp gồm ba cổng điều khiển và một trạng thái bộ nhớ. Tại mỗi bước thời gian t, LSTM sử dụng đầu vào hiện tại t và trạng thái ẩn từ bước trước h_{t-1} để điều chỉnh trạng thái bộ nhớ ct và đầu ra ht

➢ Cổng quên

Cổng quên là cơ chế đầu tiên trong LSTM, chịu trách nhiệm loại bỏ thông tin không còn cần thiết từ trạng thái bộ nhớ trước đó c_{t-1}. Hai đầu vào x_t và h_{t-1} được kết hợp và đưa vào hàm sigmoid để tạo thành vectơ $f_t \in [0,1]$ đại diện cho xác suất giữ lại thông tin tại mỗi chiều:

$$\boldsymbol{f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)}$$

Trong đó:

W_f: Trọng số cổng quên.

b_f: Bias của cổng quên

σ: hàm kích hoạt sigmoid.

Nếu một thành phần trong f_t gần bằng 0, thông tin tương ứng tại vị trí đó trong c_{t-1} sẽ bị "quên". Ngược lại, nếu gần bằng 1, thông tin sẽ được giữ lại cho bước kế tiếp.

- Cổng vào

Cổng vào kiểm soát việc bổ sung thông tin mới vào trạng thái bộ nhớ. Quá trình này gồm hai bước:

Đầu tiên, một cổng sigmoid được sử dụng để xác định mức độ quan trọng của thông tin mới:

$$\boldsymbol{i_t}=\sigma(\boldsymbol{W_i}\cdot[\boldsymbol{h_{t-1}},\boldsymbol{x_t})]+\boldsymbol{b_i}))$$

Tiếp theo, một vectơ ứng viên trạng thái mới $\tilde{c}_t$ tạo ra qua hàm tanh:

$$\tilde{\mathbf{c}}_\mathbf{t}=\mathbf{tanh}(\boldsymbol{W_C}\cdot[\boldsymbol{h_{t-1}},\boldsymbol{x_t})]+\boldsymbol{b_C}).$$

Sau đó, trạng thái bộ nhớ được cập nhật bằng cách kết hợp thông tin đã chọn từ c_{t-1}với ứng viên mới:

$$\boldsymbol{C_t}=\boldsymbol{f_t}\cdot\boldsymbol{C_{t-1}}+\boldsymbol{i_t}\cdot\tilde{\mathbf{c}}_\mathbf{t}$$

Trong đó, ⊙ biểu thị phép nhân từng phần tử (element-wise multiplication). Cơ chế này cho phép LSTM "quên" có chọn lọc và "học" những thông tin mới có giá trị.

- Cổng đầu ra

Cổng đầu ra xác định phần thông tin nào từ trạng thái bộ nhớ c_t sẽ được sử dụng để tạo ra đầu ra h_t, đồng thời cung cấp trạng thái ẩn cho bước thời gian tiếp theo:

Tính toán hệ số cổng:

$$\boldsymbol{o_t}=\sigma(\boldsymbol{W_o}\cdot[\boldsymbol{h_{t-1}},\boldsymbol{x_t}]+\boldsymbol{b_o})$$

Sau đó, đầu ra được tính bằng cách nhân oto_tot với trạng thái bộ nhớ được chuẩn hóa qua hàm tanh:

$$\boldsymbol{h_t}=\boldsymbol{o_t}\cdot\mathbf{tanh}(\boldsymbol{C_t})$$

Kết quả là một vectơ đầu ra có chọn lọc, đồng thời giữ lại thông tin quan trọng từ trạng thái nội tại.

1.3.3. Ứng dụng của LSTM trong dự báo giá cổ phiếu

Trong những năm gần đây, mô hình mạng nơ-ron hồi tiếp với bộ nhớ dài ngắn hạn (Long Short-Term Memory – LSTM) đã trở thành một công cụ nổi bật trong việc xử lý và dự báo chuỗi thời gian phi tuyến. Trong lĩnh vực tài chính, đặc biệt là trong bài toán dự báo giá cổ phiếu, LSTM được đánh giá là một phương pháp có hiệu quả cao nhờ khả năng ghi nhớ thông tin trong quá khứ và giảm thiểu hiện tượng mất thông tin dài hạn – một điểm yếu cố hữu của các mạng hồi tiếp truyền thống (simple RNNs).

Bản chất của dữ liệu tài chính, chẳng hạn như giá cổ phiếu, thường mang tính chất nhiễu, phi tuyến và phụ thuộc mạnh mẽ vào yếu tố thời gian. Việc sử dụng các mô hình tuyến tính cổ điển như ARIMA, GARCH hoặc các phương pháp học máy truyền thống như Random Forest hay SVM thường gặp giới hạn trong việc nắm bắt các mối quan hệ phức tạp và phi tuyến giữa các yếu tố đầu vào và biến mục tiêu. Trong bối cảnh đó, mô hình LSTM, với cấu trúc đặc biệt bao gồm các cổng quên (forget gate), cổng vào (input gate) và cổng ra (output gate), cho phép kiểm soát luồng thông tin trong quá trình huấn luyện, từ đó có thể học được các mối quan hệ dài hạn mà không làm mất đi tính ổn định trong mạng.

Các nghiên cứu gần đây đã chỉ ra rằng LSTM có khả năng ứng dụng hiệu quả trong việc dự báo các chỉ số giá cổ phiếu như giá đóng cửa (close price), giá mở cửa (open price), giá cao nhất/thấp nhất trong phiên (high/low), hoặc các chỉ báo kỹ thuật (technical indicators) như RSI, MACD, v.v. Bên cạnh đó, một số công trình cũng áp dụng LSTM để dự đoán xu hướng biến động (tăng/giảm), xác định điểm đảo chiều hoặc đánh giá rủi ro thông qua độ biến động (volatility) của cổ phiếu trong tương lai.

Đặc biệt, trong môi trường thực tế nơi các yếu tố vĩ mô, tin tức kinh tế và tâm lý thị trường có thể tác động đồng thời đến biến động giá cổ phiếu, LSTM còn có khả năng tích hợp các yếu tố đầu vào đa chiều để tăng cường tính toàn diện trong dự báo. Ngoài ra, mô hình cũng có thể mở rộng theo chiều sâu thông qua việc xếp chồng nhiều lớp LSTM (stacked LSTM) để trích xuất các đặc trưng ở nhiều cấp độ khác nhau trong dữ liệu.

Từ những phân tích trên có thể thấy, mô hình LSTM không chỉ là một công cụ tiên tiến trong xử lý chuỗi thời gian, mà còn là một giải pháp tiềm năng trong dự báo giá cổ phiếu – vốn là một bài toán có độ phức tạp cao, đòi hỏi khả năng học sâu, linh hoạt và thích nghi tốt với sự biến động liên tục của thị trường tài chính. Trong khuôn khổ khóa

luận này, mô hình LSTM sẽ được triển khai nhằm dự báo giá cổ phiếu của Tổng Công ty Viglacera – CTCP, và so sánh hiệu quả với mô hình Random Forest nhằm đánh giá ưu điểm của học sâu trong phân tích và dự báo tài chính.

1.3.4. Ưu nhược điểm của LSTM

❖ Ưu điểm

Long Short-Term Memory (LSTM) là một kiến trúc đặc biệt của mạng nơ-ron hồi tiếp (Recurrent Neural Network – RNN), được thiết kế nhằm khắc phục các hạn chế cố hữu trong việc xử lý và học từ các chuỗi dữ liệu dài hạn. Trong bối cảnh dự báo tài chính, đặc biệt là dự báo giá cổ phiếu – một bài toán đặc trưng của chuỗi thời gian phi tuyến và nhiễu động cao – LSTM thể hiện nhiều ưu điểm đáng kể:

Trước hết, khả năng ghi nhớ dài hạn là đặc trưng cốt lõi giúp LSTM vượt trội hơn so với các mô hình RNN truyền thống. Thông qua cơ chế bộ nhớ nội tại (cell state) được điều tiết bởi ba cổng (input gate, forget gate và output gate), LSTM có thể lưu giữ các thông tin quan trọng từ quá khứ và loại bỏ các yếu tố không còn phù hợp với dự báo tương lai. Cơ chế này đặc biệt hữu ích trong bối cảnh giá cổ phiếu có thể chịu ảnh hưởng bởi các sự kiện, xu hướng diễn ra cách thời điểm hiện tại nhiều phiên giao dịch.

Tiếp theo, LSTM có khả năng học và mô hình hóa các mối quan hệ phi tuyến phức tạp trong dữ liệu đầu vào, một đặc tính phù hợp với bản chất biến động của giá cổ phiếu – vốn thường bị chi phối bởi nhiều yếu tố ẩn không thể biểu diễn bằng các hàm tuyến tính. Nhờ đó, LSTM cho phép mô hình hóa tốt hơn các xu hướng ẩn và biến động không rõ ràng.

Thêm vào đó, tính linh hoạt và khả năng mở rộng cao của LSTM giúp mô hình dễ dàng tích hợp vào các kiến trúc học sâu đa tầng hoặc các hệ thống lai (chẳng hạn kết hợp với mạng CNN, attention mechanism, hoặc các lớp embedding). Điều này mở ra tiềm năng nâng cao chất lượng dự báo trong các kịch bản có tính biến động cao và dữ liệu phức hợp như thị trường chứng khoán.

Cuối cùng, LSTM đã được áp dụng và kiểm chứng rộng rãi trong các nghiên cứu thực nghiệm trên lĩnh vực tài chính, bao gồm dự báo giá cổ phiếu, chỉ số thị trường, và các chỉ báo kỹ thuật. Việc kế thừa các kết quả từ các nghiên cứu trước đó giúp đảm bảo tính thực tiễn và độ tin cậy khi triển khai vào bài toán cụ thể với cổ phiếu của Tổng Công ty Viglacera – CTCP.

❖ Nhược điểm

Bên cạnh các ưu điểm nổi bật, mô hình LSTM vẫn tồn tại một số hạn chế cần được lưu ý trong quá trình triển khai thực tiễn.

Thứ nhất, do cấu trúc mạng chứa nhiều tham số cần huấn luyện, LSTM đòi hỏi một tập dữ liệu huấn luyện đủ lớn, đồng nhất và có chất lượng cao để tránh tình trạng quá khớp (overfitting) hoặc mất ổn định trong quá trình huấn luyện. Trong bối cảnh dữ liệu tài chính thường xuyên biến động và không ổn định, việc đảm bảo chất lượng dữ liệu là một yêu cầu then chốt.

Thứ hai, LSTM có thời gian huấn luyện dài hơn đáng kể so với các mô hình truyền thống (chẳng hạn ARIMA, hồi quy tuyến tính) do cấu trúc nhiều tầng và quy trình lan truyền ngược qua thời gian (Backpropagation Through Time – BPTT) phức tạp. Điều này yêu cầu hệ thống phần cứng mạnh và thời gian huấn luyện được tối ưu hợp lý.

Ngoài ra, mô hình LSTM thường được xem là hộp đen (black-box) trong phân tích – tức khó lý giải được vì sao mô hình đưa ra một dự báo cụ thể, từ đó gây khó khăn trong việc diễn giải và thuyết minh với các bên liên quan không chuyên về kỹ thuật.

❖ Tính phù hợp với đề tài

Với đặc trưng của bài toán dự báo giá cổ phiếu – vốn là bài toán chuỗi thời gian phức tạp, nhiễu động cao và đòi hỏi khả năng ghi nhớ dài hạn, mô hình LSTM được đánh giá là hoàn toàn phù hợp và có tiềm năng cao trong việc cải thiện độ chính xác của dự báo.

Cụ thể, trong đề tài "Ứng dụng học máy và học sâu để dự báo giá cổ phiếu của Tổng Công ty Viglacera – CTCP", LSTM đóng vai trò là một mô hình học sâu chủ lực, có khả năng tiếp thu chuỗi giá cổ phiếu lịch sử cùng các chỉ báo kỹ thuật (technical indicators) nhằm đưa ra các dự báo giá trong tương lai gần. Việc kết hợp LSTM với các phương pháp xử lý dữ liệu đầu vào (chuẩn hóa, kỹ thuật sliding window) và tối ưu hóa mô hình giúp đảm bảo LSTM phát huy tối đa khả năng học biểu diễn từ dữ liệu chuỗi, từ đó hỗ trợ ra quyết định đầu tư hiệu quả.

1.4. Các chỉ số đánh giá hiệu quả mô hình dự báo

1.4.1. R^2 Score (Coefficient of Determination)

Trong các mô hình dự báo giá cổ phiếu, đặc biệt là hồi quy đa biến như Random Forest hoặc LSTM xử lý như một mô hình hồi quy chuỗi thời gian, chỉ số R-squared (ký hiệu R^2) thường được sử dụng nhằm đánh giá chất lượng của mô hình. R^2 thể hiện tỷ lệ

phần trăm phương sai của biến phụ thuộc (ví dụ, giá cổ phiếu VGC) được giải thích bởi các biến độc lập có trong mô hình. Về bản chất, R^2 đo lường mức độ phù hợp giữa mô hình hồi quy và dữ liệu thực nghiệm, từ đó phản ánh năng lực dự báo tương đối của mô hình.

$$\boldsymbol{R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}}$$

R^2 dao động từ 0 đến 1, với giá trị càng cao chứng tỏ mô hình càng giải thích được nhiều biến thiên trong dữ liệu. Nếu $R^2 = 0.65$, có nghĩa là 65% biến động trong giá cổ phiếu có thể được mô hình giải thích thông qua các biến đầu vào.

❖ Ý nghĩa thực tiễn trong dự báo giá cổ phiếu

Trong bối cảnh dự báo giá cổ phiếu, R^2 giúp đánh giá mô hình không chỉ theo nghĩa tuyệt đối mà còn theo tiêu chuẩn tương đối: mô hình mang lại mức cải thiện bao nhiêu trong dự báo so với việc không có mô hình (tức sử dụng giá trị trung bình). Đây là lý do khiến R^2 thường được gọi là thước đo "năng lực dự báo tương đối".

Chẳng hạn, nếu không có thông tin nào về các yếu tố ảnh hưởng, lựa chọn dự báo tối ưu có thể chỉ là lấy giá trung bình trong quá khứ. Khi đưa vào các biến như khối lượng giao dịch, tỷ lệ P/E, chỉ số thị trường, mô hình hồi quy có thể cải thiện độ chính xác dự báo – mức độ cải thiện này được phản ánh bởi R^2.

❖ Hạn chế và các yếu tố cần điều chỉnh

Tuy nhiên, R^2 không phải là chỉ số hoàn hảo. Một trong những hạn chế phổ biến là nó luôn tăng khi ta thêm biến độc lập vào mô hình, kể cả khi các biến này không có giá trị thực tiễn. Điều này có thể dẫn đến hiện tượng quá khớp (overfitting), khiến mô hình thể hiện tốt trên dữ liệu huấn luyện nhưng kém hiệu quả trên dữ liệu mới.

Ngoài ra, R^2 không cung cấp thông tin về sai số dự báo tuyệt đối. Một mô hình có R^2 cao nhưng sai số lớn (ví dụ RMSE hoặc MAE cao) vẫn có thể không phù hợp với mục tiêu thực tế. Do đó, R^2 nên được sử dụng cùng với các chỉ số khác như MAE, RMSE và MAPE để có đánh giá toàn diện hơn.

Trong các nghiên cứu có mục tiêu giải thích tác động của biến độc lập lên biến phụ thuộc, R^2 thường không đóng vai trò quan trọng bằng các hệ số hồi quy và kiểm định ý nghĩa thống kê. Tuy nhiên, khi mục tiêu là dự báo, đặc biệt trong bối cảnh thị trường tài

chính biến động mạnh, R^2 là công cụ hữu hiệu giúp đánh giá khả năng của mô hình trong việc giảm thiểu sai số dự báo so với các phương án đơn giản như dự đoán trung bình.

Trong so sánh giữa các mô hình hồi quy khác nhau, đặc biệt khi áp dụng cùng một bộ dữ liệu, R^2 là chỉ số phù hợp để đánh giá tương đối năng lực dự báo. Tuy nhiên, nếu dữ liệu khác nhau (ví dụ giữa các giai đoạn hoặc thị trường), việc so sánh R^2 cần được thực hiện thận trọng do sự thay đổi trong phân phối và phương sai của biến đầu vào có thể ảnh hưởng đến kết quả.

1.4.2. MAE (Mean Absolute Error)

Trong lĩnh vực dự báo chuỗi thời gian tài chính, việc lựa chọn và sử dụng các chỉ số đánh giá mô hình đóng vai trò thiết yếu trong quá trình kiểm định hiệu quả và độ chính xác của các thuật toán học máy. Một trong những chỉ số phổ biến và có tính diễn giải cao là MAE (Mean Absolute Error) – sai số tuyệt đối trung bình. Chỉ số này đặc biệt hữu ích trong các bài toán dự báo thực tiễn như dự báo giá cổ phiếu, do khả năng phản ánh trực tiếp độ sai lệch trung bình giữa giá trị thực và giá trị dự báo.

$$\mathbf{MAE} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} |\, \boldsymbol{y_i} - \boldsymbol{\hat{y}_i} \,|$$

Chỉ số MAE phản ánh sai số dự báo trung bình mà không phân biệt sai lệch mang dấu dương hay âm. Việc sử dụng giá trị tuyệt đối giúp đảm bảo rằng các sai số không triệt tiêu lẫn nhau, từ đó cung cấp một thước đo rõ ràng về độ chính xác tổng thể của mô hình.

❖ Vai trò và ý nghĩa trong dự báo cổ phiếu

Trong bối cảnh dự báo giá cổ phiếu, chỉ số MAE có thể được hiểu như mức chênh lệch trung bình (tính theo đơn vị giá) giữa giá trị thực tế trên thị trường và giá trị mà mô hình dự đoán. Khác với các chỉ số phức tạp hơn như RMSE hay MAPE, MAE cho phép nhà nghiên cứu và nhà đầu tư đánh giá trực tiếp hiệu năng mô hình thông qua con số cụ thể, từ đó dễ dàng đưa ra nhận định thực tiễn.

Chẳng hạn, MAE = 1.25 có thể được diễn giải rằng trung bình mỗi phiên giao dịch, mô hình dự báo lệch khoảng 1.25 đồng so với giá cổ phiếu thực tế – một thông tin có tính định lượng rõ ràng và trực tiếp.

❖ Đặc điểm nổi bật

MAE sở hữu một số đặc điểm kỹ thuật và thống kê đáng lưu ý:

Tính dễ hiểu và dễ diễn giải: Giá trị MAE mang cùng đơn vị với biến phụ thuộc, cho phép người dùng đánh giá trực tiếp mức độ chính xác mà không cần biến đổi hay chuẩn hóa thêm.

Tính ổn định cao trước outliers: Do không sử dụng bình phương sai số như RMSE, MAE không làm phóng đại các sai số lớn, từ đó tránh hiện tượng bị chi phối bởi giá trị dị biệt trong dữ liệu tài chính – một đặc điểm thường thấy trong dữ liệu cổ phiếu.

Độ nhạy thấp với phương sai của phân phối sai số: MAE không giả định phân phối chuẩn của sai số, do đó có thể ứng dụng trong nhiều trường hợp mà các giả định cổ điển của hồi quy tuyến tính không được thỏa mãn hoàn toàn.

❖ Ứng dụng trong đánh giá mô hình dự báo cổ phiếu

Trong khuôn khổ đề tài này, chỉ số MAE được sử dụng để đánh giá và so sánh hiệu suất của hai mô hình học máy là Random Forest Regression và LSTM trong dự báo giá cổ phiếu của Tổng Công ty Viglacera – CTCP. Giá trị MAE càng nhỏ thể hiện mô hình càng chính xác, với mức độ sai số trung bình thấp hơn giữa giá trị dự báo và giá trị thực tế.

Việc phân tích chỉ số MAE song song với các thước đo khác như RMSE, MAPE và R^2 không chỉ giúp đánh giá toàn diện hiệu quả dự báo của các mô hình, mà còn cung cấp góc nhìn đa chiều về sai số theo các tiêu chí khác nhau: độ lệch trung bình tuyệt đối, sai số bình phương, sai số phần trăm tương đối, và mức độ giải thích phương sai. Cách tiếp cận này cho phép không chỉ so sánh định lượng giữa các mô hình mà còn giúp nhận diện đặc điểm sai số phổ biến, từ đó hỗ trợ nhà đầu tư hoặc người sử dụng mô hình trong việc lựa chọn thuật toán phù hợp với mục tiêu sử dụng cụ thể và đặc thù của dữ liệu.

1.4.3. RMSE (Root Mean Squared Error)

RMSE là căn bậc hai của sai số bình phương trung bình. Khác với MAE, RMSE nhấn mạnh vào các sai số lớn, do đó thường được sử dụng khi cần kiểm soát các trường hợp dự báo cực đoan.

$$\boldsymbol{RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}$$

RMSE đặc biệt hữu ích trong tài chính khi các sai lệch lớn gây rủi ro nghiêm trọng, chẳng hạn như sai dự báo mức đỉnh hoặc đáy của cổ phiếu.

Trong bối cảnh dự báo giá cổ phiếu, RMSE có thể được hiểu như mức độ sai lệch dự báo trung bình, tính theo đơn vị gốc của biến phụ thuộc (tức giá cổ phiếu). Tuy nhiên, khác với MAE – vốn đối xử các sai số như nhau – thì RMSE đặc biệt nhạy cảm với sai số lớn**,** điều này khiến nó trở thành chỉ số phù hợp trong các bối cảnh yêu cầu độ chính xác cao, hoặc khi rủi ro từ những sai lệch lớn là không thể chấp nhận được.

❖ Đặc điểm

Tính phản ứng cao với outliers: Do sử dụng bình phương sai số, RMSE phóng đại ảnh hưởng của các sai số lớn, từ đó trở nên nhạy cảm hơn với các điểm dị biệt – một đặc tính vừa là lợi thế, vừa là hạn chế, tùy theo mục tiêu phân tích.

Đơn vị cùng với biến phụ thuộc**:** RMSE có cùng đơn vị đo lường với biến yyy, điều này cho phép diễn giải kết quả một cách trực quan và dễ hiểu trong ngữ cảnh dự báo tài chính.

Thường lớn hơn MAE: Trong phần lớn các trường hợp thực nghiệm, RMSE có xu hướng lớn hơn MAE, trừ khi tất cả sai số đều có cùng độ lớn. Điều này phản ánh rõ hơn ảnh hưởng của các sai số cực đoan.

❖ Vai trò trong đánh giá mô hình dự báo cổ phiếu

Trong đề tài này, RMSE được sử dụng như một chỉ số trọng yếu để đo lường mức độ sai số trung bình có trọng số của các mô hình dự báo, đặc biệt giữa hai phương pháp Random Forest Regression và LSTM. Một mô hình có giá trị RMSE thấp hơn đồng nghĩa với việc mô hình đó có khả năng dự báo ổn định và chính xác hơn, nhất là trong bối cảnh biến động giá cổ phiếu thường không tuân theo phân phối chuẩn và dễ xuất hiện sai số lớn.

Việc sử dụng RMSE kết hợp với MAE, MAPE và R^2 giúp xây dựng một hệ thống đánh giá mô hình toàn diện – cân bằng giữa độ sai lệch trung bình, mức độ ảnh hưởng của các outliers và khả năng lý giải biến động của biến mục tiêu.

1.4.4. MAPE (Mean Absolute Percentage Error)

MAPE là chỉ số biểu thị sai số trung bình dưới dạng phần trăm, cho phép đánh giá mức độ tương đối của sai số dự báo.

$$\boldsymbol{MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|}$$

MAPE thuận tiện khi so sánh hiệu suất mô hình trên các tập dữ liệu có đơn vị hoặc quy mô khác nhau. Tuy nhiên, MAPE không ổn định khi giá trị thực tế tiệm cận về 0 – điều cần lưu ý khi dự báo các cổ phiếu có biến động thấp.

MAPE đặc biệt hữu ích trong việc chuẩn hóa sai số dự báo. Thay vì thể hiện mức độ sai lệch theo đơn vị tuyệt đối (như MAE hay RMSE), MAPE cho biết tỷ lệ sai số dự báo trung bình so với giá trị thực tế. Điều này rất có giá trị trong các bài toán dự báo giá cổ phiếu, bởi giá cổ phiếu có thể biến động trên các mức giá khác nhau theo thời gian. Việc đánh giá sai số theo phần trăm sẽ phản ánh tính hiệu quả của mô hình một cách công bằng hơn, không bị ảnh hưởng bởi quy mô của biến mục tiêu.

❖ Đặc điểm kỹ thuật

Tính dễ hiểu và trực quan: Việc biểu diễn sai số dưới dạng phần trăm giúp người sử dụng mô hình – bao gồm cả các đối tượng không chuyên – dễ dàng nắm bắt mức độ chính xác của mô hình.

Phù hợp để so sánh đa mô hình: Nhờ vào tính chuẩn hóa, MAPE cho phép so sánh hiệu quả giữa các mô hình dự báo khác nhau trên các biến mục tiêu có quy mô khác nhau. Không bị ảnh hưởng bởi đơn vị đo: Vì là một tỷ lệ phần trăm, MAPE không phụ thuộc vào đơn vị gốc của biến mục tiêu, phù hợp để đánh giá hiệu suất trên dữ liệu.

❖ Vai trò trong đánh giá mô hình dự báo cổ phiếu

Trong đề tài này, MAPE được sử dụng như một công cụ để đo lường mức độ chính xác tương đối của các mô hình dự báo giá cổ phiếu. Giá trị MAPE thấp cho thấy mô hình dự báo ổn định và hiệu quả trên phương diện tỷ lệ, phù hợp với yêu cầu đánh giá các mô hình hoạt động trên những biến động có quy mô khác nhau trong thời gian.

Việc kết hợp MAPE với các chỉ số khác như MAE, RMSE và R^2 sẽ cho phép phân tích mô hình từ nhiều góc độ – cả về sai số tuyệt đối, mức độ nhạy với outliers và khả năng lý giải sự biến động của dữ liệu. Đây là tiền đề quan trọng cho việc lựa chọn mô hình tối ưu trong dự báo tài chính.

Kết luận chương 1:

Chương 1 đã trình bày các cơ sở lý thuyết nền tảng phục vụ cho quá trình nghiên cứu và dự báo giá cổ phiếu của Tổng Công ty Viglacera – CTCP. Trước hết, các khái niệm liên quan đến cổ phiếu, phân loại, đặc điểm cũng như các yếu tố ảnh hưởng đến giá cổ phiếu đã được hệ thống hóa một cách đầy đủ. Bên cạnh đó, chương cũng đã khẳng

định vai trò quan trọng của việc dự báo giá cổ phiếu trong bối cảnh thị trường tài chính biến động mạnh và nhà đầu tư ngày càng cần thông tin chính xác để ra quyết định.

Về mặt phương pháp, học máy và học sâu – đại diện bởi mô hình Random Forest Regression và mạng LSTM – đã được giới thiệu với các đặc điểm kỹ thuật, cơ chế hoạt động, ưu nhược điểm và khả năng ứng dụng vào bài toán dự báo giá cổ phiếu. Cuối chương, các chỉ số đánh giá mô hình như R^2, MAE, RMSE và MAPE cũng đã được trình bày nhằm tạo tiền đề cho việc đánh giá khách quan hiệu suất mô hình trong các chương sau.

Từ những cơ sở lý thuyết trên, chương tiếp theo sẽ đi vào phân tích thực trạng tài chính của Tổng Công ty Viglacera – CTCP trong ba năm gần đây, làm nền tảng dữ liệu cho việc huấn luyện và đánh giá các mô hình dự báo.

CHƯƠNG 2: THỰC TRẠNG TỔNG CÔNG TY VIGLACERA – CTCP

2.1. Giới thiệu chung về tổng công ty VIGLACERA

Tổng công ty Viglacera – Công ty Cổ phần (VGC) là một trong những doanh nghiệp hàng đầu tại Việt Nam trong lĩnh vực sản xuất vật liệu xây dựng và đầu tư phát triển bất động sản công nghiệp – đô thị. Được thành lập vào năm 1974, Viglacera có lịch sử phát triển lâu đời, đóng vai trò tiên phong trong ngành công nghiệp vật liệu xây dựng của Việt Nam. Trải qua gần 50 năm hình thành và phát triển, Viglacera đã từng bước khẳng định vị thế vững chắc trên thị trường nội địa, đồng thời mở rộng hoạt động ra thị trường quốc tế, với nhiều sản phẩm được xuất khẩu sang hơn 40 quốc gia và vùng lãnh thổ trên thế giới. Với triết lý kinh doanh gắn liền giữa hiệu quả kinh tế và trách nhiệm xã hội, Viglacera hiện đang sở hữu và vận hành hơn 40 đơn vị thành viên, bao gồm các nhà máy sản xuất vật liệu xây dựng hiện đại và các khu công nghiệp, khu đô thị lớn tại nhiều tỉnh thành trên cả nước.

Trong lĩnh vực vật liệu xây dựng, Viglacera nổi bật với hệ thống sản phẩm đa dạng, bao gồm: gạch ốp lát ceramic, granite, cotto; thiết bị vệ sinh cao cấp; kính xây dựng (gồm kính nổi, kính tiết kiệm năng lượng, kính siêu trắng); gạch bê tông khí chưng áp (AAC); gạch ngói đất sét nung và các loại vật liệu xây dựng thân thiện môi trường khác. Công ty luôn chú trọng đầu tư đổi mới công nghệ, đưa vào vận hành các dây chuyền sản xuất hiện đại theo tiêu chuẩn châu Âu, đáp ứng nhu cầu ngày càng cao của thị trường và xu hướng xây dựng xanh – bền vững.

Song song với sản xuất, Viglacera còn là một trong những đơn vị tiên phong trong phát triển hạ tầng khu công nghiệp và bất động sản đô thị tại Việt Nam. Tổng công ty đã và đang đầu tư, quản lý hơn 10 khu công nghiệp trải dài từ Bắc vào Nam, với tổng diện tích lên đến hàng nghìn hecta, tạo điều kiện thuận lợi cho thu hút đầu tư trong và ngoài nước.

2.1.1. Quy mô công ty

Tính đến thời điểm hiện tại, Tổng công ty Viglacera – CTCP là một trong những doanh nghiệp có quy mô hoạt động lớn mạnh và toàn diện hàng đầu trong lĩnh vực vật liệu xây dựng và bất động sản tại Việt Nam. Viglacera sở hữu hệ thống hơn 40 đơn vị thành viên, bao gồm các công ty con, công ty liên kết và nhiều doanh nghiệp hoạt động

độc lập với vai trò là các công ty cổ phần, trong đó một số đơn vị đã được niêm yết trên thị trường chứng khoán.

Hình 2.1. Hệ thống phân phối của VGC

(Nguồn: viglacera.com.vn)

Trên toàn quốc, Viglacera hiện đang vận hành trên 20 nhà máy sản xuất vật liệu xây dựng, với công suất lớn và công nghệ tiên tiến hàng đầu khu vực. Các nhà máy này chuyên sản xuất các sản phẩm chủ lực như gạch ốp lát ceramic, granite, cotto; thiết bị vệ sinh cao cấp; kính xây dựng (kính nổi, kính tiết kiệm năng lượng, kính siêu trắng); và vật liệu xây dựng không nung thân thiện môi trường. Hệ thống cơ sở sản xuất được phân bổ rộng khắp, giúp Viglacera vừa tối ưu chi phí vận hành, vừa nhanh chóng đáp ứng nhu cầu thị trường trong nước và quốc tế.

Ở lĩnh vực đầu tư phát triển hạ tầng khu công nghiệp, Viglacera là một trong những nhà đầu tư lớn nhất tại Việt Nam, với hơn 10 khu công nghiệp đang hoạt động hiệu quả, có tổng diện tích lên tới hơn 4.000 ha. Các khu công nghiệp của Viglacera được đầu tư bài bản, đồng bộ về hạ tầng kỹ thuật và hạ tầng xã hội, tọa lạc tại những vị trí chiến lược như Bắc Ninh, Quảng Ninh, Hưng Yên, Thái Nguyên, Vĩnh Phúc, Hà Nam, Thừa Thiên – Huế và TP. Hồ Chí Minh. Đây là cơ sở quan trọng giúp Viglacera thu hút hàng trăm doanh nghiệp trong và ngoài nước đến đầu tư sản xuất, đặc biệt là các doanh nghiệp đến từ Nhật Bản, Hàn Quốc và châu Âu.

Không chỉ dừng lại ở công nghiệp, Viglacera còn mở rộng mạnh mẽ trong lĩnh vực phát triển bất động sản đô thị – nhà ở, với nhiều dự án nhà ở xã hội, khu đô thị mới và khu nhà ở thương mại tại các tỉnh thành trọng điểm như Hà Nội, TP. Hồ Chí Minh, Quảng Ninh, Bắc Ninh... Những dự án này không chỉ mang lại hiệu quả kinh doanh mà còn góp phần thực hiện mục tiêu an sinh xã hội và phát triển đô thị văn minh, hiện đại.

2.1.2. Chức năng và nhiệm vụ của công ty

Tổng công ty Viglacera – CTCP (sau đây gọi tắt là Viglacera) là một trong những doanh nghiệp có vị trí hàng đầu tại Việt Nam trong lĩnh vực sản xuất vật liệu xây dựng và phát triển bất động sản công nghiệp – đô thị. Trải qua hơn bốn thập kỷ hình thành và phát triển, Viglacera không ngừng mở rộng quy mô sản xuất, nâng cao chất lượng sản phẩm, đồng thời đóng vai trò quan trọng trong quá trình công nghiệp hóa – hiện đại hóa đất nước. Với vai trò là một tổng công ty cổ phần hoạt động đa ngành, các chức năng và nhiệm vụ của Viglacera được xác định một cách toàn diện và có tính chiến lược lâu dài.

❖ Chức năng chính

- Sản xuất và kinh doanh vật liệu xây dựng

Đây là chức năng cốt lõi và truyền thống của Viglacera kể từ khi thành lập. Doanh nghiệp chuyên sản xuất và phân phối các loại vật liệu xây dựng chủ lực như gạch ốp lát ceramic, granite, sứ vệ sinh, ngói đất sét nung, kính xây dựng (kính nổi, kính tiết kiệm năng lượng), và các thiết bị nhà tắm cao cấp. Các sản phẩm của Viglacera được sản xuất trên dây chuyền hiện đại, đồng bộ, đạt tiêu chuẩn quốc tế, phục vụ cả thị trường trong nước và xuất khẩu ra hơn 40 quốc gia trên thế giới. Việc phát triển và cung ứng đa dạng sản phẩm giúp Viglacera đáp ứng nhu cầu ngày càng cao của ngành xây dựng, từ phân khúc bình dân đến cao cấp.

- Đầu tư và phát triển bất động sản

Bên cạnh lĩnh vực vật liệu xây dựng, Viglacera còn là nhà đầu tư chiến lược trong lĩnh vực bất động sản, đặc biệt là phát triển các khu công nghiệp và khu đô thị. Tính đến nay, Viglacera đã và đang triển khai trên 10 khu công nghiệp lớn trải dài từ Bắc vào Nam như KCN Tiên Sơn, KCN Yên Phong (Bắc Ninh), KCN Phú Hà (Phú Thọ), KCN Đông Mai (Quảng Ninh), KCN Yên Mỹ (Hưng Yên) và rất nhiều các bất động sản khác khắp cả nước. Với định hướng phát triển bền vững, Viglacera chú trọng xây dựng các khu công nghiệp đồng bộ về hạ tầng kỹ thuật và dịch vụ tiện ích nhằm thu hút đầu tư trong và ngoài nước.

Ngoài ra, công ty cũng phát triển các dự án khu đô thị, nhà ở xã hội và nhà ở thương mại trên địa bàn cả nước. Những dự án bất động sản này không chỉ đóng góp vào nguồn cung nhà ở cho người dân mà còn tạo động lực thúc đẩy phát triển kinh tế - xã hội tại các địa phương.

- Nghiên cứu, phát triển và ứng dụng khoa học – công nghệ

Một trong những chức năng quan trọng khác của Viglacera là tiên phong trong nghiên cứu, ứng dụng công nghệ tiên tiến vào sản xuất vật liệu xây dựng. Viglacera là đơn vị đầu tiên tại Việt Nam sản xuất thành công kính tiết kiệm năng lượng và gạch bê tông khí chưng áp – hai dòng sản phẩm thân thiện với môi trường và tiết kiệm năng lượng. Công ty cũng đầu tư mạnh mẽ vào công tác R&D, hợp tác quốc tế để cải tiến mẫu mã, nâng cao chất lượng sản phẩm theo xu hướng xanh, thông minh và bền vững.

- Tổ chức sản xuất kinh doanh theo mô hình công ty mẹ – công ty con

Viglacera hoạt động theo mô hình công ty cổ phần, trong đó Tổng công ty đóng vai trò công ty mẹ quản lý nhiều công ty con chuyên biệt về từng lĩnh vực sản xuất hoặc đầu tư. Mô hình này giúp tăng tính linh hoạt, chuyên môn hóa và hiệu quả trong quản trị doanh nghiệp. Các công ty thành viên hoạt động độc lập về tài chính, song vẫn phối hợp chặt chẽ với Tổng công ty trong định hướng chiến lược và kiểm soát chất lượng sản phẩm.

❖ Nhiệm vụ chủ yếu

- Tổ chức sản xuất kinh doanh hiệu quả

Tổng công ty có nhiệm vụ tổ chức quản lý, vận hành hiệu quả hệ thống nhà máy, cơ sở sản xuất và mạng lưới phân phối trên toàn quốc. Việc đảm bảo năng suất, chất lượng và chi phí hợp lý là mục tiêu xuyên suốt để nâng cao năng lực cạnh tranh trên thị trường nội địa và quốc tế. Viglacera liên tục cải tiến quy trình sản xuất, giảm thiểu thất thoát nguyên vật liệu và tăng cường quản lý theo tiêu chuẩn ISO.

- Đầu tư phát triển hạ tầng khu công nghiệp đồng bộ

Một trong những nhiệm vụ chiến lược của Viglacera là phát triển hạ tầng khu công nghiệp hiện đại, tích hợp tiện ích xanh như xử lý nước thải, năng lượng mặt trời, hệ thống logistics và nhà ở cho công nhân. Những khu công nghiệp này không chỉ tạo điều kiện thuận lợi cho nhà đầu tư mà còn đóng góp tích cực vào mục tiêu chuyển dịch cơ cấu kinh tế tại các địa phương.

- Mở rộng thị trường và tăng trưởng xuất khẩu

Viglacera có nhiệm vụ phát triển mạng lưới xuất khẩu, tìm kiếm thị trường mới, đồng thời gia tăng giá trị thương hiệu quốc tế của sản phẩm Việt. Công ty chủ động tham gia các hội chợ quốc tế, thiết lập chi nhánh và hệ thống phân phối tại nước ngoài, từ đó nâng cao tỷ trọng xuất khẩu trong cơ cấu doanh thu.

- Góp phần xây dựng đô thị hiện đại, thân thiện với môi trường

Bên cạnh mục tiêu kinh doanh, Viglacera còn chú trọng đến trách nhiệm xã hội và phát triển đô thị bền vững. Thông qua việc đầu tư xây dựng các khu đô thị mới, nhà ở xã hội, công ty không chỉ giải quyết nhu cầu chỗ ở cho người dân mà còn tạo lập môi trường sống chất lượng cao, văn minh và tiết kiệm năng lượng.

- Đảm bảo hiệu quả hoạt động tài chính và cổ phần hóa

Là một công ty niêm yết trên sàn chứng khoán HOSE với mã cổ phiếu VGC, Viglacera có trách nhiệm duy trì tính minh bạch, hiệu quả trong quản trị tài chính, đem lại lợi ích bền vững cho cổ đông và nhà đầu tư. Đồng thời, công ty thực hiện nghiêm túc các quy định pháp lý về tài chính, kiểm toán và công bố thông tin.

2.2. Phân tích cổ phiếu VGC theo báo cáo tài chính

2.1.1. Tình hình tài chính 3 năm gần đây của VGC

❖ Tình hình chung

Tình hình tài chính chung của Tổng công ty Viglacera – CTCP trong những năm gần đây được đánh giá là ổn định và có chiều hướng tăng trưởng tích cực. Nhờ sự phối hợp hiệu quả giữa hai lĩnh vực chủ lực là sản xuất vật liệu xây dựng và đầu tư bất động sản khu công nghiệp – đô thị, Viglacera đã duy trì được cơ cấu doanh thu cân đối, đồng thời cải thiện biên lợi nhuận qua từng năm. Tổng tài sản của công ty không ngừng gia tăng, phản ánh chiến lược đầu tư mở rộng quy mô sản xuất và phát triển quỹ đất khu công nghiệp một cách bài bản. Các chỉ số tài chính quan trọng như doanh thu thuần, lợi nhuận sau thuế, tỷ suất sinh lời trên vốn chủ sở hữu (ROE) và khả năng thanh toán đều ở mức an toàn và hợp lý, cho thấy khả năng kiểm soát dòng tiền tốt và năng lực tài chính vững mạnh. Bên cạnh đó, việc niêm yết cổ phiếu VGC trên sàn HOSE giúp công ty tiếp cận nguồn vốn đầu tư dài hạn và nâng cao tính minh bạch trong quản trị tài chính. Trong bối cảnh kinh tế có nhiều biến động, Viglacera vẫn duy trì được đà tăng trưởng ổn định, thể hiện qua kết quả kinh doanh tích cực, hiệu quả sử dụng vốn cao và sự tin tưởng của các nhà đầu tư trong và ngoài nước. Đây là nền tảng quan trọng giúp doanh nghiệp tiếp

tục triển khai các chiến lược phát triển bền vững trong thời gian tới.

❖ Quy mô cơ cấu tài sản của công ty

Bảng 2.2. Tình hình tài sản của VGC từ 2022-2024

Đơn vị: Đồng

Chỉ tiêu	Năm 2022	Năm 2023	Năm 2024
	Số tiền	Số tiền	Số tiền
I. Tài sản ngắn hạn	8.107.975.056.610	9.104.809.897.620	9.464.267.034.186
II. Tài sản dài hạn	14.850.946.352.686	14.995.380.193.666	9.878.612.305.906
TỔNG TÀI SẢN	22.958.921.409.296	24.100.190.091.286	19.342.879.340.092

(Nguồn: Bảng cân đối kế toán tổng công ty VGC)

Trong giai đoạn 2022 đến 2024, tổng tài sản của Tổng công ty Viglacera – CTCP có sự biến động đáng kể. Cụ thể, tổng tài sản đạt 22.958,9 tỷ đồng vào năm 2022, tăng lên 24.100,2 tỷ đồng vào năm 2023, trước khi giảm còn 19.342,9 tỷ đồng vào năm 2024. Điều này cho thấy doanh nghiệp đã có sự mở rộng quy mô tài sản vào năm 2023, tuy nhiên sau đó đã thu hẹp hoặc tái cơ cấu trong năm 2024.

Xét về cơ cấu tài sản:

Tài sản ngắn hạn có xu hướng gia tăng rõ rệt cả về giá trị tuyệt đối và tỷ trọng. Từ mức 8.108 tỷ đồng (chiếm 35,3%) năm 2022, tăng lên 9.105 tỷ đồng (37,8%) năm 2023, và đạt 9.464 tỷ đồng (48,9%) vào năm 2024. Việc gia tăng tỷ trọng tài sản ngắn hạn cho thấy doanh nghiệp đang ưu tiên tính thanh khoản, có thể là để phục vụ nhu cầu vốn lưu động hoặc để ứng phó với biến động thị trường.

Tài sản dài hạn tuy vẫn chiếm tỷ trọng lớn trong tổng tài sản nhưng có xu hướng giảm nhẹ về tỷ trọng: từ 64,7% năm 2022 xuống còn 62,2% năm 2023 và 51,1% năm 2024. Giá trị tuyệt đối của tài sản dài hạn gần như đi ngang trong hai năm đầu (14.851 tỷ đồng năm 2022 và 14.995 tỷ đồng năm 2023) nhưng giảm đáng kể còn 9.879 tỷ đồng vào năm 2024. Điều này có thể phản ánh sự chấm dứt đầu tư dài hạn hoặc doanh nghiệp đã tiến hành thanh lý, thu hồi một phần tài sản cố định nhằm cải thiện dòng tiền.

Tổng thể, sự thay đổi trong cơ cấu tài sản cho thấy một xu hướng dịch chuyển từ tài sản dài hạn sang tài sản ngắn hạn trong năm 2024, phản ánh chiến lược quản trị tài chính mang tính thận trọng hơn của doanh nghiệp trong bối cảnh có thể có nhiều biến động kinh tế hoặc định hướng tái cấu trúc lại danh mục đầu tư.

❖ Cơ cấu nguồn vốn của công ty

Bảng 2.3. Tình hình nguồn vốn VGC 2022 – 2024

Đơn vị: Đồng

Chỉ tiêu	Năm 2022	Năm 2023	Năm 2024
	Số tiền	Số tiền	Số tiền
I. Nợ phải trả	13.873.492.333.128	14.575.872.174.590	14.874.419.272.735
II. Vốn chủ sở hữu	9.085.429.076.168	9.524.317.916.696	9.952.999.655.403
TỔNG NGUỒN VỐN	22.958.921.409.296	24.100.190.091.286	24.827.418.928.138

(Nguồn: Bảng cân đối kế toán tổng công ty VGC)

Trong giai đoạn 2022–2024, tổng nguồn vốn của Tổng công ty Viglacera – CTCP liên tục gia tăng, từ mức 22.958,9 tỷ đồng năm 2022 lên 24.100,2 tỷ đồng năm 2023 và đạt 24.827,4 tỷ đồng vào năm 2024. Sự gia tăng này phản ánh xu hướng mở rộng hoạt động sản xuất – kinh doanh cũng như đầu tư của doanh nghiệp trong thời kỳ này.

Xét về cơ cấu nguồn vốn, có thể nhận thấy:

Nợ phải trả chiếm tỷ trọng lớn và có xu hướng tăng nhẹ qua các năm, từ 13.873,5 tỷ đồng năm 2022 lên 14.575,9 tỷ đồng năm 2023 và đạt 14.874,4 tỷ đồng vào năm 2024. Điều này cho thấy Viglacera vẫn đang phụ thuộc khá nhiều vào vốn vay hoặc các khoản nợ trong hoạt động tài chính. Tuy nhiên, mức tăng này tương đối hợp lý và đi kèm với sự tăng trưởng về tổng tài sản, phản ánh khả năng kiểm soát rủi ro tài chính tương đối ổn định.

Vốn chủ sở hữu cũng tăng đều qua các năm, từ 9.085,4 tỷ đồng năm 2022 lên 9.524,3 tỷ đồng năm 2023 và đạt 9.953,0 tỷ đồng vào năm 2024. Mặc dù tốc độ tăng trưởng của vốn chủ sở hữu chậm hơn so với nợ phải trả, nhưng sự gia tăng liên tục cho thấy doanh nghiệp vẫn duy trì được mức sinh lời tích cực, đồng thời đảm bảo được nguồn vốn tự có để phục vụ hoạt động sản xuất kinh doanh.

Như vậy, cơ cấu nguồn vốn của Tổng công ty Viglacera – CTCP giai đoạn 2022–2024 cho thấy doanh nghiệp đang giữ được sự cân bằng tương đối giữa nợ phải trả và vốn chủ sở hữu. Dù nợ chiếm tỷ trọng lớn, nhưng với xu hướng tăng trưởng vốn chủ sở hữu qua các năm, Viglacera đang dần nâng cao mức độ tự chủ tài chính – một yếu tố quan trọng giúp củng cố vị thế tài chính bền vững trong dài hạn.

❖ Kết quả hoạt động kinh doanh

Bảng 2.4. Kết quả hoạt động kinh doanh VGC 2022 – 2024

Đơn vị: Đồng

	Năm 2022	Năm 2023	Năm 2024
Tổng Doanh Thu	14.693.558.951.326	13.402.996.947.613	12.127.286.572.307
Tổng Chi Phí	12.526.869.929.288	11.623.972.485.615	10.305.744.862.643
Tổng Lợi Nhuận	2.305.204.152.097	1.601.938.537.417	1.630.325.650.110

(Nguồn: Kết quả hoạt động kinh doanh tổng công ty VGC)

Tổng doanh thu thuần về bán hàng và cung cấp dịch vụ liên tục giảm qua các năm, từ hơn 14.592 tỷ đồng năm 2022 xuống còn 13.193 tỷ đồng năm 2023 và tiếp tục giảm còn 11.906 tỷ đồng vào năm 2024. Sự sụt giảm này cho thấy công ty đang đối mặt với thách thức trong việc mở rộng thị trường hoặc duy trì sức mua của khách hàng. Đây là tín hiệu cần được phân tích sâu để xác định nguyên nhân như áp lực cạnh tranh, thị trường thu hẹp, hay chính sách giá chưa hiệu quả.

Chi phí vận hành (bao gồm giá vốn hàng bán, chi phí tài chính, chi phí bán hàng và quản lý doanh nghiệp) cũng giảm theo cùng xu hướng với doanh thu. Tuy nhiên, mức độ giảm của chi phí chậm hơn so với tốc độ giảm doanh thu. Cụ thể, tổng chi phí hoạt động năm 2022 là 12.526 tỷ đồng, giảm còn 11.591 tỷ đồng vào năm 2023 và 11.137 tỷ đồng vào năm 2024. Điều này cho thấy công ty có nỗ lực cắt giảm chi phí để thích nghi với mức doanh thu giảm, tuy nhiên hiệu quả tiết giảm chưa đủ mạnh để tạo ra bước đột phá về lợi nhuận.

Lợi nhuận kế toán trước thuế cũng cho thấy xu hướng giảm dần: từ 2.305 tỷ đồng năm 2022, giảm mạnh còn 1.601 tỷ đồng năm 2023, và chỉ đạt 1.630 tỷ đồng năm 2024. Mặc dù có sự cải thiện nhẹ trong năm 2024 so với năm 2023, mức tăng không đáng kể và không bù đắp được sự sụt giảm lớn từ năm 2022.

Như vậy, có thể thấy rằng công ty đang rơi vào trạng thái "doanh thu giảm, chi phí giảm chậm, lợi nhuận giảm", phản ánh những khó khăn rõ rệt trong hoạt động sản xuất – kinh doanh. Trong bối cảnh thị trường nhiều biến động, việc duy trì tốc độ tăng trưởng lợi nhuận ổn định là một thách thức lớn đối với doanh nghiệp.

2.2.2. Đánh giá các chỉ số tài chính doanh nghiệp 3 năm gần đây nhất

Chỉ số tài chính doanh nghiệp là một trong những căn cứ quan trọng hàng đầu khi nhà đầu tư cân nhắc rót vốn vào cổ phiếu của một công ty. Những chỉ số này cung cấp góc nhìn toàn diện về khả năng tạo lợi nhuận, hiệu quả sử dụng tài sản và sức mạnh nội tại của doanh nghiệp trong bối cảnh cạnh tranh thị trường ngày càng khốc liệt.

Bảng 2.5. Chỉ số tài chính VGC 3 năm gần đây

Chỉ số tài chính	Năm 2022	Năm 2023	Năm 2024
EPS 4 quý	3,855.00	2,717.00	2,464.00
BVPS cơ bản	20,173.00	21,159.00	22,122.00
P/E cơ bản	8.77	20.21	18.24
ROS	13.11	8.81	9.97
ROEA	19.82	13.09	11.34
ROAA	7.69	5.18	4.52

(nguồn: cafef.vn)

❖ EPS (Earnings per Share – Lợi nhuận trên mỗi cổ phiếu)

EPS 4 quý liên tục giảm dần từ 3.855 đồng (năm 2022) xuống còn 2.717 đồng (2023) và 2.464 đồng (2024). Xu hướng này phản ánh sự sụt giảm lợi nhuận ròng dành cho cổ đông phổ thông, cho thấy khả năng sinh lời trực tiếp trên mỗi cổ phiếu đang suy yếu. Sự suy giảm này có thể xuất phát từ các yếu tố như: chi phí sản xuất gia tăng, thị trường bất động sản giảm tốc hoặc sự biến động giá nguyên vật liệu xây dựng – ngành nghề chính của Viglacera. Đây là tín hiệu cảnh báo cho các nhà đầu tư dài hạn, đặc biệt trong bối cảnh chi phí cơ hội và yêu cầu lợi suất ngày càng cao.

❖ BVPS (Book Value per Share – Giá trị sổ sách trên mỗi cổ phiếu)

Trái ngược với EPS, BVPS cơ bản lại ghi nhận xu hướng tăng dần qua các năm: từ 20.173 đồng (2022) lên 22.122 đồng (2024). Điều này phản ánh sự tăng trưởng giá trị tài sản thuần của doanh nghiệp, phần nào cho thấy công ty vẫn duy trì được vốn chủ sở hữu tích lũy ổn định dù lợi nhuận sụt giảm. Tuy nhiên, tốc độ tăng trưởng BVPS không đủ bù đắp cho sự suy giảm EPS, điều này có thể dẫn đến áp lực giảm giá thị trường cổ phiếu nếu nhà đầu tư tập trung vào hiệu quả sinh lời hơn là giá trị sổ sách.

❖ Chỉ số P/E (Price to Earnings – Hệ số giá trên thu nhập)

Hệ số P/E cơ bản tăng mạnh từ 8.77 (2022) lên 20.21 (2023) rồi giảm nhẹ còn 18.24 (2024). Mức P/E cao cho thấy kỳ vọng của thị trường đối với tăng trưởng lợi nhuận trong tương lai, nhưng trong trường hợp này, sự tăng vọt P/E chủ yếu do sự sụt giảm của EPS – không phải từ kỳ vọng tích cực mà là hiệu ứng toán học. Nếu EPS tiếp tục suy giảm trong khi giá cổ phiếu không giảm tương ứng, P/E cao sẽ khiến cổ phiếu trở nên đắt đỏ tương đối và làm giảm sức hấp dẫn đầu tư. Đặc biệt trong ngành vật liệu xây dựng có tính chu kỳ cao, một P/E cao nhưng không bền vững có thể dẫn đến rủi ro định giá lại (revaluation risk).

❖ ROS (Return on Sales – Tỷ suất lợi nhuận trên doanh thu)

ROS giảm từ 13.11% (2022) xuống còn 9.97% (2024) sau khi chạm đáy 8.81% vào năm 2023. Tỷ suất này phản ánh biên lợi nhuận từ hoạt động kinh doanh cốt lõi và cho thấy hiệu quả kiểm soát chi phí so với doanh thu. Việc ROS giảm cho thấy Viglacera đang đối mặt với áp lực từ chi phí sản xuất, chiết khấu bán hàng hoặc nhu cầu thị trường suy yếu. Dù có dấu hiệu phục hồi nhẹ năm 2024, xu hướng chung là biên lợi nhuận đang bị co hẹp – đây là yếu tố quan trọng cần theo dõi nếu muốn đánh giá chất lượng tăng trưởng doanh nghiệp trong dài hạn.

❖ ROEA (Return on Equity – Tỷ suất sinh lời trên vốn chủ sở hữu)

ROEA giảm mạnh từ 19.82% (2022) xuống 13.09% (2023) và tiếp tục xuống còn 11.34% (2024). Đây là tín hiệu rõ ràng về việc khả năng sinh lời từ nguồn vốn của cổ đông đang bị suy yếu. ROEA phản ánh hiệu quả sử dụng vốn chủ sở hữu và là một trong những chỉ số quan trọng nhất để đánh giá mức độ hấp dẫn đầu tư. Với mức ROEA dưới 12% như hiện tại, Viglacera khó có thể duy trì sự hấp dẫn với các nhà đầu tư lớn vốn đòi hỏi mức sinh lời cao hơn lãi suất chiết khấu.

❖ ROAA (Return on Assets – Tỷ suất sinh lời trên tổng tài sản)

Tương tự ROEA, ROAA cũng cho thấy xu hướng giảm mạnh từ 7.69% (2022) xuống 4.52% (2024). Điều này phản ánh việc sử dụng tổng tài sản của doanh nghiệp đang kém hiệu quả dần theo thời gian, có thể do tăng tài sản cố định nhưng không sinh lợi tương ứng hoặc hiệu suất sử dụng vốn lưu động thấp. Trong ngành xây dựng và bất động sản – nơi yêu cầu đầu tư tài sản cao – việc duy trì ROAA thấp trong nhiều năm liền là dấu hiệu cho thấy cần có cải tổ trong chiến lược tài chính hoặc mô hình vận hành.

❖ Nhận định tổng quát

Tổng quan các chỉ số cho thấy Viglacera đang ở trong giai đoạn chững lại về hiệu quả hoạt động và lợi nhuận. Việc EPS, ROEA, ROAA và ROS đồng loạt suy giảm trong khi BVPS tăng chậm và P/E biến động mạnh cho thấy doanh nghiệp đang có dấu hiệu giảm sút nội lực sinh lời, mặc dù vẫn giữ được nền tảng tài sản tương đối ổn định. Điều này có thể ảnh hưởng đến kỳ vọng của nhà đầu tư trong trung hạn và làm giảm khả năng tăng trưởng giá cổ phiếu nếu không có cải thiện rõ rệt trong hiệu quả sử dụng vốn và chi phí hoạt động.

2.2.3. Tình hình cổ phiếu doanh nghiệp 3 năm gần đây

❖ Tình hình chung

Bảng 2.6. Bảng kết quả phân tích thống kê mô tả

Chỉ số	Giá trị
Giá cao nhất (nghìn đồng)	65,70
Giá thấp nhất (nghìn đồng)	25,40
Giá trung bình (nghìn đồng)	45,14
Biến động trung bình ngày (%)	2,14
Khối lượng trung bình (cổ phiếu)	1.105.206

(Nguồn: Tác giả tổng hợp và tính toán)

Trước hết, mức giá giao dịch dao động trong biên độ rộng, với giá cao nhất đạt 65,70 nghìn đồng/cổ phiếu, trong khi giá thấp nhất chỉ ở mức 25,40 nghìn đồng/cổ phiếu. Điều này phản ánh mức độ biến động đáng kể trong xu hướng thị trường, đồng thời cho thấy cổ phiếu đã trải qua nhiều pha tăng – giảm rõ rệt trong các giai đoạn khác nhau. Sự chênh lệch lớn giữa giá cực đại và cực tiểu (biên độ lên đến 40.300 đồng) là một yếu tố quan trọng cần được lưu ý khi áp dụng các mô hình dự báo, đặc biệt là những mô hình nhạy cảm với outliers.

Tiếp theo, giá trung bình của cổ phiếu đạt khoảng 45,14 nghìn đồng, được xem là tương đối cao so với mặt bằng chung các cổ phiếu trên thị trường niêm yết hiện nay. Đây có thể là biểu hiện của mức định giá tương đối tốt, đồng thời phản ánh kỳ vọng tích cực từ thị trường đối với hiệu quả kinh doanh và tiềm năng tăng trưởng của doanh nghiệp trong dài hạn.

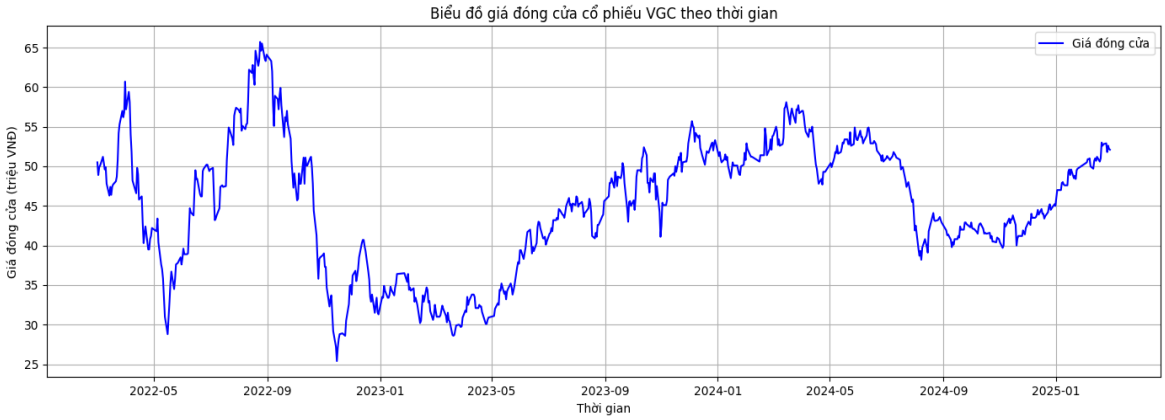
Xét về độ biến động, chỉ số biến động trung bình theo ngày đạt mức 2,14%, cho thấy cổ phiếu có mức độ dao động tương đối đều đặn giữa các phiên. Với mức biến động này, cổ phiếu VGC không thuộc nhóm quá rủi ro, nhưng cũng không hoàn toàn ổn định, phù hợp với mục tiêu giao dịch trung hạn.

Ngoài ra, khối lượng giao dịch trung bình đạt hơn 1,1 triệu cổ phiếu mỗi phiên, thể hiện tính thanh khoản khá cao. Đây là yếu tố thuận lợi cho cả nhà đầu tư và mô hình dự báo, vì dữ liệu không bị gián đoạn, không có hiện tượng "trống" do thiếu giao dịch – giúp mô hình học được các quy luật giá một cách liên tục và đáng tin cậy hơn.

❖ Xu hướng

Trong 3năm gần đây, cổ phiếu của Tổng công ty Viglacera đã ghi nhận nhiều biến động phản ánh sự tác động của cả yếu tố nội tại doanh nghiệp lẫn bối cảnh vĩ mô.

Hình 2.2. Biểu đồ giá đóng cửa cổ phiếu VGC theo thời gian



(Nguồn: Tác giả tổng hợp và tính toán)

Biểu đồ trên thể hiện diễn biến giá đóng cửa cổ phiếu VGC trong khoảng thời gian từ đầu năm 2022 đến đầu năm 2025. Quan sát tổng thể cho thấy giá cổ phiếu VGC có xu hướng biến động khá mạnh, với nhiều chu kỳ tăng – giảm rõ rệt, phản ánh sự nhạy cảm của cổ phiếu này trước các yếu tố vĩ mô và nội tại doanh nghiệp.

Trong giai đoạn từ quý I đến quý III năm 2022, giá cổ phiếu có xu hướng tăng mạnh và đạt đỉnh trên mức 65.000 đồng/cổ phiếu vào khoảng tháng 8/2022. Đây có thể là giai đoạn doanh nghiệp công bố kết quả kinh doanh tích cực hoặc hưởng lợi từ xu thế đầu tư vào bất động sản khu công nghiệp sau đại dịch COVID-19. Tuy nhiên, sau khi đạt đỉnh, cổ phiếu nhanh chóng bước vào chu kỳ điều chỉnh sâu, rơi xuống dưới mức 30.000 đồng/cổ phiếu vào đầu năm 2023, cho thấy sự sụt giảm niềm tin của nhà đầu tư và ảnh hưởng tiêu cực từ bối cảnh vĩ mô như siết tín dụng bất động sản, lạm phát và lãi suất tăng cao.

Từ giữa năm 2023 trở đi, giá cổ phiếu dần phục hồi và hình thành xu hướng tăng trung hạn khá ổn định, duy trì trong khoảng 35.000–55.000 đồng/cổ phiếu. Biểu đồ cũng cho thấy nhiều nhịp điều chỉnh nhỏ xen kẽ trong quá trình đi lên, thể hiện tâm lý thị trường vẫn còn thận trọng. Đặc biệt, trong quý III/2024, giá cổ phiếu tiếp tục chịu áp lực giảm sâu nhưng sau đó đã nhanh chóng phục hồi trở lại trong quý IV/2024 và duy trì xu hướng tăng vào đầu năm 2025.

- Biểu đồ nến (candlestick chart)

Trong phân tích kỹ thuật thị trường tài chính, đặc biệt là thị trường chứng khoán, biểu đồ nến Nhật (candlestick chart) là một trong những công cụ trực quan và phổ biến nhất để thể hiện biến động giá cổ phiếu theo thời gian.

Hình 2.3. Biểu đồ nến cổ phiếu VGC

(Nguồn: Tác giả tổng hợp và tính toánh)

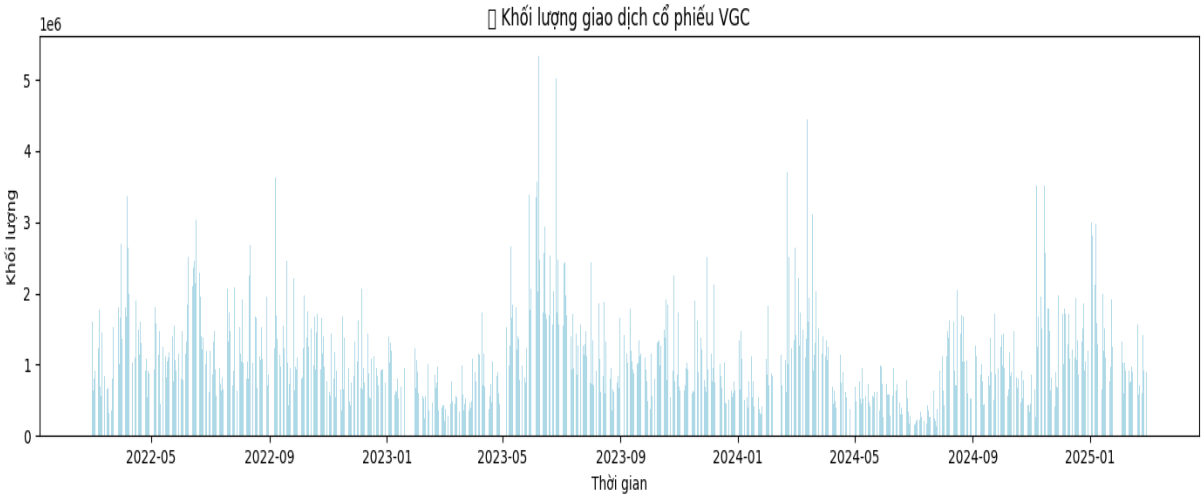
Biểu đồ trên thể hiện diễn biến giá cổ phiếu VGC thông qua mô hình nến Nhật (candlestick chart) trong giai đoạn từ đầu năm 2022 đến đầu năm 2025. Mỗi cây nến thể hiện biến động giá trong một phiên giao dịch, bao gồm các mức giá mở cửa, đóng cửa, cao nhất và thấp nhất. Màu xanh (hoặc xanh lá) thể hiện phiên tăng giá (giá đóng cửa cao hơn giá mở cửa), trong khi màu đỏ thể hiện phiên giảm giá (giá đóng cửa thấp hơn giá mở cửa). Biểu đồ nến cho phép quan sát trực quan và chi tiết hơn về hành vi thị trường, đặc biệt là các điểm đảo chiều, lực mua bán và độ mạnh của xu hướng trong từng giai đoạn.

Quan sát biểu đồ cho thấy trong nửa đầu năm 2022, cổ phiếu VGC trải qua nhiều phiên biến động mạnh, thể hiện qua các cây nến dài, bóng nến lớn và sự thay đổi nhanh chóng giữa phiên tăng và phiên giảm. Sau giai đoạn này, giá cổ phiếu tăng mạnh và đạt đỉnh vào khoảng giữa năm 2022 với mức giá vượt ngưỡng 65.000 đồng/cổ phiếu. Tuy nhiên, ngay sau đó, cổ phiếu rơi vào giai đoạn điều chỉnh mạnh và kéo dài đến đầu năm 2023, phản ánh tâm lý chốt lời và ảnh hưởng tiêu cực từ các yếu tố vĩ mô.

Từ giữa năm 2023, biểu đồ cho thấy một xu hướng tăng tương đối rõ ràng, được củng cố bởi chuỗi nến xanh liên tiếp và ít bóng nến dưới – một dấu hiệu của lực mua ổn định. Đặc biệt, trong năm 2024, cổ phiếu trải qua giai đoạn tích lũy với biên độ dao động hẹp, xen kẽ các nhịp điều chỉnh ngắn hạn. Đến cuối năm 2024 và đầu năm 2025, biểu đồ xuất hiện nhiều nến tăng mạnh liên tiếp, thể hiện đà hồi phục rõ rệt và kỳ vọng tích cực từ thị trường đối với triển vọng kinh doanh của doanh nghiệp.

❖ Khối lượng giao dịch cổ phiếu doanh nghiệp

Hình 2.4. Khối lượng giao dịch cổ phiếu VGC theo thời gian



(Nguồn: Tác giả tổng hợp và tính toán)

Biểu đồ trên thể hiện khối lượng giao dịch cổ phiếu VGC theo từng phiên giao dịch từ đầu năm 2022 đến đầu năm 2025. Mỗi cột đứng biểu thị tổng số cổ phiếu được trao tay trong một ngày giao dịch. Đây là một trong những chỉ báo quan trọng phản ánh mức độ quan tâm của nhà đầu tư đối với cổ phiếu, cũng như tính thanh khoản và dòng tiền thị trường.

Quan sát tổng thể, khối lượng giao dịch của cổ phiếu VGC biến động khá mạnh theo từng giai đoạn, thường xuất hiện các đỉnh giao dịch tại những thời điểm có sự biến động lớn về giá hoặc xuất hiện các thông tin tác động mạnh đến tâm lý thị trường.

Giai đoạn giữa năm 2022 và quý II/2023 ghi nhận nhiều phiên có khối lượng giao dịch cao, nhiều thời điểm vượt mốc 2–3 triệu cổ phiếu/phiên, thậm chí chạm đỉnh trên 5 triệu cổ phiếu. Điều này cho thấy dòng tiền đầu cơ đổ vào mạnh, có thể liên quan đến hoạt động mua bán quanh các vùng đỉnh giá hoặc thời điểm doanh nghiệp công bố kết quả kinh doanh, chia cổ tức, hoặc những thông tin vĩ mô ảnh hưởng đến toàn ngành vật liệu xây dựng – bất động sản.

Từ giữa năm 2023 đến đầu năm 2024, khối lượng giao dịch có dấu hiệu giảm sút đáng kể, phản ánh sự thận trọng của nhà đầu tư trong bối cảnh thị trường chung thiếu động lực, cùng với xu hướng đi ngang hoặc điều chỉnh của giá cổ phiếu.

Giai đoạn quý IV/2024 đến đầu 2025 chứng kiến sự phục hồi dần về thanh khoản, thể hiện qua sự gia tăng trở lại của các cột khối lượng giao dịch. Đây là dấu hiệu tích cực cho thấy thị trường đã bắt đầu có sự chú ý trở lại với mã cổ phiếu này, đồng thời phản ánh kỳ vọng về triển vọng phục hồi của doanh nghiệp.

2.2.4. Các yếu tố ảnh hưởng đến giá cổ phiếu doanh nghiệp

Trước tiên, yếu tố ảnh hưởng trực tiếp và mạnh mẽ nhất đến giá cổ phiếu VGC chính là kết quả hoạt động kinh doanh. Các chỉ số tài chính như doanh thu thuần, lợi nhuận sau thuế, biên lợi nhuận gộp và thu nhập trên mỗi cổ phiếu (EPS) có ảnh hưởng lớn đến tâm lý nhà đầu tư. Giai đoạn 2022–2024, dù doanh nghiệp vẫn duy trì được lợi nhuận, song mức giảm liên tục trong doanh thu và lợi nhuận ròng đã tạo áp lực giảm giá lên cổ phiếu. Nhà đầu tư thường phản ứng nhạy bén trước sự sụt giảm trong hiệu suất kinh doanh, đặc biệt là ở các ngành có tính chu kỳ như xây dựng và bất động sản.

Thứ hai, tình hình chung của thị trường bất động sản và vật liệu xây dựng cũng là yếu tố ảnh hưởng rõ rệt. Khi thị trường bất động sản rơi vào trạng thái chững lại do các biện pháp kiểm soát tín dụng, siết dòng vốn đầu tư công và tăng lãi suất, nhu cầu tiêu thụ vật liệu xây dựng giảm kéo theo ảnh hưởng đến doanh thu cốt lõi của Viglacera. Đồng thời, việc chậm triển khai hoặc bán chậm các dự án khu công nghiệp cũng khiến nguồn thu từ mảng bất động sản bị ảnh hưởng, tác động đến kỳ vọng của nhà đầu tư trên sàn.

Một yếu tố không kém phần quan trọng là môi trường kinh tế vĩ mô, bao gồm chính sách tiền tệ, lãi suất ngân hàng, tỷ giá hối đoái và lạm phát. Khi lãi suất tăng, chi phí vốn vay của doanh nghiệp tăng theo, đồng thời nhà đầu tư có xu hướng rút vốn khỏi thị trường cổ phiếu để chuyển sang các kênh đầu tư an toàn hơn như trái phiếu hoặc gửi tiết kiệm, khiến cầu cổ phiếu giảm. Bên cạnh đó, biến động tỷ giá cũng ảnh hưởng đến chi phí nhập khẩu nguyên vật liệu, đặc biệt là đối với các sản phẩm sử dụng công nghệ cao, thiết bị ngoại nhập như kính tiết kiệm năng lượng – một mặt hàng chủ lực của Viglacera.

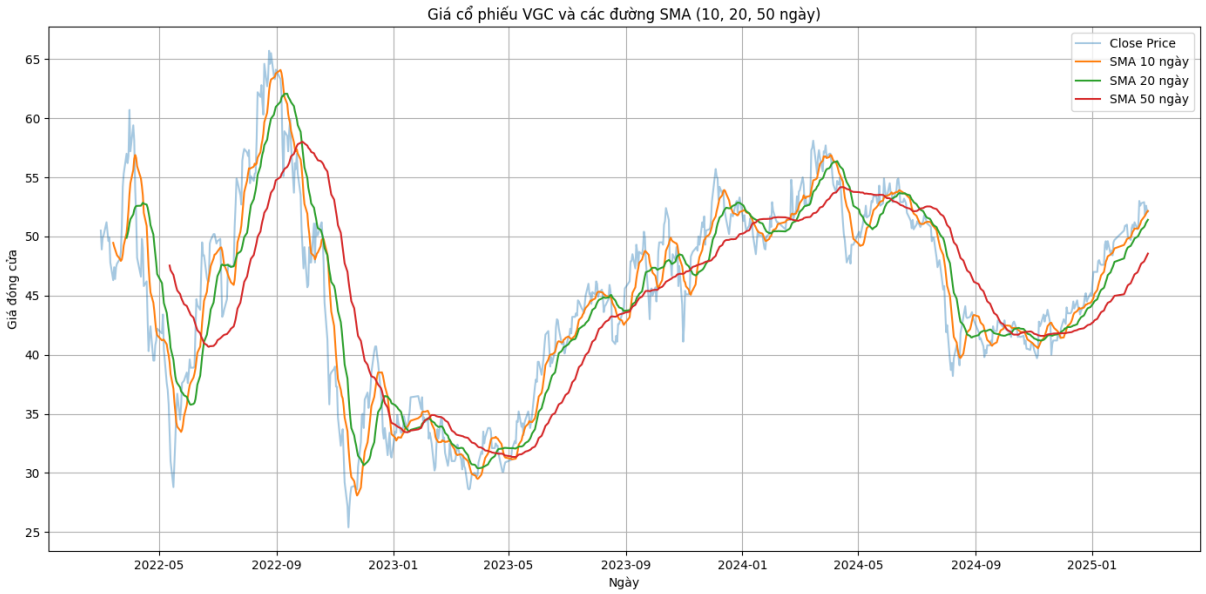
Ngoài ra, các yếu tố chiến lược nội tại như chính sách cổ tức, hoạt động M&A, mở rộng thị trường xuất khẩu hay đầu tư vào sản phẩm mới cũng có tác động không nhỏ. Ví dụ, việc công bố cổ tức đều đặn hoặc có kế hoạch phát triển các khu công nghiệp mới thường giúp gia tăng niềm tin của cổ đông và nâng cao giá trị cổ phiếu trên thị trường. Ngược lại, nếu doanh nghiệp chưa công bố rõ ràng kế hoạch đầu tư hoặc mở rộng thị trường, cổ phiếu có thể rơi vào trạng thái thiếu động lực tăng trưởng.

Cuối cùng, tâm lý nhà đầu tư và xu hướng dòng tiền trên thị trường chứng khoán cũng góp phần tạo ra biến động ngắn hạn cho cổ phiếu VGC. Những thời điểm thị trường "bullish" (tăng mạnh), nhà đầu tư có xu hướng đẩy giá cổ phiếu vượt xa giá trị nội tại, và ngược lại, trong giai đoạn "bearish" (giảm điểm).

2.3. Phân tích cổ phiếu VGC theo các chỉ báo kĩ thuật

2.3.1. Đường trung bình trượt SMA

Hình 2.5. Biểu đố giá cổ phiếu VGC và các đường SMA (10, 20, 50)



(nguồn: Tác gỉa tổng hợp và tính toán)

Biểu đồ trên thể hiện rõ xu hướng biến động giá cổ phiếu VGC trong giai đoạn từ tháng 3/2022 đến đầu năm 2025, kết hợp cùng ba đường trung bình động đơn giản (SMA) với các chu kỳ 10, 20 và 50 ngày – tương ứng với các mốc thời gian ngắn hạn, trung hạn và dài hạn.

Trước hết, có thể quan sát rằng SMA 10 ngày (đường cam) phản ứng nhanh nhất với các dao động giá, thể hiện rõ các pha tăng – giảm ngắn hạn. Đường này thường xuyên giao cắt với giá thực tế, tạo ra nhiều tín hiệu đảo chiều, tuy nhiên cũng dễ bị nhiễu trong các giai đoạn biến động mạnh.

Trong khi đó, SMA 20 ngày (đường xanh lá) đóng vai trò như một bộ lọc nhiễu hiệu quả hơn, cung cấp góc nhìn ổn định hơn về xu hướng trung hạn. Đường này thường bám sát xu hướng giá nhưng ít bị biến động ngắn hạn làm lệch hướng, nhờ đó trở thành tham chiếu hữu ích trong các chiến lược giữ lệnh trung hạn.

Đáng chú ý nhất là SMA 50 ngày (đường đỏ) – đường trung bình dài hạn, cho thấy sự ổn định rõ rệt và phản ánh xu hướng chung của thị trường theo chu kỳ rộng hơn. Ở những giai đoạn thị trường tích lũy hoặc điều chỉnh mạnh, SMA 50 có độ trễ cao nhưng lại giúp loại bỏ hoàn toàn các tín hiệu nhiễu. Cụ thể, giai đoạn từ quý I/2023 đến quý III/2023 chứng kiến một đợt tăng giá mạnh mẽ khi cả ba đường SMA đồng loạt hướng

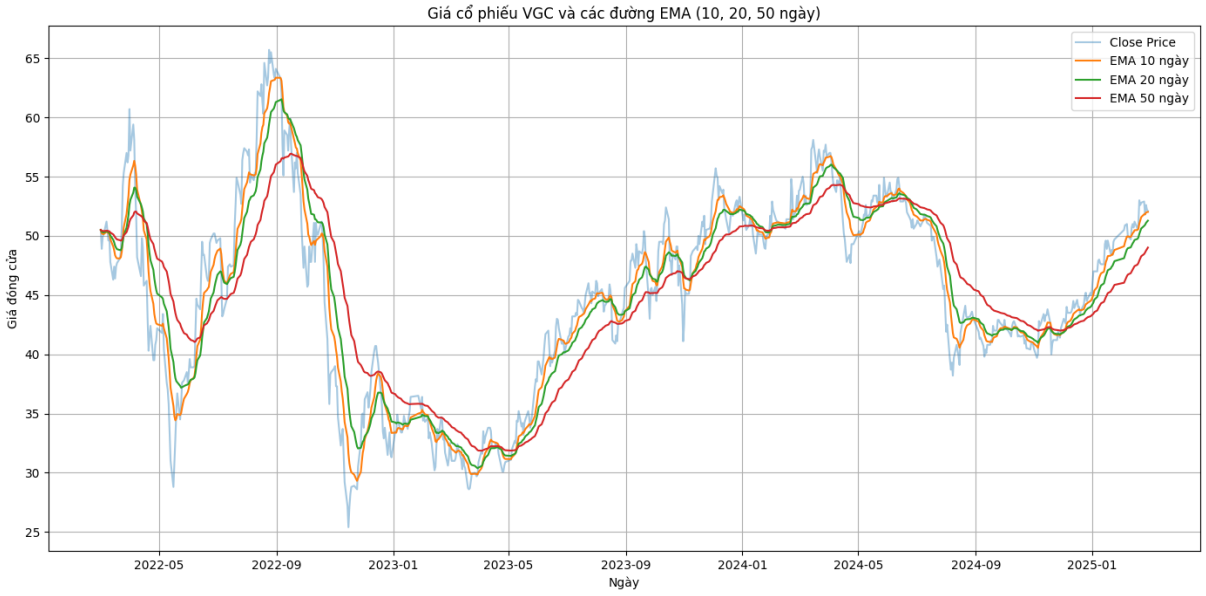
lên và sắp xếp theo đúng thứ tự tăng dần (SMA10 > SMA20 > SMA50) – dấu hiệu đặc trưng của một uptrend bền vững.

Ngược lại, trong các pha điều chỉnh như giữa năm 2022 hoặc giữa năm 2024, ta quan sát thấy hiện tượng giao cắt ngược (death cross), khi SMA ngắn hạn cắt xuống dưới SMA dài hạn, đồng thời khoảng cách giữa các đường thu hẹp lại – cho thấy tín hiệu suy yếu rõ rệt về mặt động lượng.

Tổng thể, biểu đồ SMA cho thấy cổ phiếu VGC trải qua các chu kỳ tăng – giảm rõ rệt, đồng thời các chỉ báo SMA phản ánh tương đối chính xác các giai đoạn chuyển tiếp xu hướng. Việc phân tích trực quan các đường SMA không chỉ giúp đánh giá mức độ ổn định của thị trường theo từng giai đoạn, mà còn cung cấp nền tảng lý luận định tính bổ sung cho các mô hình định lượng trong quá trình dự báo.

2.3.2. Đường trung bình động EMA

Hình 2.6. Giá cổ phiếu VGC và các đường EMA(10, 20, 50)



(nguồn: Tác gia tổng hợp và tính toán)

Biểu đồ thể hiện giá đóng cửa của cổ phiếu VGC trong giai đoạn từ tháng 3/2022 đến đầu năm 2025, kèm theo ba đường trung bình động hàm mũ (EMA) với các chu kỳ 10, 20 và 50 ngày. Đây là những chỉ báo kỹ thuật thường được sử dụng để theo dõi động lượng và xu hướng giá với độ nhạy cao hơn so với các đường trung bình động đơn giản (SMA).

Đầu tiên có thể nhận thấy, EMA 10 ngày (màu cam) phản ứng rất nhanh với các dao động giá trong ngắn hạn. Đường này thường cắt lên hoặc cắt xuống đường giá tại

các thời điểm thị trường biến động mạnh, do đó thường tạo ra nhiều tín hiệu sớm về đảo chiều xu hướng. Tuy nhiên, cũng như các chỉ báo có độ trễ thấp, EMA 10 dễ bị ảnh hưởng bởi nhiễu, đặc biệt trong các giai đoạn thị trường dao động không rõ ràng.

EMA 20 ngày (màu xanh lá) mang lại cái nhìn ổn định hơn so với EMA 10, nhưng vẫn giữ được độ nhạy cần thiết để bám sát xu hướng trung hạn. Đường này tỏ ra hiệu quả trong việc xác định các giai đoạn tích lũy hoặc tăng trưởng vừa phải. Ví dụ, trong nửa sau năm 2023, EMA 20 liên tục nằm dưới đường giá và đóng vai trò là đường hỗ trợ mạnh mẽ trong một xu thế tăng.

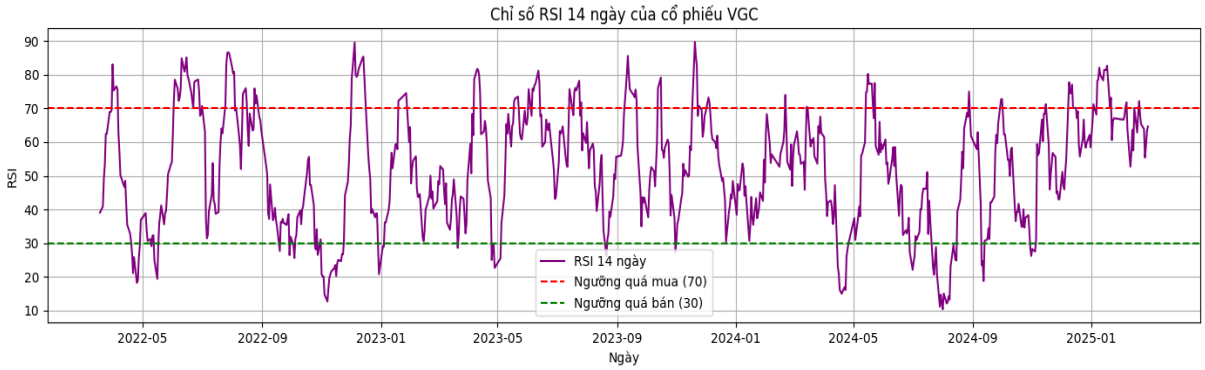
EMA 50 ngày (màu đỏ) phản ánh xu hướng dài hạn một cách ổn định và ít biến động hơn. Đặc biệt, trong giai đoạn từ quý I/2023 đến quý III/2023, khi đường EMA 10 và EMA 20 nằm trên EMA 50, thị trường bước vào một pha tăng trưởng rõ rệt – đây là tín hiệu cổ điển của một xu hướng tăng mạnh mẽ và bền vững. Ngược lại, trong giai đoạn từ giữa năm 2022 đến đầu năm 2023, sự giao cắt ngược giữa các đường EMA cùng với khoảng cách mở rộng giữa chúng phản ánh trạng thái thị trường tiêu cực kéo dài.

So với các đường SMA, các đường EMA có độ nhạy cao hơn với giá gần nhất, do đó thường báo hiệu đảo chiều sớm hơn, đặc biệt hữu ích khi thị trường đang bước vào giai đoạn chuyển động mới. Điều này làm cho EMA trở thành công cụ phù hợp trong các mô hình dự báo yêu cầu độ phản ứng nhanh, chẳng hạn như các chiến lược giao dịch theo xu hướng hoặc chiến lược định thời điểm thị trường.

Tổng kết lại, biểu đồ các đường EMA không chỉ cung cấp cái nhìn sâu sắc về động lượng thị trường theo từng giai đoạn, mà còn góp phần hỗ trợ trực quan và lý giải kết quả đầu ra từ các mô hình dự báo định lượng như Random Forest và LSTM.

2.3.3. Sức mạnh tương đối RSI

Hình 2.7. Chỉ số RSI 24 ngày của VGC



(nguồn: Tác giả tổng hợp và tính toán)

Để đánh giá động lượng giá cổ phiếu và nhận diện các vùng thị trường có nguy cơ đảo chiều, đề tài đã sử dụng chỉ báo kỹ thuật RSI (Relative Strength Index) với chu kỳ 14 ngày. Biểu đồ RSI được trình bày phản ánh biến động của chỉ số này từ tháng 3/2022 đến đầu năm 2025, cho thấy nhiều đặc điểm đáng chú ý trong hành vi thị trường của cổ phiếu VGC.

Trước hết, có thể nhận thấy rằng chỉ số RSI dao động mạnh và thường xuyên vượt qua các ngưỡng chuẩn là 70 và 30 – vốn được sử dụng phổ biến để xác định các trạng thái quá mua và quá bán. Cụ thể, trong nhiều thời điểm (đặc biệt là nửa cuối năm 2022 và giữa năm 2023), RSI đã vượt trên ngưỡng 70. Điều này cho thấy thị trường rơi vào trạng thái hưng phấn ngắn hạn, có khả năng kích hoạt các đợt điều chỉnh hoặc tích lũy giá sau đó.

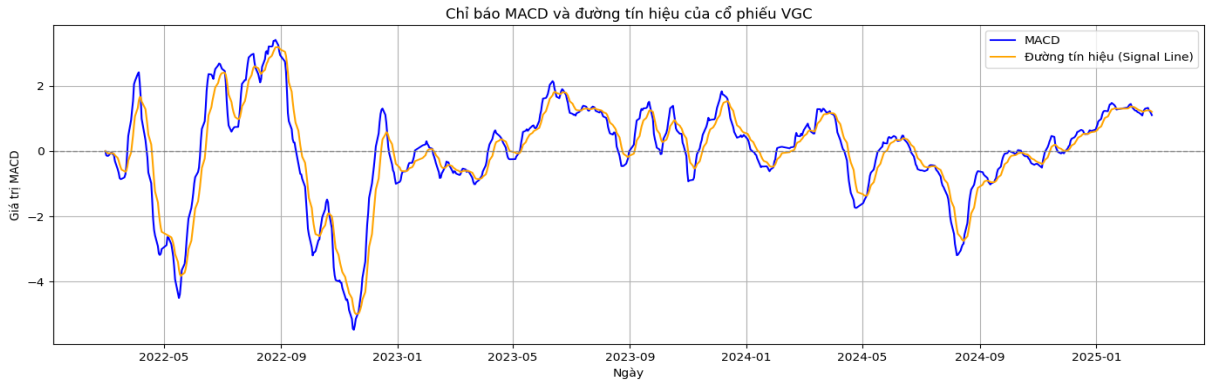
Ngược lại, trong một số giai đoạn thị trường giảm sâu, RSI đã tụt xuống dưới ngưỡng 30 – một tín hiệu cho thấy cổ phiếu có thể đang bị bán tháo quá mức. Những thời điểm này thường xuất hiện trước các pha hồi phục kỹ thuật, tuy nhiên biên độ phục hồi còn phụ thuộc vào yếu tố nền tảng thị trường và dòng tiền.

Tiếp theo, nếu quan sát kỹ giai đoạn từ giữa năm 2023 đến giữa năm 2024, có thể thấy chỉ số RSI dao động chủ yếu quanh vùng trung tính (từ 40 đến 60), phản ánh thị trường không có xu hướng rõ ràng. Đây là biểu hiện của trạng thái tích lũy hoặc giằng co, khi lực mua và lực bán tương đối cân bằng và chưa có sự bứt phá rõ rệt.

Đáng chú ý hơn, từ quý IV/2024 đến đầu năm 2025, RSI có xu hướng duy trì trên ngưỡng 50 và thường xuyên tiệm cận vùng 70. Điều này hàm ý rằng thị trường đã có sự cải thiện rõ nét về tâm lý và xu hướng tăng trưởng trung hạn có thể đang hình thành.

2.3.4. Trung bình động hội tụ phân kì

Hình 2.8. Chỉ báo MACD và đường tín hiệu cổ phiếu VGC



(nguồn: Tác giả tổng hợp và tính toán)

Trong quá trình khai phá dữ liệu, việc trực quan hóa các chỉ báo kỹ thuật đóng vai trò quan trọng trong việc cung cấp cái nhìn định tính ban đầu về hành vi thị trường. Trong số đó, chỉ báo MACD (Moving Average Convergence Divergence) là một công cụ hữu hiệu trong việc đánh giá động lượng và xu hướng giá của tài sản tài chính. Biểu đồ trên thể hiện sự biến động của đường MACD và đường tín hiệu (Signal Line) trong suốt giai đoạn từ đầu năm 2022 đến đầu năm 2025, áp dụng cho cổ phiếu VGC.

Trước hết, có thể thấy rằng các đường MACD và đường tín hiệu dao động quanh trục 0, phản ánh trạng thái cân bằng giữa các lực mua và bán trên thị trường. Các pha giao cắt giữa hai đường này thường được xem là tín hiệu cảnh báo về khả năng đảo chiều xu hướng. Cụ thể, khi đường MACD cắt lên trên đường tín hiệu, đó là dấu hiệu cho thấy lực mua đang gia tăng và thị trường có khả năng bước vào một xu hướng tăng. Ngược lại, khi MACD cắt xuống dưới đường tín hiệu, điều này thường cho thấy áp lực bán đang chiếm ưu thế và thị trường có thể rơi vào pha điều chỉnh hoặc giảm giá.

Quan sát biểu đồ trong giai đoạn năm 2022, dễ nhận thấy rằng MACD có những biến động khá mạnh và thường xuyên vượt qua các mức ±2. Điều này phản ánh một thị trường có độ biến động cao và tâm lý nhà đầu tư còn thiếu ổn định. Các giao cắt giữa MACD và đường tín hiệu trong thời gian này cũng diễn ra liên tục, cho thấy thị trường chưa thiết lập được một xu hướng bền vững trong dài hạn.

Bước sang giai đoạn năm 2023, biểu đồ cho thấy sự ổn định hơn trong chuyển động của MACD. Mặc dù vẫn có những đợt điều chỉnh ngắn hạn, nhưng nhìn chung, MACD duy trì quanh vùng dương trong phần lớn thời gian từ quý II/2023 đến đầu năm 2024. Điều này phản ánh một xu thế tăng giá chủ đạo và tâm lý tích cực từ phía nhà đầu tư. Đáng chú ý, các điểm MACD cắt lên đường tín hiệu trong giai đoạn này thường trùng khớp với các nhịp tăng mạnh của giá cổ phiếu, cho thấy độ chính xác tương đối cao của chỉ báo này trong việc phát hiện tín hiệu mua trong bối cảnh thị trường ổn định.

Tuy nhiên, từ giữa năm 2024 trở đi, MACD bắt đầu có dấu hiệu suy yếu và nhiều lần cắt xuống dưới đường tín hiệu. Dù biên độ dao động không lớn như trước, nhưng tần suất xuất hiện các tín hiệu bán lại tăng lên. Điều này có thể phản ánh sự phân hóa trong kỳ vọng của nhà đầu tư hoặc ảnh hưởng từ các yếu tố kinh tế vĩ mô. Khi MACD tiến gần về mức 0 và dao động trong biên độ hẹp, thị trường có xu hướng đi ngang hoặc thiếu động lực rõ ràng để hình thành xu thế mới.

Một điểm cần lưu ý là bản chất của MACD là chỉ báo trễ (lagging), do được tính toán từ các đường trung bình động hàm mũ (EMA). Vì vậy, tín hiệu từ MACD không mang tính dự báo trước, mà chủ yếu phản ánh các xu hướng đã và đang hình thành. Do đó, trong quá trình áp dụng vào thực tế hoặc kết hợp với mô hình học máy, MACD nên được sử dụng như một công cụ hỗ trợ trực quan, kết hợp cùng các chỉ báo sớm (leading indicators) như RSI, khối lượng giao dịch hoặc dữ liệu định lượng để tăng độ tin cậy.

Tóm lại, biểu đồ MACD của cổ phiếu VGC thể hiện rõ các giai đoạn tăng – giảm luân phiên của thị trường trong hơn ba năm qua. Mặc dù MACD không cung cấp khả năng dự đoán mạnh mẽ như các mô hình học máy, nhưng việc sử dụng chỉ báo này trong bước khai phá dữ liệu vẫn giúp định hướng tốt cho quá trình lựa chọn mô hình, đánh giá chu kỳ thị trường, và hỗ trợ diễn giải kết quả mô phỏng sau này.

Kết luận chương 2:

Chương 2 đã cung cấp cái nhìn toàn diện về thực trạng hoạt động và tình hình tài chính của Tổng Công ty Viglacera – CTCP trong giai đoạn 2022–2024, làm rõ các xu hướng biến động và yếu tố ảnh hưởng đến giá cổ phiếu. Những phân tích này không chỉ phản ánh bức tranh tài chính hiện tại mà còn đặt nền tảng quan trọng cho việc xây dựng mô hình dự báo chính xác hơn trong phần nghiên cứu tiếp theo.

Tiếp nối cơ sở này, Chương 3 sẽ trình bày kết quả và phương pháp nghiên cứu được sử dụng nhằm xây dựng và đánh giá mô hình dự báo giá cổ phiếu cho Viglacera – CTCP.

CHƯƠNG 3: QUY TRÌNH VÀ KẾT QUẢ NGHIÊN CỨU

3.1. Mô tả dữ liệu

3.1.1. Nguồn dữ liệu và phạm vi thu thập

Trong khuôn khổ nghiên cứu này, dữ liệu đầu vào được thu thập từ hai nguồn thông tin tài chính phổ biến và có độ tin cậy cao tại Việt Nam, đó là Cafef.vn và Vietstock.vn. Đây là những nền tảng chuyên cung cấp thông tin về thị trường chứng khoán, dữ liệu giao dịch cổ phiếu, báo cáo tài chính, và các chỉ số kinh doanh của các doanh nghiệp niêm yết trên thị trường chứng khoán Việt Nam. Việc sử dụng các nguồn dữ liệu này đảm bảo tính minh bạch, đầy đủ và cập nhật – các yếu tố cần thiết cho quá trình huấn luyện và kiểm định mô hình dự báo giá cổ phiếu.

Đối tượng nghiên cứu chính là cổ phiếu của Tổng Công ty Viglacera – CTCP (mã chứng khoán: VGC), một doanh nghiệp niêm yết trên sàn HOSE, hoạt động chủ yếu trong lĩnh vực vật liệu xây dựng và đầu tư bất động sản công nghiệp.

Khoảng thời gian thu thập dữ liệu kéo dài từ ngày 01/03/2022 đến ngày 28/02/2025, tương ứng với 750 phiên giao dịch, tạo thành một chuỗi dữ liệu thời gian có độ dài phù hợp để triển khai các mô hình học máy (Random Forest Regression) và học sâu (LSTM). Việc lựa chọn khoảng thời gian ba năm liên tiếp được thực hiện nhằm đảm bảo dữ liệu có tính đại diện cao, phản ánh được xu hướng thị trường sau giai đoạn ảnh hưởng bởi đại dịch COVID-19 cũng như sự phục hồi kinh tế giai đoạn 2022–2025.

Hơn nữa, khoảng thời gian này cũng bao hàm đủ các biến động lên xuống của thị trường, bao gồm cả các pha điều chỉnh, tích lũy và tăng trưởng, từ đó giúp đánh giá được hiệu năng mô hình dự báo trong nhiều điều kiện thị trường khác nhau. Bên cạnh đó, dữ liệu thu thập có định dạng thời gian liên tục theo ngày giao dịch thực tế, giúp thuận lợi trong việc xây dựng chuỗi thời gian và áp dụng các thuật toán hồi quy chuỗi thời gian.

3.1.2. Các biến quan sát ban đầu và ý nghĩa

Trong nghiên cứu này, tập dữ liệu được sử dụng bao gồm các thông tin giao dịch cổ phiếu của Tổng Công ty Viglacera – CTCP, thu thập liên tục theo từng phiên giao dịch từ ngày 01/03/2022 đến ngày 28/02/2025. Bộ dữ liệu bao gồm bảy biến quan sát, trong đó có một biến mục tiêu và sáu biến đầu vào. Các biến này được lựa chọn dựa trên cơ sở lý thuyết tài chính và thực tiễn nghiên cứu các yếu tố ảnh hưởng đến giá cổ phiếu.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 750 entries, 0 to 749
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Ngày        750 non-null    object
 1   Lần cuối    750 non-null    object
 2   Mở          750 non-null    object
 3   Cao         750 non-null    object
 4   Thấp        750 non-null    object
 5   KL          750 non-null    object
 6   % Thay đổi  750 non-null    object
dtypes: object(7)
memory usage: 41.1+ KB
(None,
          Ngày  Lần cuối        Mở       Cao      Thấp       KL % Thay đổi
 0  28/02/2025  52,100.0  52,000.0  52,800.0  51,600.0  888.40K     -0.19%
 1  27/02/2025  52,200.0  52,600.0  52,600.0  51,200.0  801.30K     -0.76%
 2  26/02/2025  52,600.0  51,800.0  53,000.0  51,800.0  918.40K      1.54%
 3  25/02/2025  51,800.0  53,000.0  53,200.0  51,400.0    1.42M     -2.08%
 4  24/02/2025  52,900.0  53,100.0  53,200.0  52,400.0  605.80K      0.19%)
```

Cụ thể, biến mục tiêu trong mô hình là giá đóng cửa (Close) – đại diện cho giá trị cổ phiếu tại thời điểm kết thúc phiên giao dịch. Đây là thông số được thị trường đặc biệt quan tâm, thường dùng để xác định xu hướng giá trong ngắn và trung hạn, đồng thời đóng vai trò là đầu ra cần dự báo của bài toán hồi quy.

Về phía các biến độc lập, đầu tiên là giá mở cửa (Open) – thể hiện mức giá khởi điểm trong phiên giao dịch. Biến này phản ánh kỳ vọng ban đầu của thị trường và là chỉ báo quan trọng cho xu hướng trong ngày. Tiếp theo là giá cao nhất (High) và giá thấp nhất (Low) trong phiên, lần lượt ghi nhận mức giá tối đa và tối thiểu cổ phiếu đạt được trong phiên đó. Sự chênh lệch giữa hai giá trị này góp phần thể hiện biên độ dao động và mức độ biến động thị trường trong ngày.

Một biến quan sát quan trọng khác là khối lượng giao dịch (Volume), thể hiện số lượng cổ phiếu được trao tay trong phiên. Đây là chỉ báo về thanh khoản và mức độ quan tâm của nhà đầu tư đối với mã cổ phiếu, đồng thời đóng vai trò là yếu tố nội tại phản ánh sức mạnh hoặc độ yếu của một xu hướng giá.

Cuối cùng, biến tỷ lệ thay đổi giá hằng ngày (% Daily_Change) được tính bằng phần trăm chênh lệch giữa giá đóng cửa phiên hiện tại và giá đóng cửa phiên liền trước. Biến này cung cấp thông tin về mức độ biến động tương đối của giá cổ phiếu theo thời gian, từ đó hỗ trợ mô hình trong việc nhận diện các chu kỳ tăng giảm ngắn hạn.

Việc kết hợp đồng thời các biến phản ánh giá trị giao dịch (Open, High, Low, Close), mức biến động (Daily_Change) và yếu tố thanh khoản (Volume) được kỳ vọng

sẽ giúp mô hình Random Forest Regression khai thác tốt các đặc trưng nội tại của dữ liệu, từ đó nâng cao độ chính xác trong việc dự báo giá cổ phiếu trong các phiên tiếp theo.

Trong quá trình thu thập dữ liệu lịch sử giá cổ phiếu từ các nguồn dữ liệu tài chính uy tín như Cafef.vn và Vietstock.vn, dữ liệu ban đầu ghi nhận các biến số ở định dạng `object`. Cụ thể, toàn bộ các cột như "Lần cuối", "Mở", "Cao", "Thấp", "KL" và "% Thay đổi" đều được hệ thống nhận diện dưới dạng chuỗi ký tự thay vì dạng số học. Đây là đặc trưng phổ biến của dữ liệu tài chính thô, khi các số liệu thường được biểu diễn dưới định dạng dễ đọc dành cho người dùng cuối, bao gồm các ký tự đặc biệt như dấu phẩy để phân cách hàng nghìn, đơn vị khối lượng (chẳng hạn như K hoặc M), dấu phần trăm hoặc ký hiệu âm dương phản ánh chiều hướng biến động giá.

Nếu giữ nguyên trạng thái này, dữ liệu sẽ không thể trực tiếp tham gia vào quá trình xử lý số học, chuẩn hóa hay huấn luyện các mô hình dự báo. Việc chuẩn hóa dữ liệu do đó là bước thiết yếu để đưa tập dữ liệu về định dạng số học thống nhất, phù hợp với yêu cầu của các thuật toán học máy và học sâu.

Đầu tiên, toàn bộ các ký tự không liên quan đến giá trị số được loại bỏ hoàn toàn. Những ký hiệu phân tách hàng nghìn như dấu phẩy, đơn vị biểu thị khối lượng như K (nghìn) và M (triệu), cũng như dấu phần trăm được xử lý nhằm đảm bảo giữ lại phần giá trị số học nguyên bản. Tiếp theo, cột "Ngày" được chuyển đổi sang định dạng thời gian chuẩn Datetime, điều này có ý nghĩa quan trọng giúp đảm bảo dữ liệu chuỗi thời gian được sắp xếp chính xác về mặt trình tự và hỗ trợ trực tiếp cho các phương pháp phân chia dữ liệu sau này.

Để đảm bảo chất lượng dữ liệu đầu vào, các bản ghi thiếu dữ liệu, hoặc có dấu hiệu bất thường được rà soát và xử lý nhằm loại bỏ ảnh hưởng tiêu cực đến hiệu quả huấn luyện mô hình. Các thao tác kiểm tra lỗi định dạng, kiểm tra thiếu sót và xác nhận đồng bộ dữ liệu giữa các cột cũng được thực hiện cẩn trọng.

Sau khi hoàn thiện toàn bộ quy trình xử lý dữ liệu, các biến số ban đầu vốn có định dạng object đã được chuyển đổi hoàn toàn sang định dạng float64 cho các biến số học, và datetime64 cho biến thời gian. Tập dữ liệu đầu ra sau chuẩn hóa có thể sẵn sàng đưa vào giai đoạn chuẩn hóa tiếp theo cũng như huấn luyện trên các thuật toán như Random Forest và LSTM.

Quá trình xử lý dữ liệu ban đầu giữ vai trò hết sức quan trọng, không chỉ nhằm đảm bảo sự chính xác và đồng nhất trong cấu trúc dữ liệu, mà còn giúp tối ưu hóa khả năng học tập và khai thác mẫu dữ liệu của các mô hình dự báo hiện đại. Đây là một trong những yếu tố then chốt quyết định đến độ ổn định và hiệu quả của toàn bộ hệ thống dự báo giá cổ phiếu được triển khai trong nghiên cứu.

3.2. Tiền xử lí dữ liệu

3.2.1. Làm sạch và định dạng dữ liệu

```
df.columns = ['Date', 'Close', 'Open', 'High', 'Low', 'Volume', '% Daily_Change']
df['Date'] = pd.to_datetime(df['Date'], dayfirst=True)
```

Đầu tiên, dữ liệu được đọc từ tệp CSV gốc và hiển thị thông tin tổng quan nhằm kiểm tra định dạng các cột và phát hiện các lỗi tiềm ẩn. Sau đó, các tên cột được chuẩn hóa theo quy chuẩn quốc tế bao gồm: Date, Close, Open, High, Low, Volume, và % Daily_Change, cột Date được chuyển đổi về định dạng ngày tháng (datetime) để đảm bảo khả năng xử lý thời gian trong mô hình chuỗi thời gian.

```
cols_to_clean = ['Close', 'Open', 'High', 'Low', 'Volume', '% Daily_Change']
for col in cols_to_clean:
    df[col] = df[col].apply(convert_price)
```

Tiếp theo các giá trị số (bao gồm giá cổ phiếu và phần trăm thay đổi) được làm sạch khỏi ký hiệu %, dấu ,, cũng như hậu tố K, M để thống nhất đơn vị. Một hàm xử lý giá trị số đã được xây dựng để thực hiện thao tác chuyển đổi này một cách linh hoạt.

```
for col in ['Close', 'Open', 'High', 'Low']:
    df[col] = (df[col] / 1000).round(1)
df['Volume'] = df['Volume'].round(0).astype(int)
```

Sau quá trình làm sạch, các cột Close, Open, High, Low được chuyển đổi về đơn vị nghìn đồng để tiện cho việc phân tích tài chính, đồng thời được làm tròn đến một chữ số thập phân. Riêng cột Volume được làm tròn về số nguyên để phản ánh sát thực tế giao dịch cổ phiếu.

Các bước làm sạch và chuẩn hóa dữ liệu này là tiền đề quan trọng nhằm đảm bảo độ chính xác, nhất quán và khả năng tương thích với các mô hình dự báo trong các bước tiếp theo.

3.2.2. Chuẩn hóa dữ liệu

Trong quá trình huấn luyện mô hình dự báo giá cổ phiếu, đặc biệt là đối với các mô hình học sâu như LSTM, việc chuẩn hóa dữ liệu là bước tiền

xử lý rất quan trọng nhằm đảm bảo hiệu quả huấn luyện và độ chính xác của mô hình.

Thứ nhất, dữ liệu tài chính như giá cổ phiếu, khối lượng giao dịch thường có các thang đo khác nhau (ví dụ: giá cổ phiếu có thể dao động từ vài chục đến vài trăm, còn khối lượng giao dịch có thể hàng triệu). Nếu không chuẩn hóa, các biến có giá trị lớn sẽ chi phối quá trình tính toán hàm mất mát và quá trình cập nhật trọng số, dẫn đến mô hình học lệch về các biến có quy mô lớn hơn, bỏ qua các biến có giá trị nhỏ hơn nhưng lại có ý nghĩa quan trọng về mặt dự báo.

Thứ hai, chuẩn hóa giúp tăng tốc quá trình hội tụ của các thuật toán tối ưu hóa như gradient descent. Khi các đầu vào có phân phối đồng đều hơn (chẳng hạn có trung bình bằng 0 và độ lệch chuẩn bằng 1), các bước nhảy trong không gian tham số trở nên ổn định hơn, tránh hiện tượng "bước nhảy quá lớn" hoặc "bước nhảy quá nhỏ" gây kẹt tại các điểm cực tiểu cục bộ hoặc hội tụ chậm.

Thứ ba, một số mô hình học sâu sử dụng các hàm kích hoạt như sigmoid hoặc tanh có đầu ra giới hạn trong khoảng [-1, 1] hoặc [0, 1]. Nếu dữ liệu đầu vào chưa chuẩn hóa, các giá trị lớn có thể đẩy hàm kích hoạt vào vùng bão hòa, khiến đạo hàm gần bằng 0, gây ra hiện tượng vanishing gradient (tiêu biến gradient), làm cho mô hình khó học được mối quan hệ giữa đầu vào và đầu ra.

Cuối cùng, việc chuẩn hóa cũng đảm bảo tính ổn định khi so sánh các mô hình khác nhau trên cùng một tập dữ liệu, bởi khi dữ liệu được đưa về cùng một thang đo, ta có thể đánh giá khách quan hiệu quả mô hình dựa trên cùng điều kiện huấn luyện.

```
close_prices = df[['Date', 'Close']].dropna().copy()
close_prices.set_index('Date', inplace=True)
scaler = MinMaxScaler()
scaled_data = scaler.fit_transform(close_prices)
```

3.2.3. Tạo tập dữ liệu huấn luyện theo cửa sổ thời gian và chia dữ liệu

Trong bài toán dự báo chuỗi thời gian tài chính như giá cổ phiếu, việc xây dựng tập dữ liệu huấn luyện đóng vai trò then chốt, bởi các mô hình học máy và học sâu không thể xử lý trực tiếp toàn bộ chuỗi giá trị mà cần chuyển đổi thành các cặp dữ liệu đầu vào – đầu ra phù hợp. Do đặc trưng dữ liệu tài chính có sự phụ thuộc mạnh mẽ vào lịch sử, phương pháp cửa sổ thời gian (sliding window) đã được áp dụng nhằm khai thác các thông tin từ những phiên giao dịch trước đó để dự báo giá trị tương lai.

Việc chia dữ liệu tuần tự theo thứ tự thời gian, thay vì chia ngẫu nhiên, nhằm đảm bảo tuân thủ đúng bản chất chuỗi thời gian và tránh hiện tượng "rò rỉ dữ liệu" (data leakage) trong quá trình huấn luyện, vốn có thể dẫn đến kết quả đánh giá sai lệch.

Phương pháp trượt cửa sổ đã được áp dụng đồng nhất cho cả hai mô hình: Random Forest Regression (sử dụng dữ liệu dạng bảng sau xử lý đặc trưng) và LSTM (sử dụng dữ liệu dạng chuỗi 3D). Nhờ đó, đảm bảo tính nhất quán khi so sánh hiệu quả hai mô hình trên cùng một bộ dữ liệu gốc.

```
def create_dataset(dataset, window_size=30):
    X, y = [], []
    for i in range(window_size, len(dataset)):
        X.append(dataset[i - window_size:i, 0])
        y.append(dataset[i, 0])
    return np.array(X), np.array(y)

X, y = create_dataset(scaled_data, 30)
```

Cụ thể, với một biến đầu vào như giá đóng cửa (Close), tập dữ liệu được tái cấu trúc thành dạng mà mỗi mẫu đầu vào Xt bao gồm các giá trị giá đóng cửa của w ngày liên tiếp trước thời điểm T1 , và đầu ra tương ứng Yt là giá đóng cửa tại ngày thứ t. Trong nghiên cứu này, độ dài cửa sổ được lựa chọn là w = 30, nghĩa là mô hình sẽ sử dụng dữ liệu của 30 phiên giao dịch gần nhất để dự đoán giá cổ phiếu của phiên tiếp theo.

Phương pháp tạo tập dữ liệu này mặc dù thường được áp dụng trong các mô hình học sâu như LSTM, nhưng hoàn toàn có thể mở rộng cho các mô hình hồi quy truyền thống như Random Forest, bằng cách coi mỗi chuỗi 30 ngày là một vector đặc trưng đầu vào. Điều này cho phép mô hình khai thác mối quan hệ giữa các quan sát thời gian mà không cần thiết kế đặc trưng thủ công cho từng ngày.

```
split = int(len(X) * 0.8)
X_train, y_train = X[:split], y[:split]
X_test, y_test = X[split:], y[split:]
```

Sau khi tạo tập dữ liệu theo cửa sổ thời gian, bộ dữ liệu được chia thành hai phần: tập huấn luyện (80%) và tập kiểm tra (20%) theo đúng thứ tự thời gian, nhằm đảm bảo tính nhất quán trong dự báo chuỗi thời gian và tránh hiện tượng rò rỉ dữ liệu từ tương lai vào quá trình huấn luyện.

Cách tiếp cận này giúp mô hình học được quy luật vận động theo thời gian của giá cổ phiếu VGC và tạo điều kiện cho việc đánh giá mô hình một cách thực tiễn chính xác.

3.3. Khai phá dữ liệu

3.3.1. Thống kê mô tả

	Date	Close	Open	High	Low	Volume	D
count	750	750.000000	750.000000	750.000000	750.000000	7.500000e+02	
mean	2023-08-28 06:24:00.000000256	45.144800	45.200000	46.058267	44.325467	1.105206e+06	
min	2022-03-01 00:00:00	25.400000	23.600000	25.500000	23.600000	1.282000e+05	
25%	2022-11-24 06:00:00	40.025000	40.000000	40.825000	39.300000	6.566750e+05	
50%	2023-08-26 12:00:00	45.650000	45.700000	46.500000	44.800000	9.447000e+05	
75%	2024-05-30 18:00:00	51.200000	51.275000	51.900000	50.600000	1.387500e+06	
max	2025-02-28 00:00:00	65.700000	65.700000	67.900000	64.000000	5.340000e+06	
std	NaN	8.098302	8.107219	8.208474	8.037896	6.646091e+05	

Trong quá trình phân tích dữ liệu giá cổ phiếu VGC, thống kê mô tả được thực hiện trên tập dữ liệu gồm 750 phiên giao dịch từ ngày 01/03/2022 đến ngày 28/02/2025. Kết quả thống kê mô tả đã phản ánh rõ nét đặc điểm biến động giá cổ phiếu trong giai đoạn nghiên cứu.

Trước hết, xét về giá đóng cửa (Close), mức giá trung bình trong toàn bộ giai đoạn đạt khoảng 45,14 nghìn đồng, với giá trị thấp nhất ghi nhận ở mức 25,4 nghìn đồng và giá trị cao nhất đạt 65,7 nghìn đồng. Biên độ dao động lớn này cho thấy cổ phiếu VGC đã trải qua các pha biến động mạnh trong suốt thời gian nghiên cứu. Độ lệch chuẩn của giá đóng cửa là 8,09, phản ánh mức biến động tương đối cao quanh giá trị trung bình.

Xét thêm các mức giá khác như giá mở cửa (Open), giá cao nhất (High) và giá thấp nhất (Low) đều có xu hướng tương đồng với giá đóng cửa. Giá mở cửa trung bình đạt 45,2 nghìn đồng; giá cao nhất trung bình là 46,05 nghìn đồng; trong khi giá thấp nhất trung bình là 44,32 nghìn đồng. Điều này cho thấy sự đồng pha trong dao động giá giữa các chỉ số giá trong từng phiên, phù hợp với đặc trưng thị trường cổ phiếu.

Về khối lượng giao dịch (Volume), mức trung bình đạt 1,1 triệu cổ phiếu mỗi phiên, với mức thấp nhất chỉ khoảng 128,200 cổ phiếu và mức cao nhất lên tới hơn 5,34 triệu cổ phiếu. Chênh lệch đáng kể về khối lượng giao dịch giữa các phiên phản ánh sự

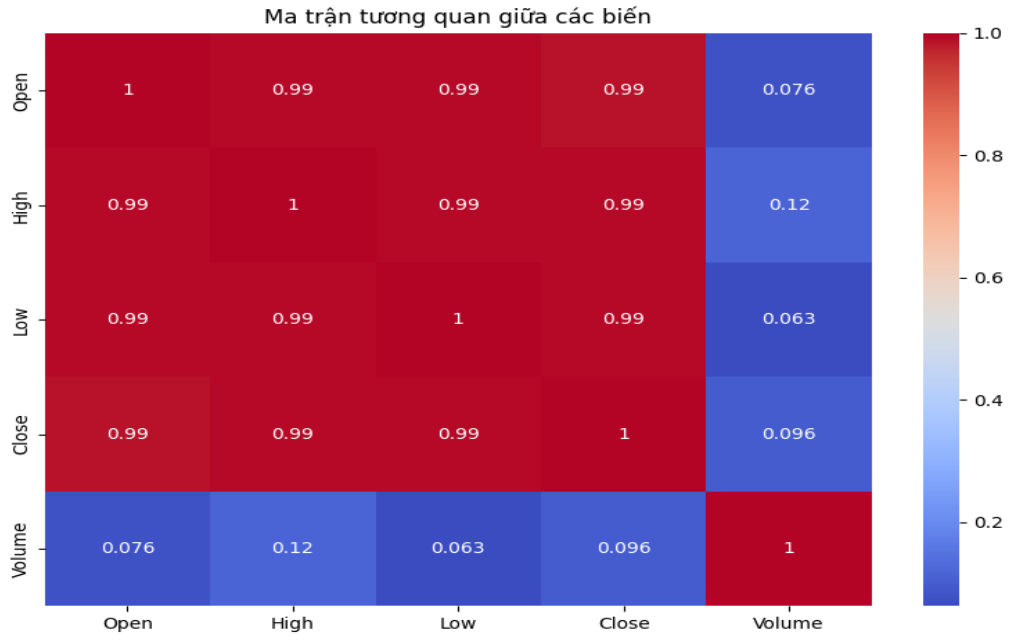
tham gia thị trường có tính chu kỳ, phụ thuộc vào các thông tin vĩ mô và vi mô tác động đến doanh nghiệp.

Ngoài ra, khi xét các phân vị thống kê, có thể thấy giá đóng cửa tại phân vị 25% (Q1) là 40,03 nghìn đồng, trong khi phân vị 75% (Q3) đạt 51,2 nghìn đồng. Điều này cho thấy phần lớn các phiên giao dịch tập trung trong khoảng 40 - 51 nghìn đồng, còn những mức giá thấp hơn 30 hay cao hơn 60 nghìn đồng chỉ xuất hiện trong các giai đoạn thị trường có biến động mạnh.

Như vậy, thông qua thống kê mô tả, có thể nhận định rằng dữ liệu giá cổ phiếu VGC trong giai đoạn nghiên cứu mang đầy đủ các đặc điểm điển hình của chuỗi tài chính: dao động mạnh, có xu hướng chu kỳ, xuất hiện những pha tăng giảm rõ nét và biến động thanh khoản theo từng giai đoạn thị trường.

3.3.2. Phân tích sự tương quan giữa các biến

Hình 3.1. Ma trận tương quan giữa các biến



(nguồn: Tác giả tổng hợp và tính toán)

Hình trên thể hiện ma trận tương quan Pearson giữa các biến định lượng trong tập dữ liệu cổ phiếu VGC, bao gồm các biến giá trong ngày (Open, High, Low, Close) và biến khối lượng giao dịch (Volume). Phân tích ma trận này là một bước quan trọng trong quá trình khai phá dữ liệu, nhằm đánh giá mức độ liên kết tuyến tính giữa các biến độc

lập và biến mục tiêu, từ đó đưa ra các quyết định phù hợp trong khâu chọn biến cho mô hình.

Trước hết, dễ dàng nhận thấy rằng các biến Open, High, Low và Close có hệ số tương quan rất cao với nhau, đều dao động quanh ngưỡng 0.99. Đây là mức tương quan gần như tuyệt đối, cho thấy mối quan hệ tuyến tính mạnh mẽ giữa các mức giá trong cùng một phiên giao dịch. Hiện tượng này không phải là điều bất thường, bởi lẽ trong thị trường tài chính, các mức giá trong ngày thường được xác lập dựa trên một chuỗi giao dịch liên tục và chịu ảnh hưởng chung từ xu hướng thị trường, thông tin kinh tế và dòng tiền tại thời điểm cụ thể. Do đó, việc các biến giá trong ngày cùng biến động theo một hướng là điều có thể lý giải về mặt bản chất dữ liệu.

Tuy nhiên, cũng cần lưu ý rằng mức độ tương quan cao giữa các biến độc lập có thể dẫn đến hiện tượng đa cộng tuyến (multicollinearity) trong các mô hình hồi quy truyền thống. Khi các biến như Open, High, Low và Close được sử dụng đồng thời trong một mô hình tuyến tính (ví dụ như hồi quy tuyến tính đa biến), sẽ rất khó để xác định ảnh hưởng riêng biệt của từng biến lên biến mục tiêu, do các biến này thay đổi gần như đồng thời. Hậu quả là các hệ số hồi quy có thể trở nên không ổn định, độ tin cậy giảm sút, và mô hình trở nên nhạy cảm với nhiễu.

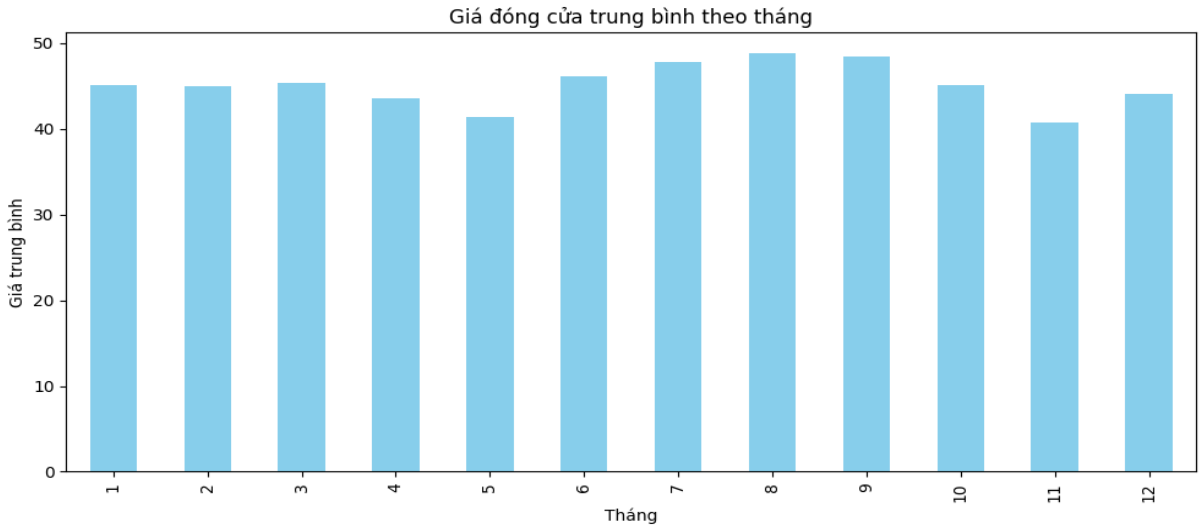
Ngược lại, nếu quan sát mối tương quan giữa biến Volume (khối lượng giao dịch) với các biến còn lại, ta thấy rằng hệ số tương quan đều ở mức rất thấp, dao động trong khoảng từ 0.04 đến 0.12. Điều này cho thấy khối lượng giao dịch không có mối quan hệ tuyến tính rõ ràng với các mức giá, phản ánh rằng yếu tố này có thể mang thông tin bổ sung độc lập, không bị đồng biến quá mức với các biến còn lại. Đây là một đặc điểm quan trọng trong quá trình xây dựng mô hình, bởi lẽ việc đưa Volume vào tập biến đầu vào có thể giúp mô hình học được chiều sâu hành vi giao dịch từ một khía cạnh khác – không chỉ dựa vào giá – đồng thời giảm nguy cơ đa cộng tuyến nghiêm trọng.

Ngoài ra, cũng cần phân biệt giữa tác động của đa cộng tuyến trong các loại mô hình khác nhau. Đối với các mô hình hồi quy tuyến tính cổ điển, việc tồn tại tương quan cao giữa các biến độc lập là vấn đề đáng quan ngại. Tuy nhiên, trong các mô hình phi tuyến và phi tham số như Random Forest Regression – mô hình chính được sử dụng trong đề tài này – tác động của đa cộng tuyến thường không gây ảnh hưởng nghiêm trọng. Điều này là nhờ vào cơ chế ngẫu nhiên hóa khi chọn tập biến tại mỗi cây con,

cùng với quá trình trung bình hóa kết quả giữa nhiều cây quyết định, giúp mô hình giảm phụ thuộc vào một tập biến cố định và nâng cao tính tổng quát.

3.3.2. Phân tích giá đóng cửa theo tháng

Hình 3.2. Giá đóng cửa trung bình theo tháng



(nguồn: Tác giả tổng hợp và tính toán)

Biểu đồ trên thể hiện giá đóng cửa trung bình hàng tháng của cổ phiếu Tổng công ty Viglacera – CTCP (VGC) trong khoảng thời gian nghiên cứu. Việc tổng hợp và trực quan hóa giá theo từng tháng là một bước phân tích quan trọng trong khai phá dữ liệu chuỗi thời gian, cho phép nhận diện các mô hình theo mùa (seasonality), chu kỳ hoạt động và hành vi lặp lại tiềm ẩn trong dữ liệu tài chính.

Trước hết, có thể thấy rằng giá cổ phiếu VGC không biến động ngẫu nhiên theo thời gian mà thể hiện những mẫu hình theo tháng tương đối rõ ràng. Cụ thể, các tháng giữa năm như tháng 7, 8 và 9 ghi nhận mức giá đóng cửa trung bình cao hơn hẳn so với mặt bằng chung, dao động quanh mức 48.5–49 nghìn đồng/cổ phiếu. Đây có thể là giai đoạn tích cực của thị trường hoặc của riêng doanh nghiệp, thường trùng với thời điểm công bố báo cáo tài chính quý II hoặc các thông tin vĩ mô tích cực liên quan đến hoạt động xây dựng – bất động sản, lĩnh vực cốt lõi của Viglacera.

Ngược lại, một số tháng như tháng 5 và tháng 11 lại cho thấy mức giá trung bình thấp hơn rõ rệt, chỉ dao động trong khoảng 41–42 nghìn đồng/cổ phiếu. Điều này gợi mở khả năng tồn tại các giai đoạn điều chỉnh ngắn hạn hoặc tâm lý bi quan mang tính chu kỳ. Ví dụ, tháng 5 thường là thời điểm sau mùa đại hội cổ đông và trước mùa công bố kết quả kinh doanh quý II, khi thị trường thiếu động lực rõ ràng. Trong khi đó, tháng

11 lại trùng với thời điểm nhà đầu tư thường có xu hướng chốt lời cuối năm, hoặc thận trọng trước kỳ nghỉ Tết và kế hoạch tái cơ cấu danh mục.

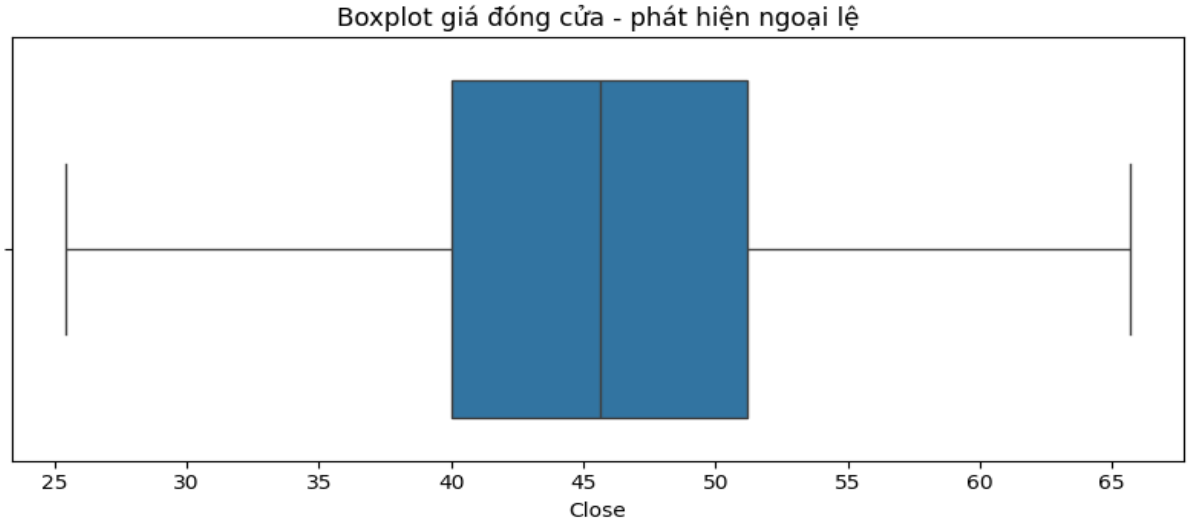
Điều đáng chú ý là độ lệch giữa các tháng trong năm là tương đối rõ ràng, thay vì dao động ngẫu nhiên, điều này cho thấy sự tồn tại của yếu tố mùa vụ trong biến động giá cổ phiếu VGC. Việc nhận diện được yếu tố chu kỳ như vậy mang ý nghĩa đặc biệt quan trọng trong bối cảnh dự báo chuỗi thời gian, bởi vì hầu hết các mô hình dự báo truyền thống như ARIMA hoặc mô hình học sâu như LSTM đều hoạt động hiệu quả hơn khi được cung cấp các đặc trưng thời gian có tính lặp lại theo định kỳ.

Hơn nữa, việc phân tích xu hướng trung bình theo tháng cũng có thể hỗ trợ các quyết định đầu tư thực tiễn. Các tháng có xu hướng tăng mạnh có thể được xem là thời điểm chiến lược để nắm giữ hoặc gia tăng vị thế, trong khi những tháng yếu hơn cần được quan sát kỹ lưỡng để phòng ngừa rủi ro hoặc điều chỉnh kỳ vọng.

Tổng kết lại, biểu đồ giá trung bình theo tháng đã làm nổi bật một mô hình theo mùa khá đặc trưng của cổ phiếu VGC. Việc khai thác hiệu quả thông tin này không chỉ giúp nâng cao chất lượng mô hình dự báo mà còn góp phần lý giải sâu hơn hành vi thị trường từ góc nhìn hành vi nhà đầu tư và yếu tố chu kỳ kinh tế.

3.3.3. Giá trị ngoại lai

Hình 3.3. Boxplot giá đóng cửa



(nguồn: Tác giả tổng hợp và tính toán)

Biểu đồ hộp (boxplot) trên thể hiện phân bố của biến giá đóng cửa (Close) của cổ phiếu VGC trong giai đoạn từ tháng 3/2022 đến tháng 3/2025. Đây là một công cụ trực quan hữu ích để xác định các điểm ngoại lệ (outliers) – những giá trị khác biệt đáng kể

so với phần còn lại của dữ liệu.Quan sát từ boxplot cho thấy:

- Khoảng giữa của dữ liệu (IQR – interquartile range) tập trung chủ yếu trong khoảng từ khoảng 40.000 đến 52.000 đồng/cổ phiếu, đây là vùng giá phổ biến nhất.
- Giá trị trung vị (median) của dữ liệu xấp xỉ 46.000 đồng, cho thấy mức cân bằng của phân phối giá.

Hai ngưỡng cực trị (whiskers) kéo dài từ khoảng dưới 30.000 đến trên 60.000 đồng, cho thấy biên độ biến động tương đối lớn.Tuy nhiên, không xuất hiện các dấu hiệu rõ ràng của điểm ngoại lệ nghiêm trọng, tức là không có điểm dữ liệu nào vượt quá giới hạn ngoài 1.5 lần IQR.

Việc không phát hiện các ngoại lệ rõ rệt giúp tăng độ tin cậy của mô hình dự báo, do dữ liệu huấn luyện không bị sai lệch bởi các điểm bất thường. Tuy nhiên, nếu trong quá trình cập nhật dữ liệu mới xuất hiện các điểm dị biệt (do sự kiện thị trường hoặc hoạt động nội bộ công ty), cần áp dụng các kỹ thuật xử lý như winsorization hoặc robust scaling để bảo toàn chất lượng mô hình.

3.4. Xây dựng mô hình dự báo LSTM

3.4.1. Chuẩn bị dữ liệu đầu vào

Mô hình LSTM được thiết kế đặc biệt cho dữ liệu chuỗi thời gian, do đó quá trình chuẩn bị dữ liệu đầu vào cần đảm bảo tính tuần tự và có tính phụ thuộc thời gian. Trong đề tài này, biến đầu vào được sử dụng là giá đóng cửa (Close) của cổ phiếu VGC, do đây là biến phản ánh trực tiếp giá trị giao dịch kết thúc mỗi phiên và thường được sử dụng để dự báo xu hướng giá. Tất cả dữ liệu đã được xử lí tại phần trước.

```
X_train = X_train.reshape((X_train.shape[0], X_train.shape[1], 1))
X_test = X_test.reshape((X_test.shape[0], X_test.shape[1], 1))
```

Cuối cùng, để tương thích với kiến trúc LSTM, tập dữ liệu X được chuyển đổi về định dạng 3 chiều (samples, timesteps, features), với timesteps = 30 và features = 1. Việc định dạng lại này là bắt buộc nhằm đảm bảo đúng yêu cầu đầu vào của các lớp LSTM trong Keras hoặc TensorFlow.

3.4.2. Xây dựng và huấn luyện mô hình LSTM

Mô hình LSTM được xây dựng với kiến trúc đơn giản và hiệu quả, bao gồm: Một lớp LSTM với 64 đơn vị nơ-ron, không trả về chuỗi (return_sequences=False), có khả năng ghi nhớ thông tin dài hạn trong chuỗi dữ liệu. Một lớp Dense đầu ra với 1 nơ-ron

nhằm dự báo giá đóng cửa cổ phiếu tại thời điểm kế tiếp.

```
set_seed(42)
model = Sequential()
model.add(LSTM(64, return_sequences=False, input_shape=(X_train.shape[1], 1)))
model.add(Dense(1))
model.compile(optimizer='adam', loss='mean_squared_error')
```

Mô hình LSTM trong đề tài được xây dựng bằng cách sử dụng API Sequential từ thư viện Keras, cho phép sắp xếp các lớp theo thứ tự tuyến tính. Trước khi khởi tạo mô hình, tác giả tiến hành cố định hạt giống ngẫu nhiên bằng lệnh `set_seed(42)` nhằm đảm bảo tính tái lập của quá trình huấn luyện. Tiếp theo, một lớp LSTM với 64 đơn vị được thêm vào mô hình, trong đó tham số input_shape=(X_train.shape[1], 1) cho biết mỗi mẫu đầu vào có độ dài chuỗi là X_train.shape[1] bước thời gian, với 1 đặc trưng tại mỗi bước. Việc thiết lập return_sequences=False thể hiện rằng lớp LSTM chỉ trả về đầu ra tại bước cuối cùng, phù hợp với mục tiêu dự báo một giá trị duy nhất. Sau lớp LSTM, mô hình sử dụng thêm một lớp Dense với một nút đầu ra để thực hiện nhiệm vụ hồi quy, tức là dự đoán giá trị giá cổ phiếu trong tương lai gần. Cuối cùng, mô hình được biên dịch với thuật toán tối ưu Adam và hàm mất mát là mean squared error (MSE). Việc lựa chọn Adam giúp quá trình huấn luyện đạt hiệu quả cao nhờ cơ chế tự điều chỉnh tốc độ học, trong khi MSE là hàm mất mát phổ biến đối với các bài toán hồi quy, phản ánh mức độ chênh lệch bình phương giữa giá trị dự đoán và thực tế. Nhìn chung, cấu trúc mô hình LSTM trong đề tài được thiết kế đơn giản nhưng hợp lý, đảm bảo khả năng học được các đặc điểm chuỗi thời gian mà không gây phức tạp hóa mô hình ngay từ giai đoạn đầu.

```
model.fit(X_train, y_train, epochs=50, batch_size=32, verbose=0)
```

```
<keras.src.callbacks.history.History at 0x7de75e3c5850>
```

Sau khi mô hình LSTM được khởi tạo và biên dịch, quá trình huấn luyện được tiến hành thông qua phương thức model.fit(), trong đó dữ liệu huấn luyện được cung cấp dưới dạng X_train và y_train. Mô hình được huấn luyện trong 50 vòng lặp (epochs), nghĩa là toàn bộ tập dữ liệu được đưa qua mạng LSTM 50 lần liên tiếp để tối ưu hóa trọng số. Kích thước lô (batch size) được thiết lập là 32, đồng nghĩa với việc dữ liệu huấn luyện sẽ được chia thành các lô nhỏ gồm 32 mẫu mỗi lần cập nhật tham số. Việc lựa chọn batch size ở mức trung bình như vậy là nhằm cân bằng giữa tốc độ huấn luyện và độ ổn định của quá trình cập nhật trọng số. Ngoài ra, tham số verbose=0 được sử

dụng để tắt hiển thị thông tin chi tiết trong quá trình huấn luyện, phù hợp khi chạy mô hình trong môi trường notebook mà không cần theo dõi trực tiếp tiến trình huấn luyện. Kết quả của hàm fit() trả về một đối tượng lịch sử huấn luyện (history object), chứa thông tin về giá trị hàm mất mát sau mỗi epoch, từ đó có thể được sử dụng để trực quan hóa hoặc đánh giá mức độ hội tụ của mô hình trong các bước phân tích sau. Tổng thể, quá trình huấn luyện này đóng vai trò cốt lõi trong việc giúp mô hình LSTM học được quy luật tiềm ẩn trong chuỗi thời gian giá cổ phiếu, từ đó phục vụ cho bài toán dự báo trong giai đoạn kiểm định mô hình.

❖ Dự báo và đánh giá mô hình

```
predicted_scaled = model.predict(X_test)
predicted = scaler.inverse_transform(np.concatenate([np.zeros((len(predicted_scaled), 1)), predicted_scaled], axis=1))[:, 1]
real = scaler.inverse_transform(np.concatenate([np.zeros((len(y_test), 1)), y_test.reshape(-1, 1)], axis=1))[:, 1]
```

Sau khi huấn luyện, mô hình tiến hành dự báo trên tập kiểm tra (X_test). Kết quả dự báo được đưa về giá trị gốc bằng cách đảo chuẩn hóa (inverse_transform). Cuối cùng, kết quả được so sánh với giá trị thực tế để đánh giá hiệu suất mô hình bằng các chỉ số như MAE, RMSE, và MAPE (nếu cần)

3.4.3. Kết quả dự báo

❖ Biểu đồ so sánh kết quả dự báo của LSTM và thực tế

Hình 3.4.. Biểu đồ so sánh kết quả dự báo của LSTM và thực tế

(Nguồn: Tác giả tổng hợp và tính toán)

Biểu đồ trên thể hiện kết quả so sánh giữa giá thực tế của cổ phiếu VGC (đường màu xanh, nét liền) và giá dự báo từ mô hình LSTM (đường màu cam, nét đứt) trong khoảng thời gian từ tháng 8/2024 đến tháng 3/2025. Nhìn tổng thể, mô hình LSTM đã

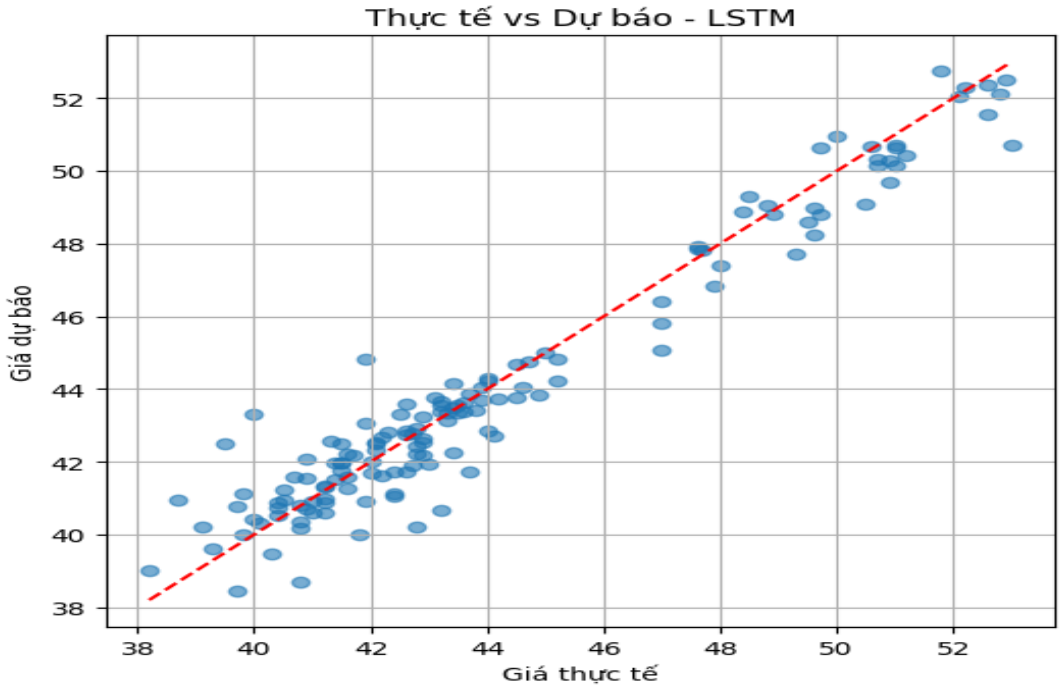
thể hiện khả năng bám sát khá tốt xu hướng thực tế của giá cổ phiếu, đặc biệt trong các giai đoạn tăng trưởng liên tục hoặc dao động ổn định.

Cụ thể, trong giai đoạn từ tháng 8 đến tháng 10/2024, khi thị trường có những pha dao động nhẹ, mô hình LSTM phản ánh chính xác các nhịp điều chỉnh giá ngắn hạn. Mặc dù tại một vài thời điểm ban đầu, mô hình có độ trễ nhẹ so với giá thực tế – đặc điểm thường gặp do cơ chế nhớ ngắn hạn – nhưng sự sai lệch này không đáng kể và nhanh chóng được điều chỉnh khi mô hình tiếp tục được cập nhật với dữ liệu mới.

Từ khoảng tháng 12/2024 đến tháng 2/2025 – giai đoạn mà giá cổ phiếu tăng mạnh và có xu hướng rõ rệt – mô hình LSTM đã dự báo đúng chiều hướng giá và thể hiện được khả năng nhận diện chu kỳ tăng tương đối mượt mà. Điều này cho thấy cơ chế bộ nhớ của LSTM phát huy hiệu quả khi đối mặt với dữ liệu có tính chu kỳ cao hoặc xu hướng rõ ràng, giúp mô hình duy trì độ chính xác ổn định trong suốt chuỗi thời gian kiểm tra.

❖ Biểu đồ tương quan giữa dự báo của LSTM so với thực tế

Hình 3.5. Biểu đồ tương quan giữa dự báo của LSTM so với thực tế



(Nguồn: Tác giả tổng hợp và tính toán)

Biểu đồ trên thể hiện mối tương quan giữa giá thực tế và giá dự báo từ mô hình LSTM, trong đó mỗi điểm biểu diễn một cặp giá trị tại cùng một thời điểm trong tập

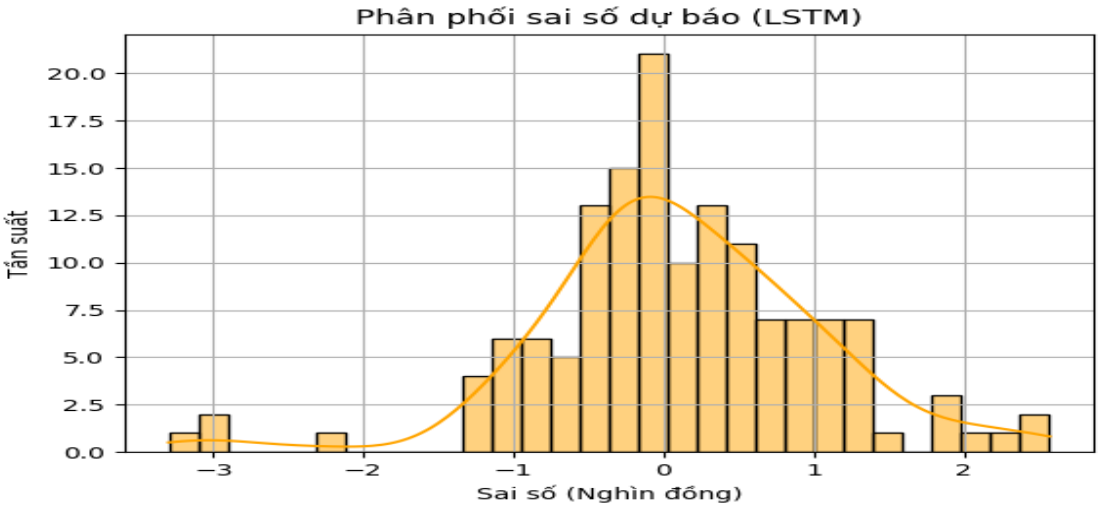
kiểm tra. Đường chấm đỏ thể hiện đường hồi quy lý tưởng (đường chéo 45 độ), tức là khi giá dự báo hoàn toàn trùng khớp với giá thực tế.

Quan sát tổng thể cho thấy, các điểm dữ liệu phân bố khá tập trung xung quanh đường lý tưởng, cho thấy mô hình LSTM có khả năng dự báo giá cổ phiếu với độ chính xác cao và ổn định. Đặc biệt, ở vùng giá từ khoảng 40.000 đến 50.000 đồng, mô hình cho kết quả rất sát với thực tế, khi đa số các điểm nằm rất gần hoặc ngay trên đường chéo.

Ngoài ra, không xuất hiện rõ các cụm điểm lệch hoàn toàn khỏi xu hướng hoặc hiện tượng "đuôi lệch" (outliers), điều này phản ánh tính nhất quán trong hiệu suất của mô hình và cho thấy LSTM không bị sai lệch nghiêm trọng tại các phiên có giá trị cực đoan. Một số điểm lệch nhẹ ở vùng giá thấp hơn có thể là kết quả của các pha điều chỉnh đột ngột của thị trường hoặc nhiễu ngắn hạn – vốn là những yếu tố khó mô hình hóa tuyệt đối chính xác.

Đường xu hướng chéo đỏ không chỉ là mốc tham chiếu để đánh giá độ lệch giữa giá dự báo và thực tế, mà còn cho thấy mức đồng biến tuyến tính mạnh giữa hai biến. Biểu đồ phân bố sai số dự báo của LSTM

Hình 3.6. Biểu đồ phân bố sai số dự báo của LSTM



(Nguồn: Tác giả tổng hợp và tính toán)

Biểu đồ trên mô tả phân phối sai số dự báo của mô hình LSTM, được tính bằng hiệu số giữa giá dự báo và giá thực tế tại từng thời điểm trong tập kiểm tra. Trục hoành thể hiện mức sai số (tính theo đơn vị nghìn đồng), còn trục tung biểu thị tần suất xuất hiện của mỗi khoảng sai số. Đường cong màu cam thể hiện phân phối chuẩn ước lượng từ dữ liệu thực tế.

Quan sát tổng thể cho thấy, phân phối sai số có dạng chuông gần chuẩn, với tâm phân bố nằm rất gần điểm 0. Điều này cho thấy mô hình không có xu hướng dự báo thiên lệch về một phía, tức là không nghiêng nhiều về việc dự báo cao hơn hay thấp hơn thực tế một cách hệ thống. Đây là một đặc điểm quan trọng, phản ánh tính khách quan và ổn định trong khả năng dự đoán của mô hình.

Ngoài ra, phần lớn các sai số đều nằm trong khoảng từ -1 đến +1 nghìn đồng, tương đương với mức sai số tương đối nhỏ so với mặt bằng giá cổ phiếu dao động quanh ngưỡng 45–50 nghìn đồng. Tần suất tập trung cao quanh vùng sai số ±0,5 cho thấy mô hình đạt độ chính xác tốt với phần lớn dự báo sai lệch rất ít so với thực tế.

Tuy nhiên, cũng có thể nhận thấy một số giá trị ngoại biên (outliers) ở cả hai phía, trong đó sai số lớn nhất dao động từ khoảng -3 đến +2,5 nghìn đồng. Những điểm này nhiều khả năng rơi vào các phiên có biến động mạnh, phản ánh độ nhạy kém của mô hình với các thay đổi đột ngột, vốn là đặc điểm chung của các mô hình học sâu trong điều kiện dữ liệu tài chính biến động cao.

3.4.5. Đánh giá hiệu năng

Sau quá trình huấn luyện và kiểm tra, mô hình mạng bộ nhớ dài ngắn hạn (LSTM) được đánh giá bằng bốn chỉ số hiệu năng phổ biến gồm: R^2 (hệ số xác định), MAE (Sai số tuyệt đối trung bình), RMSE (Căn bậc hai sai số bình phương trung bình) và MAPE (Sai số phần trăm tuyệt đối trung bình). Kết quả như sau:

Với giá trị R^2 đạt 0.9384 (tương đương 93.84%), mô hình đã thể hiện năng lực lý giải rất cao đối với biến động của giá cổ phiếu trong tập kiểm tra. Điều này có nghĩa là gần 94% phương sai của biến mục tiêu đã được mô hình giải thích thông qua các đặc trưng đầu vào, cho thấy mối quan hệ học được là đáng tin cậy và có tính khái quát hóa cao. Một hệ số R-squared gần bằng 1 không chỉ phản ánh mức độ phù hợp (goodness-of-fit) mạnh mẽ giữa kết quả dự báo và dữ liệu thực tế, mà còn là minh chứng rõ ràng cho việc mô hình đã thành công trong việc khai thác mối quan hệ động trong chuỗi thời gian. Trong bối cảnh dữ liệu tài chính thường chứa nhiều nhiễu và có tính phi tuyến cao, việc đạt được mức R^2 vượt ngưỡng 0.9 cho thấy tiềm năng ứng dụng thực tiễn của mô hình trong các hệ thống phân tích đầu tư hoặc cảnh báo rủi ro ngắn hạn.

MAE = 0.6920. Sai số tuyệt đối trung bình ở mức 0.6920 (nghìn đồng) cho thấy trung bình mỗi phiên, giá dự báo lệch dưới 700 đồng so với giá đóng cửa thực tế. Đây là mức sai số thấp và dễ chấp nhận trong ngữ cảnh dự báo tài chính ngắn hạn.

RMSE = 0.9417. Sai số bình phương trung bình căn bậc hai là 0.9417, phản ánh rằng mô hình ít bị ảnh hưởng bởi các sai lệch lớn. Việc RMSE không chênh lệch quá nhiều so với MAE chứng tỏ dữ liệu không có quá nhiều điểm dị biệt (outliers), và mô hình duy trì được độ ổn định trong toàn bộ chuỗi kiểm tra.

MAPE = 0.0158 (tức 1.58%). Sai số phần trăm tuyệt đối trung bình chỉ ở mức 1.58% cho thấy mô hình dự báo khá chính xác ở góc nhìn tương đối. Đây là một chỉ số đặc biệt quan trọng trong dự báo tài chính, vì nó thể hiện độ lệch nhỏ so với giá trị thực tế bất kể quy mô giá.

Tổng hòa bốn chỉ số đánh giá, có thể kết luận rằng mô hình LSTM đạt hiệu năng rất tốt trong việc dự báo giá cổ phiếu VGC. Đặc biệt, việc đạt được đồng thời R^2 cao và MAPE thấp chứng minh mô hình không chỉ dự báo chính xác theo nghĩa tuyệt đối, mà còn có khả năng tổng quát hóa và khả năng ổn định trên toàn chuỗi dữ liệu.

3.5. Mô hình hồi quy Random Forest

3.5.1. Chuẩn bị dữ liệu đầu vào

Trong nghiên cứu này, mô hình Random Forest Regression được triển khai với cùng cách thiết lập dữ liệu đầu vào như mô hình LSTM, nhằm đảm bảo sự đồng nhất khi so sánh hiệu quả dự báo. Cụ thể, dữ liệu đầu vào của mô hình được xây dựng dựa trên chiến lược cửa sổ thời gian trượt (sliding time window).

Mỗi quan sát đầu vào bao gồm 30 giá trị liên tiếp của biến giá đóng cửa (Close) trong 30 ngày gần nhất. Biến mục tiêu tương ứng là giá đóng cửa của ngày kế tiếp (ngày thứ 31). Toàn bộ dữ liệu đã được chuẩn hóa theo phương pháp Min-Max để đưa về cùng một thang đo, tăng hiệu quả học của mô hình. Sau khi tạo tập dữ liệu đặc trưng và mục tiêu, dữ liệu được chia thành hai tập:

- Tập huấn luyện (80%): Dùng để huấn luyện mô hình Random Forest.
- Tập kiểm tra (20%): Dùng để đánh giá hiệu quả dự báo của mô hình.

Việc đồng nhất cấu trúc dữ liệu và quy trình xử lý cho cả hai mô hình LSTM và Random Forest giúp quá trình đánh giá trở nên minh bạch và khách quan, đảm bảo sự khác biệt về hiệu quả đến từ bản chất thuật toán thay vì khác biệt trong dữ liệu đầu vào.

3.5.2. Xây dựng mô hình Random Forest Regression

```
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train_rf, y_train_rf)
```

Mô hình được khởi tạo thông qua đối tượng RandomForestRegressor với tham số n_estimators=100, tức là mô hình sẽ bao gồm 100 cây quyết định (decision trees). Việc lựa chọn số lượng cây ở mức vừa phải giúp đảm bảo sự cân bằng giữa hiệu quả mô hình và chi phí tính toán. Ngoài ra, để đảm bảo tính tái lập của kết quả, random_state được thiết lập bằng 42 – đây là một bước cần thiết trong các thí nghiệm học máy nhằm tránh biến động ngẫu nhiên giữa các lần chạy. Sau khi được khởi tạo, mô hình được huấn luyện bằng phương thức fit() với tập dữ liệu đầu vào X_train_rf và nhãn mục tiêu tương ứng y_train_rf. Quá trình huấn luyện này cho phép Random Forest học được mối quan hệ phức tạp, có thể phi tuyến giữa các đặc trưng đầu vào và biến mục tiêu, thông qua cơ chế tổ hợp nhiều cây quyết định độc lập. Nhờ vào kỹ thuật bagging và cơ chế lấy mẫu ngẫu nhiên theo đặc trưng tại mỗi nút chia, Random Forest có khả năng giảm thiểu hiện tượng quá khớp (overfitting) và thường mang lại hiệu suất ổn định trên dữ liệu có nhiễu cao như dữ liệu tài chính. Đây chính là lý do mô hình được lựa chọn làm phương pháp đối sánh trong đề tài.

❖ Dự báo và đánh giá mô hình

```
predicted_rf_scaled = rf_model.predict(X_test_rf)
predicted_rf = scaler.inverse_transform(np.concatenate([np.zeros((len(predicted_rf_scaled), 1)), predicted_rf_scaled.reshape(-1, 1)], axis=1))[:, 1]
real_rf = scaler.inverse_transform(np.concatenate([np.zeros((len(y_test_rf), 1)), y_test_rf.reshape(-1, 1)], axis=1))[:, 1]
```

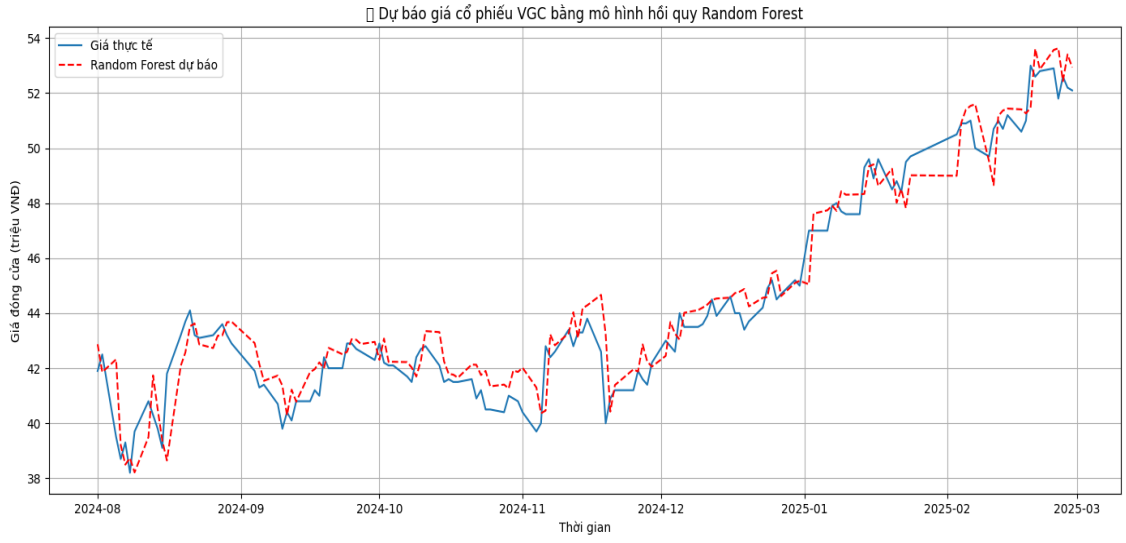
Sau khi hoàn tất quá trình huấn luyện, mô hình Random Forest được sử dụng để thực hiện dự đoán trên tập kiểm tra thông qua lệnh rf_model.predict(X_test_rf), kết quả đầu ra được lưu vào biến predicted_rf_scaled. Do mô hình đã được huấn luyện trên dữ liệu đã được chuẩn hóa (scaled), nên kết quả dự đoán cũng đang ở dạng tỷ lệ, chưa phản ánh đúng đơn vị giá trị thực tế. Để đưa kết quả về đúng ngữ cảnh ban đầu (giá cổ phiếu), quá trình chuẩn hóa ngược (inverse transform) được thực hiện bằng cách sử dụng phương thức inverse_transform của đối tượng scaler. Tuy nhiên, do scaler được huấn luyện với nhiều đặc trưng (features), trong khi mô hình chỉ trả về giá trị một chiều, ta cần tạo ra một mảng hai chiều giả lập có cùng số chiều với dữ liệu ban đầu. Việc này được xử lý thông qua hàm np.concatenate() kết hợp với np.zeros() để tạo ra các cột giá trị 0 làm nền, sau đó ghép nối với kết quả dự đoán đã reshape về dạng (n,1). Lệnh thứ hai trong đoạn mã thực hiện quá trình tương tự đối với biến y_test_rf nhằm chuẩn hóa ngược giá trị thực tế từ tập kiểm tra, lưu kết quả vào biến real_rf. Như vậy, cả hai mảng predicted_rf và real_rf đều đã được đưa về đơn vị gốc, sẵn sàng để so sánh, tính toán sai

số hoặc trực quan hóa nhằm đánh giá hiệu suất dự báo của mô hình. Việc chuẩn hóa ngược là bước đặc biệt quan trọng trong các bài toán hồi quy có dữ liệu được xử lý trước, giúp đảm bảo kết quả đánh giá mang tính thực tiễn và có thể diễn giải.

3.5.3. Kết quả dự báo

❖ Biểu đồ so sánh dự báo của Random Forest với giá thực tế

Hình 3.7. Biểu đồ so sánh dự báo của Random Forest với giá thực tế



(Nguồn: Tác giả tổng hợp và tính toán)

Biểu đồ trên thể hiện kết quả so sánh giữa giá thực tế và giá dự báo của cổ phiếu VGC trong giai đoạn từ tháng 8/2024 đến tháng 3/2025, sử dụng mô hình hồi quy Random Forest. Đường màu xanh biểu thị giá thực tế theo từng phiên, trong khi đường màu đỏ, nét đứt, biểu thị giá được dự báo bởi mô hình.

Quan sát tổng thể, có thể nhận thấy rằng mô hình Random Forest đã tái hiện được tương đối chính xác xu hướng biến động chính của giá cổ phiếu, đặc biệt trong các pha tăng đều hoặc giai đoạn thị trường ổn định. Cụ thể, từ tháng 11/2024 đến cuối kỳ dự báo, mô hình đã phản ánh rõ xu hướng tăng trưởng liên tục của giá, với độ lệch nhỏ và đường dự báo bám sát chặt chẽ đường giá thực tế.

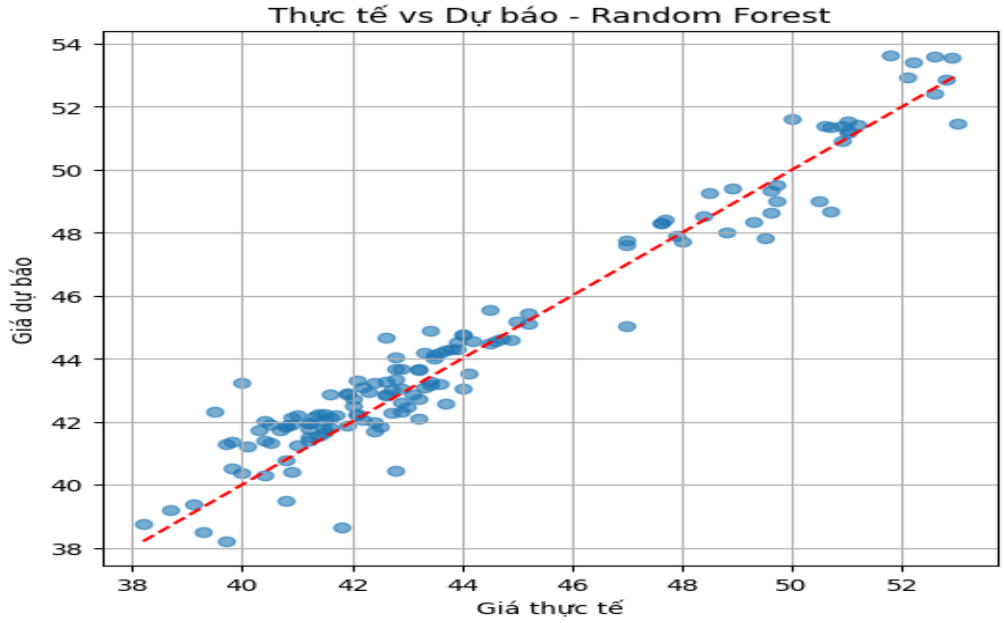
Tuy nhiên, trong những giai đoạn thị trường có biến động mạnh hoặc đảo chiều đột ngột, ví dụ khoảng cuối tháng 9 và đầu tháng 11/2024, mô hình có xu hướng phản ứng trễ hoặc bị sai lệch biên độ. Điều này được lý giải bởi bản chất của mô hình Random Forest là tổ hợp của các cây quyết định, vốn chia dữ liệu thành các vùng rời rạc. Do đó, mô hình thường có xu hướng làm mượt hoặc bậc thang hóa dự báo, dẫn đến việc bỏ qua một số biến động ngắn hạn có tính bất ngờ.

Ngoài ra, tại một số thời điểm, đặc biệt trong khoảng đầu năm 2025, có thể thấy mô hình tạo ra các đoạn "nấc" trong dự báo, phản ánh đặc trưng của phương pháp hồi quy bằng cây quyết định: đầu ra không liên tục mà thường thay đổi theo từng khoảng giá đã học. Mặc dù điều này không làm sai lệch xu hướng tổng thể, nhưng phần nào làm giảm tính mượt mà và trực quan của dự báo khi so sánh với các mô hình hồi tiếp như LSTM.

Tóm lại, biểu đồ cho thấy mô hình Random Forest có khả năng mô phỏng tốt xu hướng giá cổ phiếu, đặc biệt trong các giai đoạn có tính ổn định cao. Tuy nhiên, độ phản ứng còn hạn chế trong các phiên đảo chiều mạnh, và dự báo có tính bậc thang là đặc điểm cần được cân nhắc khi áp dụng trong thực tế. Mô hình này vẫn là lựa chọn phù hợp cho các bài toán yêu cầu tính đơn giản, tốc độ nhanh và khả năng xử lý dữ liệu không tuyến tính một cách hiệu quả.

❖ Biểu đồ tương quan giữa dự báo của Random Forest so với thực tế

Hình 3.8. Biểu đồ tương quan giữa dự báo Random Forest với thực tế



(Nguồn: Tác giả tổng hợp và tính toán)

Biểu đồ trên thể hiện mối quan hệ giữa giá thực tế và giá dự báo của cổ phiếu VGC từ mô hình Random Forest Regressor. Mỗi điểm trên đồ thị đại diện cho một cặp giá trị tại cùng một thời điểm trong tập kiểm tra. Đường chéo màu đỏ thể hiện đường lý tưởng (đường y = x), tức là khi dự báo hoàn toàn khớp với thực tế.

Có thể thấy rằng, phần lớn các điểm dữ liệu phân bố tập trung gần đường hồi quy lý tưởng, đặc biệt là trong khoảng giá từ 42.000 đến 52.000 đồng. Điều này cho thấy mô

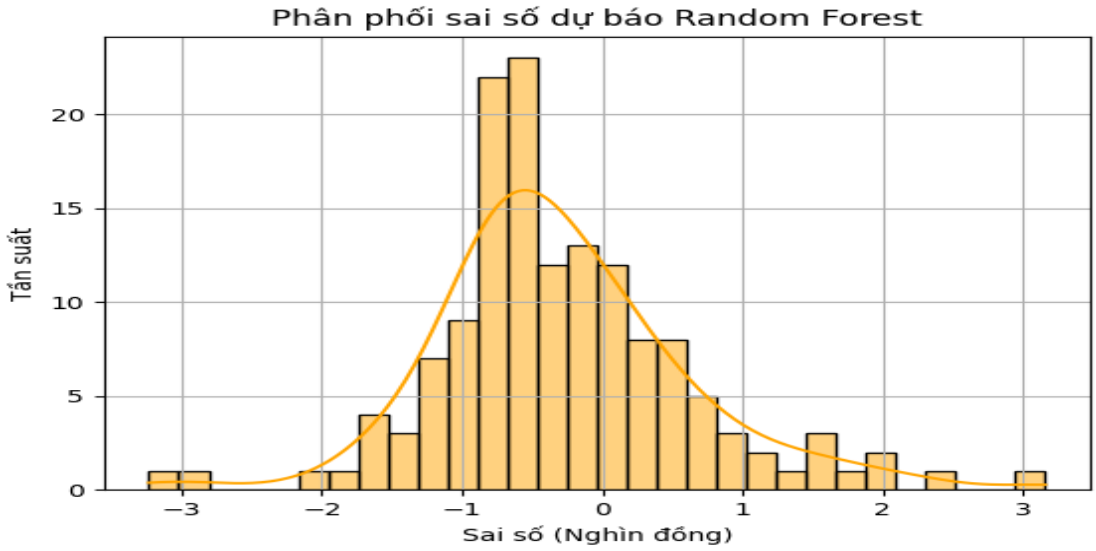
hình có khả năng tái hiện chính xác mối quan hệ tổng thể giữa dữ liệu đầu vào và giá cổ phiếu đầu ra. Độ phân tán quanh đường lý tưởng tương đối thấp, phản ánh độ chính xác dự báo ổn định và không có thiên lệch rõ rệt.

Đáng chú ý, so với LSTM, các điểm trong biểu đồ Random Forest có xu hướng rải đều hơn ở vùng giá thấp và trung bình, trong khi ở vùng giá cao, sai số có dấu hiệu tăng nhẹ với một số điểm vượt lên trên đường lý tưởng. Điều này cho thấy mô hình đôi khi có xu hướng dự báo giá cao hơn thực tế ở vùng đỉnh – một hiện tượng thường gặp với mô hình hồi quy tổ hợp trong bối cảnh dữ liệu có mức tăng nhanh.

Ngoài ra, không xuất hiện các cụm điểm nằm cách xa hoàn toàn khỏi vùng tập trung – điều này phản ánh rằng mô hình không tạo ra các dự báo bất thường hoặc ngoài phạm vi hợp lý. Tuy nhiên, tính tuyến tính của mối tương quan chưa hoàn toàn khớp tuyệt đối, phần nào cho thấy sự giới hạn của Random Forest khi mô hình hóa dữ liệu thời gian có độ trễ cao.

❖ Biểu đồ phân bố sai số dự báo của Random Forest

Hình 3.9. Biểu đồ phân bố sai số dự báo của Random Forest



(Nguồn: Tác giả tổng hợp và tính toán)

Biểu đồ trên thể hiện phân phối sai số giữa giá dự báo và giá thực tế của cổ phiếu VGC do mô hình Random Forest hồi quy tạo ra. Trục hoành là giá trị sai số (dự báo trừ thực tế, tính theo nghìn đồng), còn trục tung biểu thị tần suất xuất hiện của từng khoảng sai số. Đường cong màu cam thể hiện phân phối chuẩn ước lượng, giúp quan sát tính đối xứng và mức độ phân tán của sai số.

Tổng quan, phân phối sai số có xu hướng hội tụ mạnh quanh điểm 0, tức là sai số phần lớn tập trung ở mức rất nhỏ, cho thấy mô hình dự báo với độ chính xác tương đối cao. Tuy nhiên, không giống mô hình LSTM – vốn có phân phối sai số gần chuẩn hơn – phân bố trong biểu đồ này bị lệch nhẹ về bên trái. Điều này phản ánh rằng mô hình Random Forest có xu hướng dự báo thấp hơn thực tế trong một số trường hợp, đặc biệt là ở các mức giá trung bình.

Phần lớn sai số tập trung trong khoảng từ -1,5 đến +1 nghìn đồng, cho thấy các dự báo nằm trong phạm vi chấp nhận được với mức dao động không quá lớn. Những cột cao nhất gần điểm 0 thể hiện mật độ sai số nhỏ chiếm tỷ trọng cao, đây là dấu hiệu tích cực cho thấy mô hình hoạt động ổn định trên đa số phiên giao dịch.

Ngoài ra, cũng cần lưu ý rằng vẫn tồn tại một số giá trị ngoại biên (outliers) ở cả hai phía, với sai số có thể vượt quá ±3 nghìn đồng. Những điểm này thường xuất hiện trong các phiên giao dịch có biến động bất thường, như các cú đảo chiều đột ngột hoặc do ảnh hưởng của yếu tố vĩ mô mà mô hình không nhận biết được trong dữ liệu huấn luyện.

Tóm lại, biểu đồ phân phối sai số cho thấy mô hình Random Forest có độ lệch nhỏ, phân bố sai số chủ yếu tập trung quanh 0, đồng thời ít xuất hiện sai số lớn. Dù có độ trễ nhất định so với biến động thời gian thực, mô hình vẫn giữ được độ ổn định cao và tính dự báo đáng tin cậy, đặc biệt khi áp dụng trong các giai đoạn thị trường đi ngang hoặc xu hướng ổn định.

3.5.4. Đánh giá hiệu năng

Sau khi huấn luyện mô hình Random Forest Regression trên tập dữ liệu đã chuẩn hóa (với biến đầu vào là chuỗi giá đóng cửa được xử lý dưới dạng sliding window), kết quả kiểm tra hiệu suất trên tập dữ liệu kiểm định thu được như sau:

- R^2 (Hệ số xác định): 0.9374
- MAE (Sai số tuyệt đối trung bình): 0.7426
- RMSE (Sai số bình phương trung bình căn bậc hai): 0.9500
- MAPE (Sai số phần trăm tuyệt đối trung bình): 0.0170 (tương đương 1.70%)

Phân tích kết quả đánh giá:

Trước hết, hệ số R^2 đạt giá trị 0.9374 cho thấy mô hình có khả năng giải thích hơn 93% phương sai trong dữ liệu mục tiêu (giá cổ phiếu VGC) thông qua các biến đầu vào là các mức giá quá khứ. Điều này phản ánh mức độ khớp tương đối tốt giữa mô hình và

dữ liệu thực tế, đồng thời khẳng định rằng mô hình có thể nắm bắt phần lớn thông tin cấu trúc trong chuỗi dữ liệu tài chính đã được định hình theo thời gian.

Chỉ số MAE = 0.7426 và RMSE = 0.9500, đều ở mức thấp, cho thấy sai số dự báo trung bình thấp, tức là các dự đoán gần sát với giá trị thực tế. Đặc biệt, RMSE lớn hơn MAE cho thấy mô hình có một mức độ nhạy nhất định với các sai số lớn (outliers), nhưng mức chênh lệch không quá lớn, phản ánh tính ổn định trong dự báo.

Chỉ số MAPE đạt 1.70%, tức là sai số dự báo trung bình chỉ chiếm chưa đến 2% giá trị thực tế, là mức rất thấp trong bối cảnh dữ liệu thị trường tài chính vốn có đặc tính biến động và nhiều nhiễu. Điều này cho thấy mô hình có khả năng dự báo hiệu quả cả về tuyệt đối lẫn tương đối, góp phần khẳng định tính ứng dụng thực tiễn của Random Forest trong bài toán dự báo giá cổ phiếu.

Một ưu điểm lớn của Random Forest là khả năng xử lý tốt các mối quan hệ phi tuyến, phân phối dữ liệu không chuẩn và tính chất dị biệt trong thị trường chứng khoán. Với cấu trúc bao gồm nhiều cây quyết định hoạt động song song, mô hình có khả năng hội tụ và khái quát hóa mạnh mẽ, đồng thời giảm thiểu rủi ro quá khớp nhờ cơ chế bagging và lấy mẫu ngẫu nhiên.

Tuy nhiên, Random Forest không khai thác được cấu trúc chuỗi thời gian một cách nội tại, bởi bản chất của nó là mô hình học có giám sát phi tuần tự. Do đó, nó không tự động học được xu hướng, mùa vụ hay độ trễ – các yếu tố quan trọng trong dữ liệu tài chính. Trong đề tài này, vấn đề được khắc phục phần nào thông qua việc tái cấu trúc dữ liệu đầu vào theo kỹ thuật sliding window, giúp cung cấp cho mô hình ngữ cảnh thời gian từ quá khứ. Tuy nhiên, cách tiếp cận này vẫn mang tính thủ công và không tối ưu như các mô hình mạng nơ-ron hồi tiếp (RNN) hoặc LSTM – vốn được thiết kế chuyên biệt cho chuỗi thời gian.

3.6. So sánh hiệu quả hai mô hình

Bảng 3.1. Hiệu năng 2 mô hình

Mô hình	R^2	MAE	RMSE	MAPE (%)
Random Forest	0,9374	0.7426	0.9500	1.70%
LSTM	0,9384	0.6920	0.9417	1.58%

(nguồn : Tác giả tổng hợp và tính toán)

Dựa trên bảng trên, có thể thấy rằng mô hình LSTM vượt trội nhẹ so với Random Forest ở cả 4 chỉ số. Cụ thể, chỉ số MAE của LSTM thấp hơn Random Forest khoảng 0.0506 nghìn đồng, cho thấy sai số trung bình của LSTM nhỏ hơn trong quá trình dự báo. Tương tự, chỉ số RMSE – vốn nhấn mạnh các sai số lớn – của LSTM cũng thấp hơn, phản ánh rằng mô hình này kiểm soát sai số tốt hơn, đặc biệt ở các phiên có biến động bất thường. Đặc biệt, MAPE của LSTM đạt mức 1.58%, thấp hơn mức 1.70% của Random Forest, cho thấy LSTM dự báo chính xác hơn về mặt phần trăm sai lệch so với thực tế. R^2 (Hệ số xác định)**:** Cả hai mô hình đều cho giá trị R^2 rất cao (>0.93), chứng tỏ khả năng giải thích biến động của giá cổ phiếu là tương đối tốt. LSTM nhỉnh hơn một chút với R^2 = 0.9384 so với Random Forest (R^2 = 0.9374), cho thấy LSTM có khả năng bắt được xu hướng chuỗi thời gian tốt hơn một chút.

Xét về bản chất mô hình, LSTM là mạng nơ-ron hồi tiếp có khả năng ghi nhớ chuỗi thời gian, nên có ưu thế rõ rệt trong việc nắm bắt các mối quan hệ dài hạn, chu kỳ và xu hướng nối tiếp trong dữ liệu tài chính. Điều này giúp LSTM phản ứng linh hoạt hơn trong các giai đoạn có biến động liên tục hoặc chuyển hướng nhanh chóng. Trong khi đó, Random Forest – với bản chất là một mô hình học máy tổ hợp không phụ thuộc thời gian – lại hoạt động ổn định và dễ triển khai, nhưng phần nào kém hiệu quả hơn trong việc học chuỗi liên tục và thường phản ứng trễ với các chuyển động ngắn hạn của thị trường.

Tuy nhiên, cũng cần lưu ý rằng khoảng cách giữa hai mô hình là không quá lớn. Mô hình Random Forest vẫn thể hiện khả năng dự báo đáng tin cậy, và đặc biệt thích hợp trong các điều kiện thị trường ổn định hoặc cho các ứng dụng yêu cầu đơn giản, dễ huấn luyện và ít chi phí tính toán.

Tóm lại, trong bối cảnh nghiên cứu này, có thể khẳng định rằng LSTM là mô hình dự báo tối ưu hơn, khi xét đến độ chính xác theo cả ba tiêu chí MAE, RMSE và MAPE. Đây là lựa chọn phù hợp cho các bài toán chuỗi thời gian tài chính có yêu cầu cao về độ

Kết luận chương 3:

Chương 3 đã trình bày chi tiết quy trình thực nghiệm dự báo giá cổ phiếu VGC bằng hai mô hình: Random Forest Regression và LSTM. Dữ liệu đầu vào được xử lý và khai phá nhằm xây dựng các đặc trưng kỹ thuật phù hợp cho bài toán chuỗi thời gian. Kết quả cho thấy mô hình LSTM có khả năng nắm bắt tốt xu hướng dài hạn nhờ kiến trúc ghi nhớ, trong khi Random Forest hoạt động ổn định và hiệu quả với dữ liệu không

chuẩn hóa. Cả hai mô hình đều được đánh giá bằng MAE, RMSE, MAPE và R^2, giúp so sánh trực quan hiệu năng. Từ đó, có thể thấy rằng mỗi mô hình đều có thế mạnh riêng, phụ thuộc vào đặc điểm dữ liệu và mục tiêu dự báo cụ thể. Những phát hiện trong chương này là cơ sở để đưa ra kết luận tổng thể và đề xuất ứng dụng thực tiễn trong chương tiếp theo.

CHƯƠNG 4: KẾT LUẬN VÀ GIẢI PHÁP

4.1. Kết luận kết quả nghiên cứu

4.1.1. Đánh giái hiệu quả mô hình dự báo

Thông qua quá trình thực nghiệm, nghiên cứu đã triển khai và đánh giá hai mô hình dự báo thuộc hai nhóm phương pháp tiếp cận khác nhau: mô hình học máy Random Forest Regression và mô hình học sâu LSTM. Đây là hai thuật toán có sự khác biệt đáng kể về nguyên lý kỹ thuật, từ cách xây dựng dữ liệu đầu vào cho tới cơ chế học và cập nhật tham số, nhờ đó cung cấp những ưu thế riêng biệt trong bài toán dự báo chuỗi thời gian tài chính.

Random Forest Regression hoạt động dựa trên việc xây dựng tập hợp nhiều cây quyết định, từ đó trung bình hóa kết quả dự báo nhằm giảm phương sai và cải thiện độ chính xác. Trong khi đó, mô hình LSTM - đại diện của nhóm mạng nơ-ron hồi tiếp sâu - lại được thiết kế đặc biệt để xử lý chuỗi thời gian phụ thuộc dài hạn. Thông qua ba cổng điều tiết thông tin (gồm cổng quên, cổng đầu vào và cổng đầu ra), LSTM có khả năng ghi nhớ các mối liên hệ phức tạp giữa các bước thời gian cách xa nhau, từ đó mô hình hóa hiệu quả xu hướng giá cổ phiếu theo chuỗi biến động. Điểm mạnh nổi bật của LSTM là khả năng tự động học các mẫu hình phức tạp trong dữ liệu tài chính mà không cần quá trình thủ công thiết kế biến đầu vào.

Về mặt độ chính xác, kết quả thực nghiệm cho thấy mô hình LSTM đạt sai số thấp hơn trên cả các chỉ số MAE, RMSE và MAPE so với Random Forest. Điều này phản ánh LSTM có khả năng bám sát xu hướng thực tế tốt hơn, đặc biệt ở những thời kỳ thị trường xuất hiện các biến động bất ngờ. Tuy nhiên, đổi lại LSTM yêu cầu thời gian huấn luyện dài hơn đáng kể, khâu xử lý dữ liệu phức tạp hơn và cần cấu hình phần cứng mạnh hơn, vốn có thể gây cản trở khi triển khai thực tế tại nhiều doanh nghiệp nhỏ.

4.1.2. Đánh giá sai số, ưu nhược điểm và khả năng ứng dụng thực tiễn

Khi xét đến độ chính xác tuyệt đối**,** mô hình LSTM cho kết quả tốt hơn với giá trị MAE và RMSE thấp hơn, cho thấy khả năng mô hình nắm bắt tốt hơn các đặc trưng biến động giá. Ngoài ra, MAPE – chỉ số phản ánh sai số tương đối – cũng ở mức thấp hơn ở mô hình LSTM, điều này đặc biệt hữu ích trong việc đánh giá hiệu năng trên các biến động có quy mô khác nhau.

Tuy nhiên, Random Forest Regression lại có ưu điểm về tốc độ huấn luyện nhanh, dễ triển khai, và ít đòi hỏi xử lý phức tạp về chuẩn hóa hay định hình dữ liệu như LSTM. Hơn nữa, mô hình này không cần nhiều tài nguyên tính toán, điều này khiến nó trở nên phù hợp hơn trong bối cảnh các doanh nghiệp chưa có điều kiện tiếp cận các hạ tầng tính toán mạnh.

Về khả năng ứng dụng thực tiễn, Việc xây dựng mô hình dự báo giá cổ phiếu bằng học máy và học sâu có thể hỗ trợ nhà đầu tư ra quyết định trên nhiều phương diện:

Thứ nhất, mô hình giúp nhà đầu tư nhận diện sớm xu hướng giá cổ phiếu. Thay vì chỉ dựa vào phân tích cảm tính hoặc các chỉ báo kỹ thuật truyền thống, mô hình dự báo có thể cung cấp các tín hiệu cảnh báo sớm về khả năng đảo chiều, xu hướng tăng hoặc giảm giá trong tương lai gần, từ đó giúp nhà đầu tư chủ động lựa chọn thời điểm mua vào hoặc bán ra tối ưu hơn.

Thứ hai, mô hình hỗ trợ kiểm tra và xây dựng các chiến lược giao dịch định lượng (quantitative strategies). Nhà đầu tư có thể sử dụng các mô hình đã huấn luyện để backtest (kiểm định lại) hiệu quả các chiến lược giao dịch trên dữ liệu lịch sử, từ đó điều chỉnh chiến thuật phù hợp hơn với mục tiêu lợi nhuận và mức độ chấp nhận rủi ro của bản thân.

Thứ ba, hệ thống dự báo khi kết hợp với các nền tảng phân tích dữ liệu tài chính có thể trở thành công cụ ra quyết định trong thời gian thực. Ví dụ: với các tổ chức tài chính, mô hình có thể cảnh báo nhanh về biến động bất thường, giúp bộ phận quản lý danh mục điều chỉnh tỷ trọng cổ phiếu phù hợp theo biến động thị trường.

Thứ tư, mô hình cũng giúp giảm thiểu yếu tố cảm tính trong quá trình đầu tư. Thông qua việc dựa vào kết quả dự báo định lượng, nhà đầu tư có thể tránh được những quyết định mang tính chủ quan, cảm xúc, vốn là nguyên nhân phổ biến dẫn đến thua lỗ trên thị trường chứng khoán.

Cuối cùng, mô hình có thể tích hợp vào các hệ thống tư vấn tự động (robo-advisor) dành cho nhà đầu tư cá nhân, giúp ngay cả những nhà đầu tư thiếu kinh nghiệm cũng có thể tiếp cận công cụ hỗ trợ phân tích mang tính chuyên sâu mà trước đây chỉ dành cho các tổ chức lớn.

4.1.3. So sánh với các nghiên cứu trước

Kết quả đạt được trong khuôn khổ đề tài nhìn chung phù hợp với các nghiên cứu trước đây về ứng dụng mô hình học sâu và học máy trong lĩnh vực dự báo tài chính. Cụ

thể, nghiên cứu của Zhang et al. (2021) chỉ ra rằng mô hình LSTM thường vượt trội về độ chính xác so với các phương pháp học máy truyền thống khi áp dụng vào dữ liệu chứng khoán có đặc điểm biến động mạnh, phi tuyến và không ổn định. Năng lực ghi nhớ dài hạn và khả năng mô hình hóa quan hệ chuỗi phức tạp giúp LSTM thích nghi tốt với môi trường thị trường nhiều biến động, điều mà các mô hình như Random Forest khó thực hiện nếu không có biến đầu vào chất lượng cao.

Tuy nhiên, ở chiều ngược lại, các nghiên cứu như của Patel et al. (2015) và Fischer & Krauss (2018) cũng chỉ ra rằng Random Forest vẫn là một lựa chọn hiệu quả trong các môi trường thị trường có độ ổn định tương đối cao, với dữ liệu ít nhiễu và biến động thấp. Điểm mạnh của Random Forest nằm ở khả năng kháng nhiễu, không yêu cầu xử lý chuẩn hóa dữ liệu và tính dễ triển khai trên quy mô lớn, đặc biệt phù hợp với các hệ thống phân tích kỹ thuật nội bộ hoặc ứng dụng ra quyết định nhanh.

Do đó, kết quả thực nghiệm trong đề tài không chỉ củng cố các kết luận đã có trong các tài liệu nghiên cứu trước mà còn minh chứng rằng việc lựa chọn mô hình cần phải linh hoạt, dựa trên tính chất dữ liệu và mục tiêu ứng dụng cụ thể thay vì tuyệt đối hóa bất kỳ phương pháp nào.

4.1.3. Đóng góp học thuật

Từ góc độ nghiên cứu, đề tài đã góp phần làm rõ hiệu quả của việc áp dụng các mô hình dự báo hiện đại – cụ thể là Random Forest và LSTM – trong ngữ cảnh của một thị trường mới nổi như Việt Nam. Thị trường chứng khoán Việt Nam đặc trưng bởi sự biến động nhanh, tính thanh khoản chưa đồng đều giữa các nhóm cổ phiếu và dữ liệu lịch sử còn mang tính phân mảnh. Việc chứng minh khả năng vận hành hiệu quả của cả hai mô hình trong điều kiện dữ liệu như vậy là một đóng góp có ý nghĩa không nhỏ đối với cộng đồng nghiên cứu trong và ngoài nước.

Điểm nổi bật của đề tài là không chỉ dừng lại ở việc ứng dụng từng mô hình riêng lẻ, mà còn đề xuất khả năng kết hợp hai hướng tiếp cận – học máy truyền thống và học sâu hiện đại – vào cùng một quy trình dự báo, nhằm tận dụng điểm mạnh của từng mô hình để bù đắp điểm yếu của nhau. Trong khi phần lớn các nghiên cứu về học sâu trong tài chính hiện nay vẫn đang tập trung ở các thị trường phát triển (như Mỹ, EU hoặc Nhật Bản), nơi dữ liệu dồi dào và hệ sinh thái tài chính số hóa cao, thì đề tài này góp phần chứng minh rằng các kỹ thuật tiên tiến hoàn toàn có thể được điều chỉnh và triển khai hiệu quả tại các thị trường đang phát triển nếu được thiết kế đúng hướng.

Qua đó, đề tài không chỉ làm giàu thêm kho tàng nghiên cứu ứng dụng trí tuệ nhân tạo vào tài chính mà còn mở ra tiền đề cho các nghiên cứu tiếp theo trong bối cảnh địa phương, nơi mà nhu cầu hiện đại hóa phân tích dữ liệu tài chính ngày càng cấp thiết nhưng nguồn lực triển khai vẫn còn hạn chế.

4.2. Đề xuất giải pháp

Dựa trên toàn bộ quá trình triển khai, thử nghiệm và phân tích mô hình trong chương trước, luận văn xin đề xuất một số giải pháp cụ thể nhằm nâng cao hiệu quả ứng dụng của mô hình dự báo giá cổ phiếu VGC trong thực tế cũng như mở rộng nghiên cứu trong tương lai.

4.2.1. Đối với doanh nghiệp và nhà đầu tư

Trong bối cảnh thị trường tài chính ngày càng biến động mạnh và đòi hỏi quyết định đầu tư kịp thời, việc áp dụng các mô hình dự báo định lượng trở nên đặc biệt quan trọng. Từ kết quả nghiên cứu, luận văn đề xuất một số giải pháp cụ thể như sau:

❖ Ứng dụng mô hình LSTM trong phân tích đầu tư chủ động

Mô hình LSTM, với khả năng ghi nhớ và xử lý chuỗi thời gian có độ dài lớn, đặc biệt phù hợp để dự báo những xu hướng phức tạp, mang tính chu kỳ hoặc bất ngờ – thường gặp trong dữ liệu giá cổ phiếu. Do đó, doanh nghiệp hoặc nhà đầu tư chuyên nghiệp có thể sử dụng LSTM làm công cụ hỗ trợ trong:

- Dự báo điểm đảo chiều ngắn hạn để tối ưu hóa điểm mua – bán.
- Phân tích hành vi giá trong các giai đoạn có nhiều thông tin tác động như mùa báo cáo tài chính, chính sách lãi suất, hoặc biến động vĩ mô.
- Thiết kế chiến lược giao dịch định lượng (quantitative strategy) như trend-following, momentum hoặc các chiến lược kiểm định lại mô hình (backtest) trên dữ liệu lịch sử.

Tuy nhiên, để triển khai thành công mô hình này trong thực tế, doanh nghiệp cần đầu tư vào hạ tầng dữ liệu, hệ thống tính toán GPU và đội ngũ có năng lực triển khai học sâu, vì LSTM đòi hỏi xử lý đầu vào chuẩn hóa, cấu trúc 3 chiều và thời gian huấn luyện dài hơn đáng kể so với các mô hình học máy truyền thống.

❖ Tận dụng mô hình Random Forest như công cụ cảnh báo xu hướng

Ngược lại, mô hình Random Forest với đặc tính đơn giản, dễ huấn luyện và độ ổn định cao, rất phù hợp để ứng dụng trong các hệ thống cảnh báo sớm. Một số ứng dụng có thể kể đến gồm:

- Tích hợp vào hệ thống dashboard phân tích nhanh xu hướng của nhóm cổ phiếu, nhằm đưa ra tín hiệu giao dịch sơ bộ.
- Phục vụ các nền tảng tư vấn đầu tư cho khách hàng cá nhân ở mức độ phổ thông (retail investors), nơi yêu cầu phản hồi nhanh và dễ giải thích.
- Ứng dụng trong các công cụ phân tích nội bộ doanh nghiệp, nhằm dự báo sơ bộ xu hướng giá cổ phiếu công ty mình hoặc ngành hàng liên quan.

Lợi thế của Random Forest là tính dễ triển khai, không yêu cầu chuẩn hóa, và thời gian huấn luyện ngắn – rất thích hợp cho các hệ thống cần cập nhật thường xuyên nhưng không có điều kiện tính toán lớn.

❖ Kết hợp hai mô hình trong một kiến trúc lai

Một hướng đi tiềm năng là xây dựng hệ thống hybrid (mô hình lai), trong đó sử dụng Random Forest để lọc tín hiệu xu hướng sơ bộ, sau đó dùng LSTM để dự báo chi tiết về mức giá. Mô hình lai có thể giúp tăng độ tin cậy trong dự báo thông qua việc:

- Giảm sai số ngẫu nhiên (noise) trong dự báo một mô hình đơn lẻ.
- Tận dụng đồng thời tính ổn định của mô hình học máy và tính linh hoạt của mô hình học sâu.
- Áp dụng voting ensemble hoặc stacking để kết hợp đầu ra từ nhiều mô hình, từ đó tối ưu hóa hiệu suất tổng thể.

4.2.2. Đối với các nghiên cứu tiếp theo

Ngoài việc ứng dụng thực tiễn, đề tài cũng mở ra các hướng phát triển nghiên cứu chuyên sâu hơn, đặc biệt trong bối cảnh dữ liệu tài chính ngày càng đa dạng và biến động phức tạp:

❖ Mở rộng nguồn và loại dữ liệu

Hiện tại, dữ liệu sử dụng chủ yếu là giá và khối lượng giao dịch của cổ phiếu. Trong tương lai, nên mở rộng thêm các yếu tố đầu vào như:

- Biến vĩ mô: lạm phát (CPI), lãi suất, tỷ giá USD/VND, giá hàng hóa (như thép, xi măng), vì đây là các yếu tố ảnh hưởng trực tiếp đến ngành nghề và kết quả kinh doanh của VGC.

- Biến định tính: dữ liệu tin tức tài chính, mạng xã hội (sentiment analysis), báo cáo phân tích doanh nghiệp.
- Biến nội bộ: số liệu tài chính theo quý như doanh thu, lợi nhuận, EPS nếu kết hợp thêm phân tích cơ bản.

Việc kết hợp đa dạng nguồn dữ liệu sẽ giúp mô hình học được bối cảnh tổng thể của thị trường thay vì chỉ dựa vào giá cổ phiếu đơn lẻ.

❖ Áp dụng các mô hình học sâu hiện đại hơn

Ngoài LSTM, các mô hình mới như GRU, Temporal Convolutional Network (TCN) hoặc Transformer đã chứng minh hiệu quả vượt trội trong nhiều bài toán dự báo chuỗi thời gian. Việc thử nghiệm các mô hình này, đặc biệt là Transformer (vốn nổi bật trong NLP) nhưng đang ngày càng được ứng dụng trong tài chính (Financial Time Series Transformer), là hướng nghiên cứu rất đáng quan tâm.

❖ Tối ưu hóa mô hình và hệ thống

Tối ưu siêu tham số (hyperparameter tuning) bằng Grid Search, Random Search hoặc Bayesian Optimization để tìm cấu hình hiệu quả nhất.

Feature selection và giảm chiều dữ liệu bằng PCA hoặc autoencoder để giảm nhiễu và rút gọn mô hình.

Xây dựng hệ thống cập nhật mô hình động (online learning hoặc tái huấn luyện định kỳ) nhằm thích nghi với thay đổi của thị trường.

❖ Phát triển hệ thống dự báo thời gian thực

Một hướng ứng dụng tiềm năng là xây dựng hệ thống real-time forecasting với: Lịch trình cập nhật dữ liệu hàng ngày hoặc hàng giờ. Mô hình được tự động tái huấn luyện định kỳ (weekly/monthly).Tích hợp hiển thị trực quan qua dashboard web-based (dùng Streamlit, Power BI, v.v.)

Hệ thống này có thể phục vụ cho bộ phận phân tích doanh nghiệp, công ty chứng khoán, hoặc nhà đầu tư tổ chức cần phân tích cập nhật và phản hồi nhanh.

4.2.3. Đề xuất hướng ứng dụng cấp hệ thống

Về dài hạn, một trong những hướng đi có giá trị ứng dụng cao là xây dựng các hệ thống hỗ trợ ra quyết định (Decision Support System – DSS) dựa trên nền tảng các mô hình học máy và học sâu đã được trình bày trong đề tài. Thay vì sử dụng các mô hình một cách riêng lẻ và rời rạc, các tổ chức tài chính có thể tích hợp chúng vào một nền

tảng phân tích tổng hợp, giúp xử lý và phân tích khối lượng lớn dữ liệu thị trường trong thời gian thực. Hệ thống DSS này có thể hoạt động theo kiến trúc đa lớp, trong đó:

- Tầng dữ liệu đầu vào bao gồm các nguồn như dữ liệu giá chứng khoán thời gian thực, dữ liệu kinh tế vĩ mô, báo cáo tài chính doanh nghiệp, tin tức và dữ liệu mạng xã hội.
- Tầng xử lý mô hình sẽ triển khai đồng thời các mô hình học máy như Random Forest để lọc tín hiệu xu hướng sơ bộ, và mô hình học sâu như LSTM hoặc Transformer để đưa ra dự báo chi tiết hơn.
- Tầng ứng dụng ra quyết định sẽ tổng hợp kết quả mô hình, đối chiếu với ngưỡng rủi ro và mục tiêu lợi nhuận đã đặt trước để đưa ra các cảnh báo thị trường, đánh giá xác suất đảo chiều giá, và gợi ý chiến lược đầu tư tối ưu.

Ưu điểm nổi bật của mô hình DSS là khả năng hoạt động liên tục, phản ứng kịp thời với biến động thị trường, đồng thời giảm thiểu rủi ro cảm tính trong ra quyết định của nhà đầu tư. Nếu được triển khai thành công, hệ thống có thể đóng vai trò như một "trợ lý phân tích tài chính" tự động, hỗ trợ các bộ phận như phân tích đầu tư, quản lý danh mục, kiểm soát rủi ro trong doanh nghiệp.

Trong bối cảnh chuyển đổi số đang là xu thế tất yếu tại các tổ chức tài chính – ngân hàng, việc tích hợp các mô hình học máy vào hệ thống DSS không chỉ mang tính cấp tiến mà còn là một chiến lược dài hạn giúp nâng cao năng lực cạnh tranh và chất lượng ra quyết định của doanh nghiệp.

PHỤ LỤC

Kết quả hoạt động kinh doanh VGC 2022 – 2024

	2022	**2023**	**2024**
1. Doanh thu bán hàng và cung cấp dịch vụ	14,607,943,556,288	13,342,467,325,243	12,051,482,639,966
2. Các khoản giảm trừ doanh thu	15,493,694,311	148,648,897,197	145,126,698,439
3. Doanh thu thuần về bán hàng và cung cấp dịch vụ (10 = 01 - 02)	14,592,449,861,977	13,193,818,428,046	11,906,355,941,527
4. Giá vốn hàng bán	10,354,300,437,633	9,674,692,360,146	8,389,049,269,758
5. Lợi nhuận gộp về bán hàng và cung cấp dịch vụ(20=10-11)	4,238,149,424,344	3,519,126,067,900	3,517,306,671,769
6. Doanh thu hoạt động tài chính	85,615,395,038	60,529,622,370	75,803,932,341
7. Chi phí tài chính	324,403,037,382	380,885,840,494	310,363,905,667
- Trong đó: Chi phí lãi vay	251,376,723,757	348,457,380,117	268,896,807,699
8. Phần lãi lỗ trong công ty liên doanh, liên kết	112,409,447,259	-36,392,135,098	-74,769,120,822
9. Chi phí bán hàng	936,334,436,050	812,377,184,581	861,838,342,910
10. Chi phí quản lý doanh nghiệp	911,832,018,223	756,017,100,394	744,493,344,308
11. Lợi nhuận thuần từ hoạt động kinh doanh{30=20+(21-22) + 24 - (25+26)}	2,263,604,774,986	1,593,983,429,703	1,601,645,890,403
12. Thu nhập khác	83,630,368,167	72,067,424,203	123,961,456,138
13. Chi phí khác	42,030,991,056	64,112,316,489	95,281,696,431
14. Lợi nhuận khác(40=31-32)	41,599,377,111	7,955,107,714	28,679,759,707
15. Tổng lợi nhuận kế toán trước thuế(50=30+40)	2,305,204,152,097	1,601,938,537,417	1,630,325,650,110
16. Chi phí thuế TNDN hiện hành	381,840,210,013	425,202,407,709	411,948,856,891
17. Chi phí thuế TNDN hoãn lại	10,325,407,122	14,496,988,030	30,758,200,639
18. Lợi nhuận sau thuế thu nhập doanh nghiệp(60=50-51-52)	1,913,038,534,962	1,162,239,141,678	1,187,618,592,580
19. Lợi nhuận sau thuế công ty mẹ	1,728,187,379,363	1,218,120,252,933	1,104,734,866,668
20. Lợi nhuận sau thuế công ty mẹ không kiểm soát	184,851,155,599	-55,881,111,255	82,883,725,912
21. Lãi cơ bản trên cổ phiếu(*)	3,854	2,717	2,464
22. Lãi suy giảm trên cổ phiếu (*)			

Bảng cân đối kế toán của VGC 2022 - 2024

TÀI SẢN	**2022**	**2023**	**2024**
A- TÀI SẢN NGẮN HẠN	8,107,975,056,610	9,104,809,897,620	9,464,267,034,186

I. Tiền và các khoản tương đương tiền	2,018,744,609,826	1,841,653,234,658	2,860,122,610,379
1. Tiền	937,207,375,606	1,142,029,494,839	942,600,718,030
2. Các khoản tương đương tiền	1,081,537,234,220	699,623,739,819	1,917,521,892,349
II. Các khoản đầu tư tài chính ngắn hạn	128,954,942,982	626,586,849,988	433,382,669,069
1. Chứng khoán kinh doanh			
2. Dự phòng giảm giá chứng khoán kinh doanh			
3. Đầu tư nắm giữ đến ngày đáo hạn	128,954,942,982	626,586,849,988	433,382,669,069
III. Các khoản phải thu ngắn hạn	1,183,294,409,897	1,117,328,239,030	1,080,575,112,080
1. Phải thu ngắn hạn của khách hàng	891,078,346,611	936,463,536,139	918,046,480,694
2. Trả trước cho người bán ngắn hạn	337,132,020,703	241,225,800,207	187,562,919,086
3. Phải thu nội bộ ngắn hạn			
4. Phải thu theo tiến độ kế hoạch hợp đồng xây dựng			
5. Phải thu về cho vay ngắn hạn	800,000,000	500,000,000	450,000,000
6. Phải thu ngắn hạn khác	277,038,242,379	225,321,705,141	274,318,526,262
7. Dự phòng phải thu ngắn hạn khó đòi	-322,754,199,796	-286,182,802,457	-299,802,813,962
8. Tài sản Thiếu chờ xử lý			
IV. Hàng tồn kho	4,235,047,120,510	4,739,829,320,287	4,375,950,687,848
1. Hàng tồn kho	4,376,027,375,202	4,964,073,996,726	4,500,170,570,634
2. Dự phòng giảm giá hàng tồn kho	-140,980,254,692	-224,244,676,439	-124,219,882,786
V.Tài sản ngắn hạn khác	541,933,973,395	779,412,253,657	714,235,954,810
1. Chi phí trả trước ngắn hạn	54,356,570,033	42,989,554,925	29,729,029,806
2. Thuế GTGT được khấu trừ	452,348,183,116	662,315,510,522	654,866,284,569
3. Thuế và các khoản khác phải thu Nhà nước	35,229,220,246		29,640,640,435
4. Giao dịch mua bán lại trái phiếu Chính phủ		74,107,188,210	
5. Tài sản ngắn hạn khác			
B. TÀI SẢN DÀI HẠN	14,850,946,352,686	14,995,380,193,666	15,363,151,893,952
I. Các khoản phải thu dài hạn	303,779,116,670	255,066,099,860	278,782,895,058
1. Phải thu dài hạn của khách hàng			4,266,810,286
2. Trả trước cho người bán dài hạn			

3. Vốn kinh doanh ở đơn vị trực thuộc			
4. Phải thu nội bộ dài hạn			
5. Phải thu về cho vay dài hạn			
6. Phải thu dài hạn khác	303,779,116,670	255,066,099,860	274,516,084,772
7. Dự phòng phải thu dài hạn khó đòi			
II.Tài sản cố định	5,383,244,682,733	5,385,365,380,110	6,020,629,770,983
1. Tài sản cố định hữu hình	5,003,312,336,780	4,977,038,486,206	5,643,555,734,279
- Nguyên giá	11,493,993,975,689	11,940,379,802,597	13,272,070,291,384
- Giá trị hao mòn lũy kế	-6,490,681,638,909	-6,963,341,316,391	-7,628,514,557,105
2. Tài sản cố định thuê tài chính	214,034,391,200	247,828,004,647	216,580,018,705
- Nguyên giá	289,975,784,215	349,264,056,993	335,589,133,907
- Giá trị hao mòn lũy kế	-75,941,393,015	-101,436,052,346	-119,009,115,202
3. Tài sản cố định vô hình	165,897,954,753	160,498,889,257	160,494,017,999
- Nguyên giá	219,107,449,512	219,105,789,512	224,529,293,460
- Giá trị hao mòn lũy kế	-53,209,494,759	-58,606,900,255	-64,035,275,461
III. Bất động sản đầu tư	1,951,881,365,444	1,942,422,317,951	1,914,237,254,178
- Nguyên giá	9,540,047,077,919	11,885,872,067,635	13,123,170,380,338
- Giá trị hao mòn lũy kế	-7,588,165,712,475	-9,943,449,749,684	-11,208,933,126,160
IV. Tài sản dở dang dài hạn	5,774,841,992,938	6,229,377,004,740	6,093,932,875,606
1. Chi phí sản xuất, kinh doanh dở dang dài hạn			
2. Chi phí xây dựng cơ bản dở dang	5,774,841,992,938	6,229,377,004,740	6,093,932,875,606
V. Đầu tư tài chính dài hạn	688,507,845,751	438,307,587,498	365,094,718,370
1. Đầu tư vào công ty con			
2. Đầu tư vào công ty liên kết, liên doanh	680,287,553,610	430,086,118,936	356,873,249,808
3. Đầu tư góp vốn vào đơn vị khác	9,332,682,344	9,332,682,344	9,332,682,344
4. Dự phòng đầu tư tài chính dài hạn	-1,214,690,203	-1,213,513,782	-1,213,513,782
5. Đầu tư nắm giữ đến ngày đáo hạn	102,300,000	102,300,000	102,300,000
VI. Tài sản dài hạn khác	748,691,349,150	744,841,803,507	690,474,379,757
1. Chi phí trả trước dài hạn	734,707,382,502	729,258,927,036	678,851,868,203
2. Tài sản thuế thu nhập hoãn lại	13,983,966,648	15,582,876,471	11,622,511,554
3. Thiết bị, vật tư, phụ tùng thay thế dài hạn			
4. Tài sản dài hạn khác			
5. Lợi thế thương mại			

TỔNG CỘNG TÀI SẢN	22,958,921,409,296	24,100,190,091,286	24,827,418,928,138
NGUỒN VỐN			
C. NỢ PHẢI TRẢ	13,873,492,333,128	14,575,872,174,590	14,874,419,272,735
I. Nợ ngắn hạn	8,390,770,390,534	8,337,206,229,771	8,746,167,408,966
1. Phải trả người bán ngắn hạn	1,590,437,105,954	1,575,970,831,903	1,753,591,495,266
2. Người mua trả tiền trước ngắn hạn	2,402,024,391,289	1,597,655,019,348	1,919,276,372,631
3. Thuế và các khoản phải nộp nhà nước	208,971,331,531	400,679,502,256	363,327,162,760
4. Phải trả người lao động	365,579,148,846	288,102,845,937	316,476,056,758
5. Chi phí phải trả ngắn hạn	1,052,948,571,329	1,036,736,254,250	1,149,344,831,000
6. Phải trả nội bộ ngắn hạn			
7. Phải trả theo tiến độ kế hoạch hợp đồng xây dựng			
8. Doanh thu chưa thực hiện ngắn hạn	38,697,241,786	41,491,006,735	41,004,429,818
9. Phải trả ngắn hạn khác	567,092,159,688	260,861,099,385	357,445,718,501
10. Vay và nợ thuê tài chính ngắn hạn	1,959,414,545,347	2,897,483,366,729	2,571,970,866,987
11. Dự phòng phải trả ngắn hạn	19,003,828,492	23,083,194,750	31,516,682,782
12. Quỹ khen thưởng phúc lợi	186,602,066,272	215,143,108,478	242,213,792,463
13. Quỹ bình ổn giá			
14. Giao dịch mua bán lại trái phiếu Chính phủ			
II. Nợ dài hạn	5,482,721,942,594	6,238,665,944,819	6,128,251,863,769
1. Phải trả người bán dài hạn			
2. Người mua trả tiền trước dài hạn			
3. Chi phí phải trả dài hạn	188,387,114,899	338,801,485,090	238,323,318,020
4. Phải trả nội bộ về vốn kinh doanh			
5. Phải trả nội bộ dài hạn			
6. Doanh thu chưa thực hiện dài hạn	2,717,939,404,426	2,629,204,017,021	2,538,976,435,988
7. Phải trả dài hạn khác	44,057,480,912	39,337,326,486	58,783,034,981
8. Vay và nợ thuê tài chính dài hạn	1,657,144,167,196	2,237,289,981,401	2,240,226,202,142
9. Trái phiếu chuyển đổi			
10. Cổ phiếu ưu đãi			
11. Thuế thu nhập hoãn lại phải trả	144,422,683,836	160,518,581,689	187,316,417,409
12. Dự phòng phải trả dài hạn	408,463,901,133	426,497,213,428	435,226,216,857

13. Quỹ phát triển khoa học và công nghệ	322,307,190,192	407,017,339,704	429,400,238,372
D.VỐN CHỦ SỞ HỮU	9,085,429,076,168	9,524,317,916,696	9,952,999,655,403
I. Vốn chủ sở hữu	9,044,584,238,640	9,486,508,196,467	9,918,225,052,475
1. Vốn góp của chủ sở hữu	4,483,500,000,000	4,483,500,000,000	4,483,500,000,000
- Cổ phiếu phổ thông có quyền biểu quyết	4,483,500,000,000	4,483,500,000,000	4,483,500,000,000
- Cổ phiếu ưu đãi			
2. Thặng dư vốn cổ phần	929,867,056,019	929,867,056,019	929,867,056,019
3. Quyền chọn chuyển đổi trái phiếu			
4. Vốn khác của chủ sở hữu	17,162,355,346	17,162,355,346	17,162,355,346
5. Cổ phiếu quỹ	-1,713,600	-1,713,600	-1,713,600
6. Chênh lệch đánh giá lại tài sản	-211,681,407,015	-211,681,407,015	-211,681,407,015
7. Chênh lệch tỷ giá hối đoái	-3,205,804,051	6,457,877,936	27,034,728,326
8. Quỹ đầu tư phát triển	693,263,706,476	1,121,249,807,094	1,595,971,326,553
9. Quỹ hỗ trợ sắp xếp doanh nghiệp			
10. Quỹ khác thuộc vốn chủ sở hữu	6,257,939,977	6,257,939,977	6,257,939,977
11. Lợi nhuận sau thuế chưa phân phối	1,659,864,625,390	1,462,623,130,973	1,426,065,505,266
- LNST chưa phân phối lũy kế đến cuối kỳ trước	380,051,983,106	692,860,837,253	321,330,638,598
- LNST chưa phân phối kỳ này	1,279,812,642,284	769,762,293,720	1,104,734,866,668
12. Nguồn vốn đầu tư XDCB			
13. Lợi ích cổ đông không kiểm soát	1,469,557,480,098	1,671,073,149,737	1,644,049,261,603
II. Nguồn kinh phí và quỹ khác	40,844,837,528	37,809,720,229	34,774,602,928
1. Nguồn kinh phí			
2. Nguồn kinh phí đã hình thành TSCĐ	40,844,837,528	37,809,720,229	34,774,602,928
TỔNG CỘNG NGUỒN VỐN	22,958,921,409,296	24,100,190,091,286	24,827,418,928,138

TÀI LIỆU THAM KHẢO

1. Khoa Kinh doanh số, Trường Kinh Tế - Đại Học Công Nghiệp Hà Nội, *Tài liệu hướng dẫn khoá luận tốt nghiệp ngành Phân tích dữ liệu kinh doanh, 2024*
2. Đại học Công nghiệp Hà Nội, *Đề cương bài giảng Ứng dụng lập trình Python*
3. Đại học Công nghiệp Hà Nội, *Đề cương bài giảng Khai phá và phân tích dữ liệu lớn*
4. Đại học Công nghiệp Hà Nội, *Đề cương bài giảng Phân tích dữ liệu lớn trong Kế toán, Kiểm toán*
5. Tổng công ty VIGLACERA - CTCP, *Bảng báo cáo kết quả hoạt động kinh doanh 2022-2024*
6. Tổng công ty VIGLACERA - CTCP, Bảng cân đối kết toán *2022-2024*
7. Tổng công ty VIGLACERA – CTCP, Website chính thức: https://www.viglacera.com.vn/

8. Trang thông tin điện tử tổng hợp CAFEF. Website chính thức: https://cafef.vn/

9. Investing Việt Nam, *Dữ liệu giá cổ phiếu Viglacera (VGC),* https://www.investing.com/equities/viglacera-corporation-jsc

10. Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5–32.

11. Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (2013). *A comparison of random forest regression and multiple linear regression for prediction in neuroscience*. Journal of Neuroscience Methods, 220(1), 85–95.

12. Hochreiter, S., & Schmidhuber, J. (1997). *Long Short-Term Memory*. Neural Computation, 9(8), 1735–1780.

13. Fischer, T., & Krauss, C. (2018). *Deep learning with long short-term memory networks for financial market predictions*. European Journal of Operational Research, 270(2), 654–669.

14. Lewis-Beck, M. S., & Skalaban, A. (1990). *The R-Squared: Some Straight Talk*. Political Analysis, 2(1), 153–171.

15. Chai, T., & Draxler, R. R. (2014). *Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE*. Geoscientific Model Development, 7, 1247–125