TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI
TRƯỜNG KINH TẾ

NGUYỄN ĐÌNH ĐẠT

KHÓA LUẬN TỐT NGHIỆP

NGÀNH PHÂN TÍCH DỮ LIỆU KINH DOANH

Ứng dụng phân tích dữ liệu và học máy trong tối ưu danh mục đầu tư tài chính

Hà Nội – 2025

TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI

TRƯỜNG KINH TẾ

KHÓA LUẬN TỐT NGHIỆP

NGÀNH PHÂN TÍCH DỮ LIỆU KINH DOANH

Ứng dụng phân tích dữ liệu và học máy trong tối ưu danh mục đầu tư tài chính

Họ và tên sinh viên : Nguyễn Đình Đạt

Mã sinh viên : 2021607831

Lớp - Khóa : 2021DHDLKD01 – K16

Giảng viên hướng dẫn : TS. Nguyễn Thị Thu Thủy

TS. Trần Hùng Cường

Hà Nội – 2025

MỤC LỤC

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi, được thực hiện dưới sự hướng dẫn khoa học của TS. Nguyễn Thị Thu Thuỷ và TS. Trần Hùng Cường.

Các số liệu, kết quả trình bày trong khóa luận tốt nghiệp với đề tài "Ứng dụng phân tích dữ liệu và học máy trong tối ưu danh mục đầu tư tài chính" là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tôi xin cam đoan rằng mọi sự giúp đỡ cho việc thực hiện khóa luận này đã được cảm ơn và các thông tin trích dẫn trong khóa luận đều đã được chỉ rõ nguồn gốc.

Hà Nội, ngày 21 tháng 06 năm 2025

Sinh viên

Nguyễn Đình Đạt

LỜI CẢM ƠN

Để hoàn thành khóa luận tốt nghiệp này, bên cạnh sự nỗ lực của bản thân, tôi đã nhận được rất nhiều sự quan tâm, hướng dẫn và hỗ trợ quý báu từ các cá nhân, tổ chức. Tôi xin được bày tỏ lòng biết ơn chân thành và sâu sắc nhất.

Trước hết, tôi xin gửi lời tri ân sâu sắc tới TS. Nguyễn Thị Thu Thuỷ và TS. Trần Hùng Cường những người thầy cô đã tận tình hướng dẫn, định hướng và luôn đồng hành cùng tôi trong suốt quá trình thực hiện khóa luận. Những góp ý chuyên môn, sự tận tâm chỉ bảo và động viên kịp thời của các thầy cô chính đã giúp tôi hoàn thiện công trình nghiên cứu này.

Tôi xin chân thành cảm ơn Ban Giám hiệu Trường Đại học Công nghiệp Hà Nội, Ban Lãnh đạo Trường Kinh tế cùng các thầy cô trong Khoa Kinh doanh số đã tạo điều kiện thuận lợi về cơ sở vật chất, môi trường học tập và nghiên cứu trong suốt quá trình học tập và thực hiện đề tài.

Cuối cùng, tôi xin gửi lời cảm ơn chân thành đến gia đình, bạn bè và những người thân yêu đã luôn đồng hành, động viên, chia sẻ và tạo điều kiện tốt nhất để tôi có thể yên tâm học tập và hoàn thành khóa luận này.

Mặc dù đã có nhiều cố gắng trong quá trình nghiên cứu và hoàn thiện, nhưng do hạn chế về kiến thức và kinh nghiệm, khóa luận chắc chắn vẫn còn những thiếu sót nhất định. Tôi rất mong nhận được những ý kiến đóng góp quý báu từ các thầy cô và các bạn để nghiên cứu được hoàn thiện hơn.

Xin chân thành cảm ơn!

DANH MỤC CÁC TỪ VIẾT TẮT

Từ viết tắt	Ý nghĩa
XGBoost	eXtreme Gradient Boosting
MVO	Mean Variance Optimization
LSTM	Long-short term memory
MAE	Mean Absolute Error
MSE	Mean Square Error
RMSE	Root Mean Square Error
TTCK	Thị trường chứng khoán
SGDCK	Sở giao dịch chứng khoán
OTC	Over The Counter
DMĐT	Danh mục đầu tư
ETFs	Exchange-traded fund
MPT	Modern Portfolio Theory
ML	Machine Learning
BO	Bayesian Optimization
RNN	Recurrent Neural Network
DL	Deep Learning

DANH MỤC BẢNG

DANH MỤC HÌNH ẢNH

MỞ ĐẦU

1. Lý do chọn đề tài

Trong bối cảnh nền kinh tế số phát triển mạnh mẽ và hệ thống tài chính toàn cầu ngày càng phức tạp, việc quản lý danh mục đầu tư đòi hỏi những phương pháp ra quyết định hiện đại dựa trên cơ sở dữ liệu và công nghệ tiên tiến. Các phương pháp đầu tư truyền thống, chủ yếu dựa vào kinh nghiệm cá nhân và các mô hình tài chính cổ điển như lý thuyết Markowitz, đang dần bộc lộ những hạn chế đáng kể trong môi trường thị trường hiện đại với biến động nhanh chóng và khối lượng dữ liệu khổng lồ phát sinh liên tục.

Sự bùng nổ của dữ liệu lớn và những tiến bộ vượt bậc trong lĩnh vực học máy đã tạo ra cơ hội chưa từng có để cách mạng hóa việc phân tích đầu tư tài chính. Các thuật toán học máy hiện đại như XGBoost, LightGBM, CatBoost và các mô hình mạng nơ-ron sâu đã chứng minh khả năng vượt trội trong việc xử lý dữ liệu phi tuyến, phát hiện các mẫu hình phức tạp và đưa ra dự báo chính xác hơn so với các phương pháp truyền thống.

Thực tiễn cho thấy các tổ chức tài chính hàng đầu thế giới đã bắt đầu tích hợp các công nghệ này vào quy trình đầu tư của mình để nâng cao hiệu suất và kiểm soát rủi ro. Tuy nhiên, việc nghiên cứu có hệ thống về ứng dụng phân tích dữ liệu và học máy trong tối ưu hóa danh mục đầu tư vẫn còn nhiều khoảng trống, đặc biệt là trong việc tích hợp các nguồn dữ liệu đa dạng và xây dựng các mô hình phù hợp với đặc thù thị trường tài chính.

Xuất phát từ những nhu cầu thực tiễn này, đề tài "Ứng dụng phân tích dữ liệu và học máy trong tối ưu danh mục đầu tư tài chính" được lựa chọn nhằm khám phá và vận dụng tiềm năng của các kỹ thuật phân tích hiện đại để nâng cao hiệu quả ra quyết định đầu tư. Nghiên cứu này không chỉ có ý nghĩa thực tiễn quan trọng đối với các nhà đầu tư và tổ chức tài chính mà còn góp phần mở rộng cơ sở lý thuyết và ứng dụng của học máy trong lĩnh vực tài chính, đáp ứng xu thế phát triển tất yếu trong thời đại chuyển đổi số.

2. Mục tiêu nghiên cứu

- Hệ thống hóa các lý thuyết và phương pháp tối ưu danh mục đầu tư tài chính, đặc biệt là mô hình Mean-Variance Optimization.
- Áp dụng các mô hình học máy (như XGBoost, LSTM) để dự báo lợi suất tài sản và đo lường rủi ro.
- Tích hợp kết quả dự báo vào mô hình tối ưu danh mục đầu tư nhằm nâng cao hiệu quả phân bổ tài sản.
- So sánh hiệu quả danh mục đầu tư tối ưu giữa phương pháp truyền thống và phương pháp ứng dụng học máy dựa trên các chỉ tiêu tài chính như lợi suất kỳ vọng, độ lệch chuẩn, hệ số Sharpe.
- Đề xuất các chiến lược phân bổ tài sản và tái cân bằng danh mục đầu tư dựa trên kết quả mô hình và bối cảnh thị trường thực tế.

3. Phạm vi và đối tượng nghiên cứu

Đối tượng nghiên cứu: Các mô hình phân tích dữ liệu và học máy trong bối cảnh đầu tư tài chính, đặc biệt các mô hình dự báo như LSTM hay XgBoost và mô hình tối ưu danh mục đầu tư MVO.

Phạm vi dữ liệu: Dữ liệu được thu thập trên thị trường chứng khoán Việt Nam, bao gồm các thông tin về giá mở cửa, giá đóng cửa, khối lượng giao dịch. Dữ liệu được chuẩn bị, làm sạch và xử lý trong giai đoạn từ 2015 đến 2025.

Phạm vi ứng dụng: Đề tài tập trung vào việc hỗ trợ trong việc tối ưu danh mục đầu tư cho các nhà đầu tư, trong bối cảnh thị trường nhiều biến động việc lựa chọn và phân bổ danh mục vô cùng quan trọng giúp tối ưu giữa lợi nhuận và rủi ro.

4. Phương pháp nghiên cứu

Tiền xử lý dữ liệu: Làm sạch dữ liệu, xử lý giá trị thiếu, chuẩn hóa và tạo các biến dẫn xuất.

Xây dựng mô hình: Sử dụng các thư viện Python (như scikit-learn, tensorflow, keras) để huấn luyện các mô hình dự báo tỉ suất lợi nhuận và các thư viện tài chính (như ta,

math) để tối ưu danh mục đầu tư. Áp dụng các phương pháp tối ưu tham số cho các mô hình như GridSearchCV.

Đánh giá mô hình: Sử dụng các chỉ số như MAE, MSE, RMSE để đánh giá mô hình dự báo và các chỉ số như Sharp ratio, Return để đánh giá các danh mục đầu tư.

Ứng dụng kết quả: Trình bày các khuyến nghị cho các nhà đầu tư, các quỹ đầu tư về cách sử dụng dữ liệu và cách kết hợp các mô hình để hỗ trợ trong việc ra quyết định đầu tư.

5. Cấu trúc báo cáo

Ngoài phần Mở đầu, Kết luận, Tài liệu tham khảo và Phụ lục, khóa luận gồm 4 chương chính:

Chương 1: Tổng quan về đầu tư tài chính và phân tích dữ liệu

Chương 2: Mô hình nghiên cứu và phương pháp nghiên cứu

Chương 3: Thực nghiệm và đánh giá kết quả

Chương 4: Kết luận và đề xuất khuyến nghị

CHƯƠNG I: TỔNG QUAN VỀ ĐẦU TƯ TÀI CHÍNH VÀ PHÂN TÍCH DỮ LIỆU

1.1 Tổng quan về thị trường chứng khoán

1.1.1 Thị trường chứng khoán

Thị trường chứng khoán (TTCK) được định nghĩa là một bộ phận quan trọng của thị trường vốn dài hạn, đóng vai trò thực hiện cơ chế luân chuyển vốn trực tiếp từ nhà đầu tư đến nhà phát hành [18]. Thông qua cơ chế này, thị trường chứng khoán thực hiện chức năng cốt lõi của thị trường tài chính là cung ứng nguồn vốn trung và dài hạn cho nền kinh tế, góp phần thúc đẩy tăng trưởng kinh tế và phát triển bền vững [37]

Dựa theo tính chất địa lý, thị trường chứng khoán được chia thành hai loại chính. Thứ nhất là thị trường chứng khoán tập trung, đây là địa điểm hoạt động chính thức của các giao dịch chứng khoán, thường được gọi là Sở giao dịch chứng khoán (SGDCK). Tại đây, các nhà môi giới kinh doanh chứng khoán gặp gỡ, đấu giá và thương lượng mua bán chứng khoán cho khách hàng theo nguyên tắc và quy chế mà SGDCK đề ra theo luật chứng khoán [15]. Thứ hai là thị trường chứng khoán không tập trung (OTC - Over The Counter), đây là hoạt động giao dịch chứng khoán không thông qua SGDCK mà được thực hiện rải rác khắp nơi trên thế giới thông qua các thiết bị mạng điện thoại và mạng vi tính [30]

Căn cứ theo quá trình luân chuyển vốn, thị trường chứng khoán được phân thành thị trường sơ cấp và thị trường thứ cấp. Thị trường sơ cấp, còn được gọi là thị trường phát hành, là nơi diễn ra các giao dịch mua bán chứng khoán được phát hành lần đầu [12]. Trên thị trường này, nguồn vốn từ các nhà đầu tư sẽ được chuyển trực tiếp sang nhà phát hành thông qua việc các nhà đầu tư mua chứng khoán mới phát hành Ngược lại, thị trường thứ cấp là nơi diễn ra các giao dịch chứng khoán đã được phát hành trên thị trường sơ cấp, đây là thị trường chuyên nhượng quyền sở hữu các chứng khoán và đóng vai trò quan trọng trong việc đảm bảo tính thanh khoản cho các chứng khoán đã phát hành.[6]

Dựa vào đặc điểm hàng hóa giao dịch, thị trường chứng khoán có thể được phân loại thành ba nhóm chính. Thị trường trái phiếu là nơi mua bán các loại trái phiếu, bao

gồm trái phiếu chính phủ và trái phiếu doanh nghiệp. Thị trường cổ phiếu là nơi mua bán các cổ phiếu của các công ty đại chúng [21]. Cuối cùng là thị trường phái sinh, nơi mua bán các công cụ tài chính có nguồn gốc từ chứng khoán cơ sở.

a. Chức năng cơ bản của thị trường chứng khoán

Thị trường chứng khoán thực hiện bốn chức năng cơ bản quan trọng đối với nền kinh tế. Chức năng đầu tiên và quan trọng nhất là huy động vốn đầu tư cho nền kinh tế. Thị trường chứng khoán có nhiệm vụ huy động vốn đầu tư cho các doanh nghiệp thông qua việc phát hành cổ phiếu, trái phiếu và các công cụ tài chính khác [44]. Các doanh nghiệp sử dụng vốn thu được để đầu tư vào các dự án mở rộng hoặc phát triển hoạt động kinh doanh, từ đó tăng cường nguồn vốn và sức mạnh cho toàn bộ nền kinh.

Chức năng thứ hai là cung cấp môi trường đầu tư đa dạng và an toàn cho các nhà đầu tư. Thị trường chứng khoán hoạt động như một hệ thống tổ chức có cấu trúc, trong đó các nhà đầu tư có thể mua, bán và chuyển nhượng các loại tài sản tài chính cho nhau một cách thuận lợi và minh bạch [16].

Chức năng thứ ba là tạo tính thanh khoản cho các chứng khoán. Tính thanh khoản của chứng khoán được coi là một yếu tố cực kỳ quan trọng đối với các nhà đầu tư. Chứng khoán có tính thanh khoản cao giúp các nhà đầu tư thực hiện các giao dịch một cách dễ dàng và nhanh chóng, từ đó giảm thiểu rủi ro đầu tư và tăng khả năng sinh lời

Chức năng cuối cùng là tạo lập môi trường hỗ trợ Chính phủ thực hiện các chính sách kinh tế vĩ mô [16]. Thị trường chứng khoán giúp Chính phủ tác động hiệu quả đến nền kinh tế thông qua việc triển khai các chính sách kinh tế vĩ mô, bao gồm các chính sách tiền tệ và tài khóa nhằm điều chỉnh các hoạt động kinh tế trong phạm vi quốc gia.

b. Vai trò của thị trường chứng khoán

Đối với Chính phủ

Thị trường chứng khoán đóng vai trò quan trọng trong việc hỗ trợ Chính phủ thực hiện các chức năng quản lý kinh tế. Thứ nhất, thị trường này cung cấp công cụ đánh giá hiệu quả hoạt động của các doanh nghiệp thông qua giá cổ phiếu và các chỉ số tài chính [5]. Thứ hai, thị trường chứng khoán cung cấp các phương tiện hiệu quả để huy động vốn và phân bổ nguồn vốn một cách tối ưu cho nền kinh tế quốc dân. Chính phủ có thể

huy động vốn thông qua việc phát hành trái phiếu chính phủ và sử dụng số tiền thu được để đầu tư vào các dự án cơ sở hạ tầng và phát triển kinh tế xã hội [5].

Thị trường chứng khoán còn góp phần thúc đẩy quá trình cổ phần hóa và khuyến khích việc phân hóa nhanh chóng các doanh nghiệp quốc doanh, tạo điều kiện cho việc nâng cao hiệu quả hoạt động của khu vực kinh tế nhà nước [4]. Đồng thời, thị trường này còn là nơi Chính phủ thực hiện các chính sách tiền tệ thông qua việc mua bán trái phiếu chính phủ, từ đó tác động đến lãi suất thị trường. Ngoài ra, thị trường chứng khoán còn là kênh quan trọng thu hút các nguồn vốn đầu tư gián tiếp từ nước ngoài thông qua việc các nhà đầu tư nước ngoài tham gia mua bán chứng khoán trên thị trường trong nước [17].

Đối với các doanh nghiệp

Thị trường chứng khoán mang lại nhiều lợi ích quan trọng cho các doanh nghiệp trong việc huy động vốn và phát triển kinh doanh. Trước hết, thị trường này giúp các công ty có thể thoát khỏi sự phụ thuộc vào các khoản vay ngân hàng có chi phí cao bằng cách phát hành cổ phiếu hoặc trái phiếu doanh nghiệp [20]. Tính thanh khoản cao của thị trường chứng khoán tạo điều kiện cho các doanh nghiệp bán chứng khoán bất kỳ lúc nào khi cần thiết, từ đó tăng khả năng huy động vốn linh hoạt.

Thị trường chứng khoán còn đóng vai trò như một công cụ đánh giá giá trị doanh nghiệp và của cả nền kinh tế thông qua các chỉ số chứng khoán, giúp các nhà đầu tư và các bên liên quan có cái nhìn khách quan về hiệu quả hoạt động của doanh nghiệp [35]. Hơn nữa, việc niêm yết trên thị trường chứng khoán còn giúp doanh nghiệp quảng bá thương hiệu, thể hiện độ uy tín và thúc đẩy quá trình ra mắt công chúng, tạo điều kiện tiếp cận được nhiều khách hàng và đối tác tiềm năng hơn

Đối với nhà đầu tư

Thị trường chứng khoán cung cấp một môi trường đầu tư đa dạng và phong phú cho các nhà đầu tư. Đây là nơi các nhà đầu tư có thể tìm kiếm và khai thác các cơ hội đầu tư khác nhau để xây dựng danh mục đầu tư đa dạng, từ đó giảm thiểu rủi ro tổng thể trong hoạt động đầu tư của mình.

c. Các nguyên tắc hoạt động của thị trường chứng khoán

Thị trường chứng khoán hoạt động dựa trên bốn nguyên tắc cơ bản nhằm đảm bảo tính hiệu quả và công bằng. Nguyên tắc cạnh tranh đảm bảo rằng giá cả trên thị trường chứng khoán phản ánh chính xác mối quan hệ cung cầu về chứng khoán và thể hiện tương quan cạnh tranh giữa các công ty [40]. Trên thị trường sơ cấp, các nhà phát hành cạnh tranh để bán chứng khoán của mình cho các nhà đầu tư, trong khi các nhà đầu tư lựa chọn chứng khoán phù hợp với mục tiêu đầu tư của mình. Trên thị trường thứ cấp, các nhà đầu tư cạnh tranh tự do để tìm kiếm lợi nhuận cao nhất thông qua phương thức đấu giá.

Cuối cùng, nguyên tắc trung gian quy định rằng các giao dịch chứng khoán phải được thực hiện thông qua các tổ chức trung gian, hay còn được gọi là các công ty môi giới chứng khoán, nhằm đảm bảo tính chuyên nghiệp và an toàn trong các giao dịch.

1.1.2 Tổng quan về Cổ Phiếu

Cổ phiếu được định nghĩa là những giấy tờ có giá trị xác định số vốn đầu tư và đồng thời xác nhận quyền sở hữu về tài sản cũng như các điều kiện về thu nhập trong một khoảng thời gian nhất định, đồng thời có khả năng chuyển nhượng trên thị trường [23]. Cổ phiếu thể hiện quyền sở hữu một phần vốn của công ty cổ phần và mang lại cho người sở hữu những quyền lợi cụ thể liên quan đến hoạt động của doanh nghiệp. Trên thị trường chứng khoán, cổ phiếu được phân loại thành hai loại chính: cổ phiếu thường và cổ phiếu ưu đãi, mỗi loại có những đặc điểm và quyền lợi riêng biệt [36].

a. Cổ phiếu thường

Cổ phiếu thường là loại chứng khoán vốn cổ phần không có thời gian đáo hạn cố định và không đảm bảo thu nhập ổn định cho người nắm giữ [13]. Thu nhập từ cổ phiếu thường hoàn toàn phụ thuộc vào kết quả sản xuất kinh doanh của công ty cũng như chính sách chi trả cổ tức mà ban lãnh đạo công ty đề ra. Một trong những ưu điểm nổi bật của cổ phiếu thường là khả năng chuyển nhượng dễ dàng trên thị trường thứ cấp, giúp nhà đầu tư có thể linh hoạt trong việc mua bán theo nhu cầu đầu tư của mình.

Người nắm giữ cổ phiếu thường được hưởng các quyền cơ bản của cổ đông, bao gồm quyền tham gia bầu cử vào Hội đồng quản trị công ty, qua đó có tiếng nói trong việc định hướng chiến lược phát triển của doanh nghiệp. Tuy nhiên, cổ đông sở hữu cổ phiếu thường không được ưu tiên trong việc phân chia lợi nhuận cũng như thanh lý tài sản khi

công ty gặp khó khăn tài chính hoặc phá sản [9]. Bên cạnh đó, cổ phiếu thường còn mang lại cho các cổ đông quyền đặt mua cổ phiếu mới khi công ty phát hành thêm cổ phiếu, giúp duy trì tỷ lệ sở hữu của họ trong công ty.

b. Cổ phiếu ưu đãi

Cổ phiếu ưu đãi là loại chứng khoán vốn có những đặc quyền vượt trội so với cổ phiếu thường, đặc biệt trong việc phân chia lợi nhuận, chi trả cổ tức và thanh lý tài sản khi công ty phá sản [25]. Tương tự như cổ phiếu thường, cổ phiếu ưu đãi không có thời gian đáo hạn cố định và tồn tại song song với sự tồn tại của công ty phát hành. Mặc dù có khả năng chuyển nhượng, việc giao dịch cổ phiếu ưu đãi thường phải thỏa mãn những điều kiện nhất định do công ty quy định.

Cổ phiếu ưu đãi sở hữu ba đặc tính quan trọng giúp phân biệt với cổ phiếu thường. Thứ nhất, cổ phiếu ưu đãi có tính chất tham dự trong phân chia lợi nhuận khi công ty đạt được mức lãi vượt trội nhất định, cho phép cổ đông hưởng lợi từ sự tăng trưởng mạnh mẽ của doanh nghiệp. Thứ hai, trong điều kiện bình thường, cổ phiếu ưu đãi không có quyền biểu quyết trong các cuộc họp cổ đông [22]. Tuy nhiên, khi công ty gặp khó khăn tài chính hoặc làm ăn thua lỗ, cổ đông sở hữu cổ phiếu ưu đãi sẽ được trao quyền biểu quyết để tham gia vào quá trình ra quyết định quan trọng của công ty.

Thứ ba, cổ phiếu ưu đãi có thể mang tính chất tích lũy hoặc không tích lũy cổ tức, tùy thuộc vào điều khoản phát hành mà công ty đề ra. Trong trường hợp công ty làm ăn không hiệu quả và không thể chi trả cổ tức trong năm tài chính, với cổ phiếu ưu đãi có tính tích lũy, công ty có nghĩa vụ thanh toán số cổ tức chưa trả này trong các năm tiếp theo khi hoạt động kinh doanh khả quan trở lại. Ngược lại, với cổ phiếu ưu đãi không tích lũy, công ty không cần phải bù đắp những khoản cổ tức đã bỏ lỡ trong các năm trước đó [41].

1.1.3 Tổng quan các nghiên cứu về ứng dụng phân tích dữ liệu trong tối ưu danh mục đầu tư

Hiện nay, việc ứng dụng các mô hình học máy vào lĩnh vực tài chính đang thu hút sự quan tâm mạnh mẽ nhằm nâng cao hiệu quả dự báo rủi ro và tối ưu hóa danh mục đầu tư. Theo nghiên cứu của Aftab Uddin và cộng sự (2025), các thuật toán hiện đại như Random Forest, Gradient Boosting, Long Short-Term Memory (LSTM) và đặc biệt là

Transformer có khả năng vượt trội trong việc khai thác dữ liệu tài chính phức tạp, giúp cải thiện đáng kể độ chính xác dự báo lợi suất, quản lý rủi ro và tối ưu hiệu suất danh mục so với các mô hình truyền thống. Tuy nhiên, phần lớn các nghiên cứu hiện nay mới chỉ tập trung so sánh riêng lẻ từng mô hình mà chưa khai thác hiệu quả các mô hình kết hợp (hybrid) hoặc ensemble để tận dụng ưu điểm của từng thuật toán. Bên cạnh đó, việc tích hợp yếu tố hành vi thị trường và dữ liệu phi cấu trúc, cũng như mở rộng nghiên cứu sang các tài sản phi truyền thống hoặc thị trường kém thanh khoản vẫn còn là khoảng trống cần tiếp tục được khám phá nhằm nâng cao tính thực tiễn và khả năng ứng dụng của các mô hình học máy trong bối cảnh thị trường tài chính ngày càng biến động phức tạp.

Trong bối cảnh các mô hình tối ưu hóa danh mục đầu tư truyền thống như Markowitz vẫn bộc lộ những hạn chế về tính hiệu quả ngoài mẫu, đặc biệt khi dữ liệu đầu vào như lợi suất kỳ vọng và ma trận hiệp phương sai được ước lượng không chính xác, nhiều nghiên cứu gần đây đã tập trung ứng dụng các kỹ thuật học máy để cải thiện vấn đề này. Tiêu biểu, nghiên cứu của Bùi Quốc Hoàn và Phạm Thị Hương Nguyên (2023) đã vận dụng phương pháp chính quy hóa, cụ thể là kỹ thuật Lasso, nhằm nâng cao hiệu quả của mô hình Markowitz trong xây dựng danh mục đầu tư tối ưu trên thị trường chứng khoán Việt Nam. Kết quả thực nghiệm cho thấy danh mục xây dựng bằng phương pháp này đạt được mức rủi ro thấp hơn và lợi suất trung bình cao hơn so với chỉ số VN-Index, đồng thời đảm bảo tính ổn định qua nhiều giai đoạn thị trường, kể cả trước và sau đại dịch Covid-19. Tuy nhiên, nghiên cứu chủ yếu tập trung vào kỹ thuật chính quy hóa tuyến tính và chưa khai thác sâu các mô hình học máy phi tuyến hoặc mô hình kết hợp nhiều kỹ thuật hiện đại khác, mở ra khoảng trống cho các nghiên cứu tiếp theo trong việc ứng dụng các mô hình học sâu, mô hình hybrid hoặc kết hợp dữ liệu phi cấu trúc để nâng cao hơn nữa khả năng tối ưu hóa danh mục đầu tư trong điều kiện thị trường thực tế nhiều biến động.

Theo nghiên cứu của Ángel Samaniego Alcántar (2025) đã tiếp cận theo hướng sử dụng mạng nơ-ron LSTM (Long-Short Term Memory) để xây dựng danh mục tối ưu dựa trên kỳ vọng lợi suất. Khác với phần lớn các nghiên cứu trước đây vốn tập trung khai thác LSTM trong dự báo ngắn hạn giá cổ phiếu, tác giả đã áp dụng mô hình này cho mục tiêu phân bổ tài sản trong danh mục đầu tư với các kỳ hạn từ 1 đến 2 năm, sử

dụng dữ liệu từ các cổ phiếu thuộc chỉ số Dow Jones Industrial Average (DJIA) giai đoạn 2000-2021. Kết quả cho thấy, danh mục tối ưu dựa trên kỳ vọng lợi suất từ mô hình LSTM vượt trội so với DJIA truyền thống, với mức lợi nhuận cao hơn từ 3,7% đến 5% và xác suất vượt thị trường đạt tới 85,4% cho kỳ hạn 1,5 năm. Tuy nhiên, nghiên cứu vẫn còn một số giới hạn khi chỉ xét trên tập dữ liệu DJIA và chưa tích hợp yếu tố rủi ro trực tiếp vào quy trình tối ưu hóa, cũng như chưa kiểm chứng khả năng của mô hình trên các thị trường mới nổi hoặc các loại tài sản đa dạng hơn.

Lý thuyết danh mục đầu tư hiện đại (Markowitz, 1952) đóng vai trò nền tảng trong quản lý danh mục, tuy nhiên trên thực tế vẫn tồn tại nhiều thách thức chưa được giải quyết triệt để, đặc biệt là ở giai đoạn lựa chọn tài sản đầu vào cho danh mục. Nhằm khắc phục hạn chế này, nghiên cứu của Wang và cộng sự (2020) đã đề xuất mô hình kết hợp giữa mạng nơ-ron hồi tiếp dài hạn (LSTM) và mô hình Mean-Variance (MV) trong tối ưu hóa danh mục đầu tư. Điểm nổi bật của nghiên cứu là nhấn mạnh vai trò của bước tiền xử lý chọn lọc tài sản dựa trên dự báo lợi suất bằng LSTM, từ đó cung cấp dữ liệu đầu vào chất lượng cao hơn cho quá trình tối ưu hóa danh mục. Kết quả thực nghiệm trên dữ liệu dài hạn từ chỉ số FTSE 100 (1994-2019) cho thấy mô hình LSTM + MV giúp cải thiện đáng kể lợi suất tích lũy, tỷ lệ Sharpe và hiệu quả phân bổ vốn so với các mô hình truyền thống và một số thuật toán học máy khác như SVM, Random Forest hay DNN. Tuy nhiên, nghiên cứu vẫn tập trung chủ yếu vào dữ liệu thị trường Anh quốc với nhóm cổ phiếu niêm yết trên FTSE 100, chưa kiểm chứng khả năng tổng quát hóa của mô hình trên các thị trường mới nổi hoặc trong bối cảnh tài sản có tính chất khác biệt như trái phiếu, bất động sản hay tiền điện tử. Đồng thời, tác giả chưa xem xét đầy đủ tác động của yếu tố hành vi thị trường và biến động vĩ mô, đây chính là những khoảng trống cần tiếp tục nghiên cứu nhằm nâng cao hơn nữa tính thực tiễn và hiệu quả ứng dụng của các mô hình học sâu trong quản lý danh mục đầu tư.

Nghiên cứu của Van-Dai Ta và cộng sự (2020) đã tập trung khai thác mô hình mạng nơ-ron hồi tiếp LSTM nhằm dự báo giá cổ phiếu và tối ưu hóa danh mục đầu tư. Khác với các nghiên cứu trước đây chỉ sử dụng LSTM cho mục tiêu dự báo riêng lẻ, nhóm tác giả đã kết hợp kết quả dự báo từ LSTM với các kỹ thuật tối ưu hóa danh mục như phân bổ vốn đều (EQ), mô phỏng Monte Carlo (MCS) và tối ưu hóa trung bình-phương sai (MVO) để xây dựng danh mục đầu tư hiệu quả. Trên cơ sở dữ liệu lịch sử 10 năm của

các cổ phiếu thuộc chỉ số S&P500, kết quả cho thấy danh mục đầu tư dựa trên dự báo của LSTM kết hợp các kỹ thuật tối ưu hóa không chỉ đạt lợi nhuận cao hơn mà còn kiểm soát rủi ro tốt hơn so với các mô hình truyền thống như hồi quy tuyến tính (LR), máy vector hỗ trợ (SVR) hay cả benchmark S&P500. Tuy vậy, nghiên cứu chủ yếu tập trung vào thị trường Mỹ và các cổ phiếu vốn hóa lớn, chưa kiểm chứng khả năng áp dụng mô hình này trên các thị trường mới nổi, thị trường có tính thanh khoản thấp hoặc các loại tài sản phi truyền thống như vàng, bất động sản, tiền điện tử. Đồng thời, yếu tố dữ liệu phi cấu trúc (như cảm xúc thị trường, tin tức) chưa được tích hợp đầy đủ trong mô hình dự báo, đây chính là những khoảng trống tiềm năng cho các nghiên cứu tiếp theo nhằm nâng cao tính thực tiễn và khả năng ứng dụng rộng rãi của mô hình học sâu trong tối ưu hóa danh mục đầu tư.

1.2 Danh mục đầu tư

Danh mục đầu tư (DMĐT) là khoản đầu tư được phân bổ vào nhiều loại chứng khoán khác nhau nhằm đạt được lợi nhuận kỳ vọng, trong đó lợi nhuận này có mối quan hệ tỷ lệ thuận với mức độ rủi ro dự kiến của dự án đầu tư. Khác biệt với đầu tư trực tiếp, đầu tư theo danh mục đòi hỏi việc nắm giữ một tỷ lệ nhất định cổ phiếu của một hoặc nhiều công ty và có thể yêu cầu quản lý thường xuyên [3]. Danh mục đầu tư hiện đại có thể bao gồm nhiều loại tài sản đa dạng như cổ phiếu, trái phiếu chính phủ, trái phiếu doanh nghiệp, tín phiếu, bất động sản, các quỹ ETFs và chứng chỉ quỹ.

Cơ cấu của danh mục đầu tư được xác định bởi nhiều yếu tố quan trọng, trong đó ba yếu tố chính bao gồm khả năng chấp nhận rủi ro của nhà đầu tư, phạm vi đầu tư và quy mô vốn đầu tư sẵn có. Lý thuyết danh mục đầu tư hiện đại đã chứng minh rằng rủi ro tổng thể của danh mục đầu tư luôn thấp hơn tổng rủi ro theo tỷ trọng của từng chứng khoán riêng lẻ trong danh mục [10]. Chính vì lý do này, các tổ chức và cá nhân thường áp dụng chiến lược đầu tư danh mục để phân tán rủi ro trên nhiều chứng khoán khác nhau, từ đó giúp giảm thiểu tổng rủi ro trong hoạt động đầu tư.

Lý thuyết danh mục đầu tư hiện đại (Modern Portfolio Theory - MPT) do nhà kinh tế học người Mỹ Harry Markowitz đề xuất trong bài nghiên cứu "Portfolio Selection" được xuất bản trên tạp chí Journal of Finance năm 1952 Modern Portfolio Theory: What MPT Is and How Investors Use It, đã trở thành nền tảng lý thuyết cho

việc xây dựng danh mục tối ưu. MPT nhấn mạnh tầm quan trọng của việc đa dạng hóa danh mục đầu tư nhằm giảm thiểu các rủi ro phi hệ thống, đồng thời xác định danh mục tối ưu dựa trên mối quan hệ cân bằng giữa lợi suất kỳ vọng và mức độ rủi ro tương ứng [17]. Các nghiên cứu và mô hình toán học đã chỉ ra rằng việc duy trì một danh mục đầu tư được đa dạng hóa tốt với 25 đến 30 cổ phiếu sẽ mang lại mức độ giảm rủi ro hiệu quả nhất về chi phí.

1.2.1 Quản lý danh mục đầu tư

Quản lý danh mục đầu tư chứng khoán được hiểu là quá trình xây dựng một danh mục bao gồm các loại chứng khoán và tài sản đầu tư nhằm đáp ứng tốt nhất nhu cầu của nhà đầu tư, sau đó thực hiện điều chỉnh liên tục các danh mục này để đạt được những mục tiêu đã đề ra [13]. Quản lý danh mục đầu tư đòi hỏi việc lựa chọn một hỗn hợp các tài sản để giảm thiểu rủi ro tổng thể, đa dạng hóa các khoản đầu tư để tối đa hóa lợi nhuận tiềm năng, và tái cân bằng danh mục thường xuyên để duy trì sự phân bổ phù hợp.

Quản lý danh mục đầu tư là một phương pháp hiệu quả để các tổ chức quản lý sản phẩm của họ thông qua vòng đời phát triển, ưu tiên, kiểm soát và các phương pháp tiếp cận nhất quán Harry Markowitz's Modern Portfolio Theory. Đây là một quá trình năng động diễn ra liên tục và có hệ thống, bao gồm bốn thành tố cơ bản. Thành tố đầu tiên là xác định mục tiêu về mức lợi suất và rủi ro, được thực hiện thông qua việc phân tích các yêu cầu, mức độ ưu tiên và những hạn chế của chủ đầu tư. Các nhà đầu tư cá nhân có nhiều mục tiêu cá nhân khác nhau, sở thích rủi ro và nguồn lực khác nhau. Mục tiêu của họ bao gồm tiết kiệm cho việc nghỉ hưu, tích lũy tài sản cho các khoản mua lớn, tài trợ giáo dục cho con cái, hoặc xây dựng quỹ khẩn cấp What Is Diversification? Definition as Investing Strategy [19]. Quá trình này đòi hỏi việc xác định mức độ rủi ro mà nhà đầu tư có thể chấp nhận được và mức độ lợi nhuận mong muốn phù hợp với mức rủi ro tương ứng, vì mối quan hệ giữa rủi ro và lợi nhuận là một nguyên lý cơ bản trong lý thuyết tài chính, với quy luật chung là "mức độ rủi ro càng cao thì lợi nhuận tiềm năng càng lớn"

1.2.2 Lợi suất kỳ vọng và rủi ro của chứng khoán riêng lẻ

Lợi nhuận và rủi ro là hai phạm trù căn bản của nền tảng trong đầu tư chứng khoán. Mô hình Markowitz đã phát triển phương pháp lựa chọn danh mục đầu tư trên hai yếu tố lợi nhuận kỳ vọng và rủi ro. Tuy nhiên, việc tính toán và định lượng hai yếu tố này lại không phải vấn đề đơn giản. Markowitz là một trong số những người đầu tiên khởi xướng việc ứng dụng các phương pháp toán học – xác suất – thống kê vào việc tính toán lợi nhuận kỳ vọng và rủi ro, trước hết là lợi nhuận kỳ vọng và rủi ro của một chứng khoán riêng lẻ.

1.2.1 Lợi suất

Lợi suất (Return) là thu nhập hoặc số tiền thu được từ một khoản đầu tư.

Tỷ suất lợi nhuận (Rate of return) là tỷ lệ giữa thu nhập và khoản đầu tư bỏ ra.

Lợi nhuận kỳ vọng (Expected Return) của DMĐT là lợi nhuận dự kiến của một khoản đầu tư; là giá trị trung bình có trọng số của tỷ suất sinh lời mong đợi của từng tài sản trong danh mục đầu tư [16].

$$\mathrm{E}(R_P) = \sum_{i=1}^{n} w_i E(r_i)$$

Trong đó: $\mathrm{E}(R_P)$: lợi nhuận kỳ vọng của DMĐT (P)

w_i: tỷ trọng đầu tư vào tài sản i trong DMĐT (P)

$E(r_i)$: tỷ suất sinh lời kỳ vọng của tài sản i

n: số tài sản có trong DMĐT

1.2.2 Rủi ro

Rủi ro được định nghĩa như là một mối nguy hại, nguy cơ gây ra thiệt hại và tổn thương. Vì vậy, nói đến rủi ro là nói đến khả năng những sự kiện bất lợi xảy ra. Trong đầu tư tài chính, rủi ro có khả năng mà theo đó, thu nhập mà nhà đầu tư nhận được khác với thu nhập kỳ vọng.

Rủi ro có thể được chia thành 2 loại: rủi ro chứng khoán riêng lẻ và rủi ro của DMĐT.

Rủi ro có thể phân tán được bằng cách đa dạng hóa danh mục đầu tư [22] được gọi là rủi ro phi hệ thống. Rủi ro này chỉ ảnh hưởng đến một doanh nghiệp hay một ngành nào đó do các nguyên nhân nội tại như năng lực quản lý, hay do chính sách của Chính phủ. Các nghiên cứu gần đây chỉ ra rằng nếu lựa chọn đúng đắn một danh mục chỉ khoảng 10 chứng khoán là có thể bỏ được rủi ro phi hệ thống này [25].

Rủi ro của chứng khoán riêng lẻ là phương sai và độ lệch chuẩn phản ánh tổng mức rủi ro bao gồm rủi ro hệ thống và rủi ro cá biệt của chứng khoán. Phương sai là thước đo đo lường rủi ro của chứng khoán cá biệt, phương sai được định nghĩa là bình quân gia quyền của các độ lệch chuẩn của giá trị kỳ vọng, và giúp cho ta hình dung TSSL thực tế khác với TSSL kỳ vọng như thế nào.

Phương sai: $\boldsymbol{\sigma_i^2 = \sum_{i=1}^{n}[r_i - E(r_i)]^2 . P_i}$

Độ lệch chuẩn: $\boldsymbol{\sigma_i = \sqrt{\sigma_i^2}}$

Khi chứng khoán rủi ro nằm trong DMĐT, ta sẽ thấy rằng, một chứng khoán khi nằm trong DMĐT thì sẽ ít rủi ro hơn là nó nằm riêng lẻ. Vì rủi ro có thể giảm bớt bằng cách đầu tư vào DMĐT nhiều chứng khoán có tương quan ngược chiều nhau – nghĩa là đa dạng hóa đầu tư.

Khi DMĐT bao gồm nhiều chứng khoán thì tổng rủi ro của danh mục cũng bao gồm rủi ro hệ thống và rủi ro cá biệt, do có sự bù trừ lồng ghép các hiệu ứng phân tán giữa từng cặp chứng khoán trong DMĐT. Hơn nữa, mức độ biến động lợi nhuận của DMĐT còn phụ thuộc vào mức độ tương quan hay triệt tiêu giữa các cặp chứng khoán trong danh mục.

1.3 Ứng dụng phân tích dữ liệu và học máy trong tài chính

1.3.1 Giới thiệu học máy và phân tích dữ liệu

Học máy là một nhánh của trí tuệ nhân tạo, nghiên cứu về việc xây dựng và phát triển các thuật toán cho phép máy tính học từ dữ liệu thực tế trước đó từ đó đưa ra các dự đoán, phân loại hay xây dựng các mô hình giải thích ảnh hưởng của các biến độc lập

tới biến mục tiêu [8]. Có 4 nhóm học máy chính: học có giám sát, học không giám sát, học bán giám sát và học tăng cường [8]

Trong học có giám sát, mô hình được huấn luyện trên một tập dữ liệu đã được gán nhãn, với mục tiêu là dự đoán kết quả cho dữ liệu mới [8] Các mô hình học có giám sát thường có nhiệm vụ là phân loại hay dữ báo dựa trên dữ liệu có sẵn. Một số mô hình có giám sát như : logistic regression, cây quyết định, random forest.....

Học không giám sát là việc mô hình cố gắng tìm ra cấu trúc ẩn trong dữ liệu mà không cần đến nhãn. Các mô hình học không giám sát thường được sử dụng để tìm ra các cấu trúc dữ liệu, xu hướng hay các nhóm kết quả. Các tác vụ học không giám sát thường phổ biến như: phân cụm, tìm mối quan hệ, giảm chiều, tìm liên kết,.... [8] [11].

Học bán giám sát có thể được hiểu là sự lai tạo giữa hai phương pháp học có giám sát và học không giám sát, vì nó hoạt động cả trên dữ liệu có nhãn và không có nhãn [11]. Mục tiêu cuối cùng của học bán giám sát là cho ra một kết quả tốt hơn với việc chỉ sử dụng dữ liệu được gán nhãn từ mô hình [8].

Học tăng cường là một loại học máy cho phép các thuật toán được xây dựng tự động đánh giá hành trong một bối cảnh dữ liệu cụ thể thể để cải thiện hiệu quả của công việc [19]. Mục tiêu mà phương pháp học tăng cường hướng tới đó là thu được thông tin từ việc đánh giá các sai lầm trong quá khứ từ đó tối ưu hoạt động về sau [8].

Máy học có rất nhiều ứng dụng trong cuộc sống hàng ngày từ các thuật toán gợi ý đề xuất trên các nền tảng mạng xã hội [14], hay đến phát hiện gian lận trong tín dụng [19], ứng dụng trong y tế, đặc biệt trong lĩnh vực công nghệ và trí tuệ nhân tạo có thể kể đến như: tối ưu hệ thống hoạt động của robot, lái xe tự động , sản xuất và điều khiển chuỗi cung ứng,...[8].

Phân tích dữ liệu là hoạt động đóng vai trò quan trọng trong việc giúp các nhà đầu tư đưa ra quyết định dựa trên dữ liệu, tối ưu hóa hiệu suất danh mục đầu tư và quản lý rủi ro hiệu quả. Nó cung cấp những hiểu biết có giá trị giúp nâng cao quy trình đầu tư và cải thiện kết quả.

1.3.2 Các ứng dụng học máy trong tài chính

Trong thời đại dữ liệu lớn (Big Data), học máy ngày càng trở thành công cụ then chốt trong đổi mới công nghệ tài chính (FinTech) và được xem là một đòn bẩy mạnh mẽ cho các tổ chức tài chính trong việc tối ưu hóa quy trình hoạt động, tăng khả năng cạnh

tranh và giảm thiểu rủi ro. Với khả năng xử lý khối lượng dữ liệu khổng lồ, các thuật toán học máy có thể tự động hóa các quy trình phức tạp, phát hiện các mẫu ẩn trong dữ liệu, đồng thời đưa ra những dự đoán có độ chính xác cao.

Các tổ chức tài chính thu thập dữ liệu khổng lồ từ nhiều nguồn, chẳng hạn như hồ sơ giao dịch và dữ liệu thị trường, cũng như nội dung phương tiện truyền thông xã hội và các bài báo. Đánh giá dữ liệu tài chính đáng kể bằng các thuật toán học máy cho phép các tổ chức tạo ra dự báo thị trường mang tính dự đoán, phát hiện các hoạt động bất thường và tối đa hóa việc quản lý đầu tư. Nền tảng xử lý tài liệu tự động COiN tại JPMorgan Chase [34] sử dụng xử lý ngôn ngữ tự nhiên từ ML để xem xét các tài liệu pháp lý và phát hiện thông tin quan trọng trong vài giây, trước đây đòi hỏi hàng nghìn giờ làm việc.

Dự báo thị trường và tối ưu hóa danh mục đầu tư

Học máy đã mở ra một kỷ nguyên mới trong việc xây dựng danh mục đầu tư nhờ khả năng phân tích dữ liệu đa chiều và xử lý các mối quan hệ phi tuyến tính. Thay vì chỉ dựa vào Mô hình biến động trung bình (Mean-Variance Optimization – MVO) truyền thống, các thuật toán ML như LASSO, mạng thần kinh sâu, đặc biệt là Reinforcement Learning (ML tăng cường), cho phép dự báo tỷ suất lợi nhuận, tối ưu hóa tỷ trọng tài sản và mô phỏng nhiều kịch bản thị trường một cách hiệu quả hơn [9].

Phát hiện gian lận & phân tích bất thường

Với khối lượng giao dịch khổng lồ hàng ngày, ngành tài chính phải đối mặt với các hành vi gian lận ngày càng tinh vi. Các mô hình học sâu như LSTM (Long Short-Term Memory) và mạng hồi quy (RNN) có khả năng phân tích chuỗi thời gian lớn để nhận diện mẫu bất thường trong giao dịch, phát hiện sớm các dấu hiệu gian lận [2].Các nghiên cứu thực tiễn chỉ ra rằng việc triển khai hệ thống ML để theo dõi dữ liệu thời gian thực có thể ngăn chặn các hoạt động bất thường ngay khi chúng phát sinh.

Xếp hạng tín dụng & đánh giá rủi ro

Sử dụng ML để đánh giá tín dụng mang lại khả năng đánh giá sâu rộng hơn, không chỉ dựa vào lịch sử tín dụng truyền thống mà còn khai thác dữ liệu bất cấu trúc từ giao dịch tiêu dùng, phương tiện truyền thông xã hội, và hành vi cá nhân [29]. Các

mô hình ML liên tục cập nhật và phức hợp linh hoạt hơn mô hình hồi quy phổ biến, giúp tín dụng viên đưa ra quyết định chính xác hơn, giảm rủi ro tín dụng và cải thiện khả năng tiếp cận vốn cho khách hàng.

Giao dịch thuật toán & thực thi lệnh tự động

Các tổ chức tài chính hàng đầu như JPMorgan đã phát triển hệ thống LOXM – một nền tảng học máy dựa trên lịch sử giao dịch kết hợp với thuật toán Reinforcement Learning – để tối ưu tốc độ và giá thực hiện giao dịch toàn cầu. Một nghiên cứu nội bộ của JPMorgan cho thấy hệ thống tăng tỷ lệ "win rate" từ 52% lên trên 60% [21], đồng thời giảm độ trễ xử lý lệnh từ 50 ms xuống dưới 5 ms, giúp tăng lợi nhuận và cải thiện quản trị rủi ro.

Như vậy, học máy không chỉ là một công cụ phân tích mà còn đóng vai trò chiến lược trong toàn bộ chuỗi giá trị tài chính, từ dự báo danh mục, phát hiện gian lận, đánh giá tín dụng, cho đến tự động hoá xử lý tài liệu và thực thi lệnh. Các ứng dụng này đem lại hiệu quả về mặt lợi nhuận, tiết kiệm chi phí và nâng cao tính tự động, đồng thời góp phần cải thiện tuân thủ và quản trị rủi ro.

TÓM TẮT CHƯƠNG 1

Chương 1 đã trình bày cơ sở lý thuyết về thị trường chứng khóa, danh mục đầu tư và lý thuyết về phân tích dữ liệu, làm rõ vai trò quan trọng của việc khai thác dữ liệu trong việc tối ưu hóa danh mục đầu tư hiện đại. Chương cũng đã giới thiệu tổng quan một số ứng dụng của phân tích dữ liệu trong lĩnh vực tài chính. Những nội dung này là tiền đề cho các chương tiếp theo, nơi sẽ đi sâu vào triển khai và so sánh hiệu quả của các mô hình cụ thể trong ứng dụng đầu tư.

CHƯƠNG 2: MÔ HÌNH NGHIÊN CỨU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1 Các mô hình học máy trong dự báo giá tài chính

2.1.1 Mô hình LSTM

2.1.1.1 Phương pháp Time Series

Time Series (tạm dịch là Dự báo chuỗi thời gian) là quá trình phân tích dữ liệu chuỗi thời gian sử dụng số liệu thống kê và mô hình hóa để đưa ra dự đoán và thông báo cho việc ra quyết định chiến lược [20]. Nó không phải lúc nào cũng là một dự đoán chính xác và khả năng dự báo có thể rất khác nhau - đặc biệt là khi xử lý các biến thường xuyên dao động trong dữ liệu chuỗi thời gian cũng như các yếu tố nằm ngoài tầm kiểm soát của chúng tôi. Tuy nhiên, dự báo cái nhìn sâu sắc về kết quả nào có nhiều khả năng - hoặc ít khả năng hơn - xảy ra hơn các kết quả tiềm năng khác. [1] Thông thường, dữ liệu càng toàn diện thì dự báo càng chính xác. Mặc dù dự báo và "dự đoán" thường có nghĩa giống nhau, nhưng có một điểm khác biệt đáng chú ý. Trong một số ngành, dự báo có thể đề cập đến dữ liệu tại một thời điểm cụ thể trong tương lai, trong khi dự đoán đề cập đến dữ liệu tương lai nói chung.

Dự báo chuỗi thường được sử dụng cùng với phân tích chuỗi thời gian. Phân tích chuỗi thời gian liên quan đến việc phát triển các mô hình để có được sự hiểu biết về dữ liệu để hiểu được nguyên nhân cơ bản. Phân tích có thể cung cấp "lý do" đằng sau những kết quả mà bạn đang thấy. Sau đó, dự báo sẽ thực hiện bước tiếp theo về những việc cần làm với kiến thức đó và các phép ngoại suy có thể dự đoán được về những gì có thể xảy ra trong tương lai. Có những hạn chế khi đối phó với những điều không thể đoán trước và những điều chưa biết. Dự báo chuỗi thời gian không sai và không phù hợp hoặc hữu ích cho mọi tình huống. Vì thực sự không có bộ quy tắc rõ ràng nào về thời điểm bạn nên hoặc không nên sử dụng dự báo, nên các nhà phân tích và nhóm dữ liệu phải biết các hạn chế của phân tích và những gì mô hình của họ có thể hỗ trợ. Không phải mọi mô hình sẽ phù hợp với mọi tập dữ liệu hoặc trả lời mọi câu hỏi. Nhóm dữ liệu nên sử dụng dự báo chuỗi thời gian khi họ hiểu câu hỏi kinh doanh và có dữ liệu và khả năng

dự báo thích hợp để trả lời câu hỏi đó. Dự báo tốt hoạt động với dữ liệu rõ ràng, được đóng dấu thời gian và có thể xác định các xu hướng và mẫu chính xác trong dữ liệu lịch sử. Các nhà phân tích có thể cho biết sự khác biệt giữa biến động ngẫu nhiên hoặc ngoại lệ và có thể tách những thông tin chi tiết xác thực khỏi các biến thể theo 38 mùa. Phân tích chuỗi thời gian cho biết dữ liệu thay đổi như thế nào theo thời gian và dự báo tốt có thể xác định hướng dữ liệu đang thay đổi.

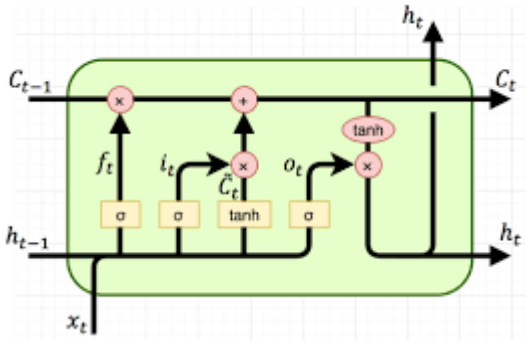
Ứng dụng của phương pháp dự báo Time Series Dự báo có một loạt các ứng dụng trong các ngành công nghiệp khác nhau. Nó có rất nhiều ứng dụng thực tế bao gồm: dự báo thời tiết, dự báo khí hậu, dự báo kinh tế, dự báo kỹ thuật dự báo chăm sóc sức khỏe, dự báo tài chính, dự báo bán lẻ, dự báo kinh doanh, dự báo nghiên cứu môi trường, dự báo nghiên cứu xã hội và hơn thế nữa.[33] Về cơ bản, bất kỳ ai có dữ liệu lịch sử nhất quán đều có thể phân tích dữ liệu đó bằng các phương pháp phân tích chuỗi thời gian và sau đó lập mô hình, dự báo và dự đoán. Đối với một số ngành, toàn bộ điểm của phân tích chuỗi thời gian là để tạo điều kiện thuận lợi cho việc dự báo. Một số công nghệ, chẳng hạn như phân tích tăng cường, thậm chí có thể tự động chọn dự báo trong số các thuật toán thống kê khác nếu nó mang lại sự chắc chắn nhất. Một số ví dụ từ một loạt các ngành để làm cho các khái niệm về phân tích chuỗi thời gian và dự báo cụ thể hơn: Dự báo giá đóng cửa của cổ phiếu mỗi ngày, dự báo doanh số bán sản phẩm theo đơn vị bán ra mỗi ngày cho một cửa hàng, dự báo thất nghiệp cho một tiểu bang mỗi quý, dự báo giá xăng trung bình mỗi ngày.

Những thứ ngẫu nhiên sẽ không bao giờ được dự báo chính xác, cho dù chúng ta thu thập bao nhiêu dữ liệu hay mức độ nhất quán. Chúng ta có thể quan sát dữ liệu hàng tuần về mọi người trúng xổ số, nhưng chúng ta không bao giờ có thể dự đoán ai sẽ thắng tiếp theo. Cuối cùng, tùy thuộc vào dữ liệu và phân tích dữ liệu chuỗi thời gian về thời điểm nên sử dụng dự báo, bởi vì dự báo rất khác nhau do các yếu tố khác nhau. Sử dụng phán đoán của bạn và biết dữ liệu của bạn.

2.1.1.2 Thuật toán LSTM

Long Short-Term Memory (LSTM) là một mô hình được đề xuất vào năm 1997 và nó chính là một loại mạng đặc biệt của RNN.[11] Đặc điểm chính là các mạng đó có

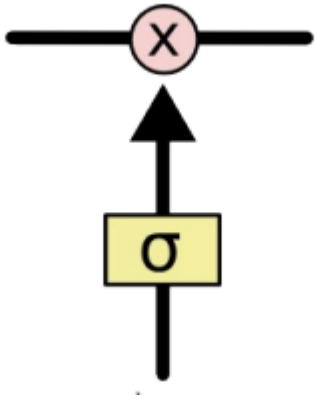
thể lưu trữ thông tin có thể được sử dụng cho quá trình xử lý cell (tạm dịch là tế bào) trong tương lai. LSTM hoạt động rất tốt trên nhiều vấn đề và hiện đang được sử dụng rộng rãi. LSTM được thiết kế rõ ràng để tránh vấn đề phụ thuộc lâu dài. Ghi nhớ thông tin trong thời gian dài thực tế là hành vi mặc định của chúng, không phải là thứ mà chúng phải vật lộn để học! Tất cả các mạng RNN đều có dạng một chuỗi các môđun lặp lại của mạng nơron. Trong các RNN tiêu chuẩn, mô-đun lặp này sẽ có cấu trúc rất đơn giản, chẳng hạn như một lớp tanh.



(Nguồn: Hochreiter và cộng sự (1997))

Hình 1.1: Cấu trúc LSTM

Ở sơ đồ trên, mỗi một đường mang một véc-tơ từ đầu ra của một nút tới đầu vào của một nút khác.[11] Các hình trong màu hồng biểu diễn các phép toán như phép cộng véc-tơ chẳng hạn, còn các ô màu vàng được sử dụng để học trong các từng mạng nơ-ron. Các đường hợp nhau kí hiệu việc kết hợp, còn các đường rẽ nhánh ám chỉ nội dung của nó được sao chép và chuyển tới các nơi khác nhau. Phần quan trọng nhất của mô hình LSTM là trạng thái tế bào - chính đường chạy thông ngang phía trên của sơ đồ hình vẽ. Trạng thái tế bào là một dạng giống như băng truyền. Nó chạy xuyên suốt tất cả các các nút mạng và chỉ tương tác tuyến tính với nhau một chút. Vì vậy mà các thông tin có thể dễ dàng truyền đi thông suốt mà không sợ bị thay đổi.

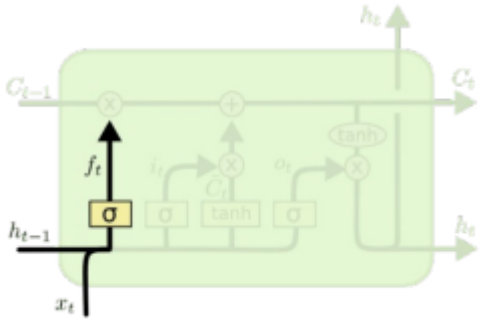


(Nguồn: Hochreiter và cộng sự (1997))

Hình 2.2: Tầng Sigmoid

Tầng sigmoid sẽ cho đầu ra là một số trong khoảng [0, 1], mô tả có bao nhiêu thông tin có thể được thông qua. Với đầu ra của tầng này là 0 thì có nghĩa là không cho thông tin nào qua cả, còn khi kết quả là 1 thì có nghĩa là cho tất cả các thông tin đi qua nó. Một LSTM gồm có 3 cổng như vậy để duy trì và điều hành trạng thái của tế bào.

Bước đầu tiên của LSTM là quyết định xem thông tin nào cần bỏ đi từ trạng thái tế bào. Quyết định này được đưa ra bởi tầng sigmoid - gọi là "tầng cổng quên" (forget gate layer) [24]. Nó sẽ lấy đầu vào là h_{t-1}và x_t rồi đưa ra kết quả là một số trong khoảng [0, 1] cho mỗi số trong trạng thái tế bào . Đầu ra là 1 thể hiện rằng nó giữ toàn bộ thông tin lại, còn 0 chỉ rằng toàn bộ thông tin sẽ bị bỏ đi.



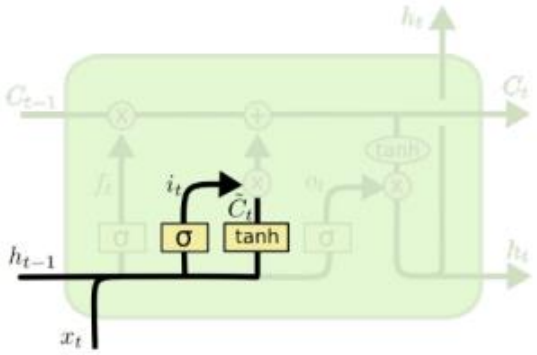
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

(Nguồn: Hochreiter và cộng sự (1997))

Hình 2.3: Hàm mô tả sigmoid

Bước tiếp theo là quyết định xem thông tin mới nào sẽ lưu vào trạng thái tế bào. Việc này gồm 2 phần. Đầu tiên là sử dụng một tầng sigmoid được gọi là "tầng cổng

vào" (input gate layer) để quyết định giá trị nào sẽ cập nhật [24]. Tiếp theo là một tầng tanh tạo ra một véc-tơ cho giá trị mới C_t nhằm thêm vào cho trạng thái. Trong bước tiếp theo, sẽ kết hợp 2 giá trị đó lại để tạo ra một cập nhật cho trạng thái.

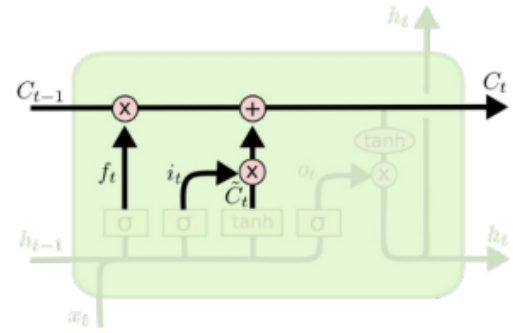


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

(Nguồn: Hochreiter và cộng sự (1997))

Hình 2.4: Mô tả hàm sigmoid kết hợp tanh

Giờ là lúc cập nhật trạng thái tế bào cũ C_{t-1}thành trạng thái mới C_t. Ở các bước trước đó đã quyết định những việc cần làm, nên giờ chỉ cần thực hiện là xong. Nhân trạng thái cũ với f_t để bỏ đi những thông tin quyết định quên lúc trước. Sau đó cộng thêm $i_t * \bar{C}_t$ .Trạng thái mới thu được này phụ thuộc vào việc quyết định cập nhập mỗi giá trị trạng thái ra sao.

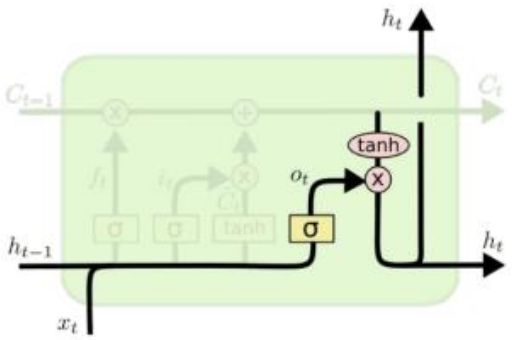


$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

(Nguồn: Hochreiter và cộng sự (1997))

Hình 2.5: Kết hợp để cho ra đầu ra

Cuối cùng, cần quyết định xem muốn đầu ra là gì. Giá trị đầu ra sẽ dựa vào trạng thái tế bào, nhưng sẽ được tiếp tục sàng lọc. Đầu tiên, chạy một tầng sigmoid để quyết định phần nào của trạng thái tế bào muốn xuất ra. Sau đó, đưa nó trạng thái tế bảo qua một hàm tanh để co giá trị nó về khoảng [-1, 1], và nhân nó với đầu ra của cổng sigmoid để được giá trị đầu ra mong muốn.



$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

(Nguồn: Hochreiter và cộng sự (1997))

Hình 2.6: Kết hợp các tầng để cho ra đầu ra là input tiếp theo của dữ liệu tiếp theo

2.1.2 Mô hình XGBoost

XGBoost là thuật toán được xây dựng dựa trên thuật toán cây quyết định tăng cường nhưng XGBoost được cải tiến và mở rộng nhằm nâng cao hiệu suất và năng lực xử lý các tập dữ liệu lớn [10]. XGBoost chọn cây quyết định làm đơn vị mô hình cơ sở và áp dụng kỹ thuật gradient boosting để tuần tự xây lên các chuỗi cây quyết định này [1].

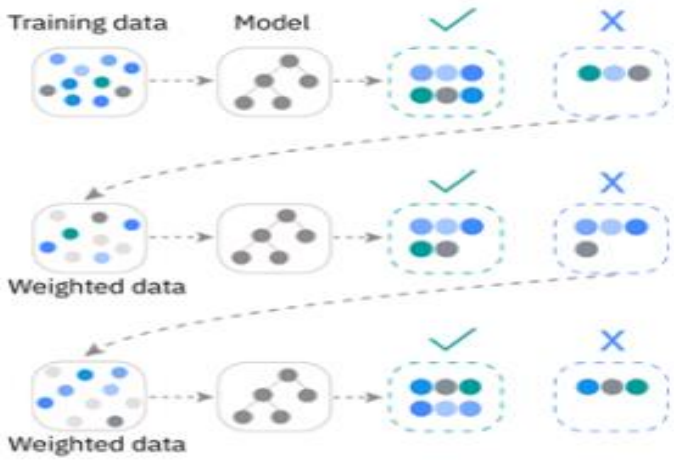
Hàm mục tiêu tổng quát của XGBoost có dạng công thức [10] [42]:

$$\text{Obj} = \sum_{i=i}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k)$$

Trong đó

- $l(y_i, \hat{y_i})$ là hàm mất mát đo sai số dự đoán với thực tế
- $\Omega(f_k)$ là hàm phạt mức độ phức tạp của cây

Điểm cốt lõi của mô hình XGBoost là giảm sai số từng chút một [10]. Thay vì tạo ra một mô hình lớn ngay từ ban đầu thì XGBoost xây một cây quyết định đơn giản đầu tiên. Tiếp đó, tại mỗi vòng lặp của mô hình thì sẽ xây thêm một cây quyết định mới. Nhiệm vụ của cây quyết định mới này không phải là trực tiếp dự đoán kết quả, mà dự đoán phần dư từ tập hợp các cây đã tạo ra trước đó.



(Nguồn: https://interdata.vn)

Hình 2.7: Cấu trúc của XGBoost

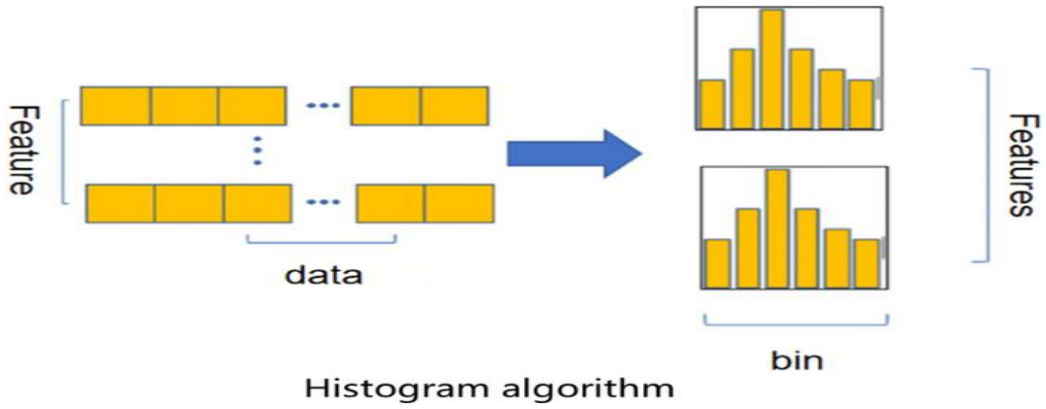
Sau mỗi vòng lặp thì mô hình sẽ được cập nhật bởi công thức:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + nf_t(x_i)$$

Trong đó:

- $\boldsymbol{f_t(x_i)}$ là cây mới được tạo ra tại thời điểm t.
- $\boldsymbol{n}$ là tốc độ học của mô hình.

Mô hình XGBoost xây các cây quyết định mới bằng kỹ thuật Level-wise Tree Growth. Khi tiến hành chia tách các lá tại cây mới này, thì XGBoost sẽ tính điểm gain (điểm lợi ích) cho sự phân tách của từng lá [10]. Những sự phân tách lá có điểm gain lớn hơn ngưỡng tối thiểu thì sẽ được phân tách. Một điểm kỹ thuật khác cũng được tích hợp trong mô hình XGBoost đó là sử dụng kỹ thuật phân lớp theo histogram để chuyển giá trị liên tục thành các "bin" rời rạc trong việc huấn luyện và xử lý các tập dữ liệu [4] [8].



(Nguồn: https://www.researchgate.net)

Hình 2.8: Nguyên lý hoạt động của Histogram Algorithm trong XGBoost

Sau đó gán giá trị vào mỗi một bin tương ứng và mô hình XGBoost sẽ tính toán trên các giá trị bin này thay vì tính toán trên giá trị thực [4]. Điều này giúp mô hình nhanh hơn trong việc xử lý dữ liệu lớn, giảm thời gian huấn luyện của mô hình, giảm tiêu thụ bộ nhớ và đặc biệt ổn định hơn trong việc tính toán do binning giúp làm trơn dữ liệu làm mô hình ít bị nhiễu bởi giá trị ngoại lai [4] [8] [13].

Chiến lược phát triển cây của mô hình XGBoost là xây dựng cây đến một độ sâu tối đa được chỉ định bởi tham số siêu điều chỉnh "max depth", sau đó bắt đầu cắt tỉa cây theo chiều ngược lại, loại bỏ các điểm phân chia không mang lại lợi ích tăng trưởng tích cực [4]. Cây quyết định mới được huấn luyện sao cho góp phần giảm thiểu hiệu quả nhất giá trị lỗi của toàn bộ mô hình tính tới thời điểm cây quyết định mới này được tạo ra [36]. Kết quả dự đoán cuối cùng là sự kết hợp có trọng số của tất cả cây quyết định đã được xây dựng trong suốt quá trình [1] và được thể hiện bởi công thức.

$$\widehat{y_i} = \sum_{k=1}^{K} nf_k(x_i)$$

Điểm khác biệt so với Gradient Boosting cũ là XGBoost không chỉ hỗ trợ giảm sai số, mà khi một cây quyết định mới được tạo ra, XGBoost còn cân nhắc đến tính phức tạp của cây mới này như độ sâu của cây hay có bao nhiêu nhánh,... bằng cách thêm hàm phạt vào hàm mục tiêu. Điều này nhằm ngăn ngừa hiện tượng overfitting từ mô hình.

Hàm phạt trong XGBoost có dạng công thức [42]:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} {w_j}^2$$

Trong công thức trên T là số lượng lá trong cây quyết định mới f_k được tạo ra. Mỗi một lá này đại diện cho một nhóm dữ liệu đầu vào có chung một kết quả dự đoán. Càng nhiều lá thì cây mới f_k tạo ra sẽ các phức tạp và xuất hiện hiện tượng overfitting. Tham số phạt số lượng lá γ được thêm vào để hạn chế độ phức tạp của cây f_k mới [42]. Khi XGBoost tính toán để chia tách một nhánh mới cho cây f_k vừa tạo ra này, XGBoost sẽ tính toán lợi ích của việc chia tách này, giá trị lợi ích này gọi là điểm Gain [10]. Nếu điểm Gain này nhỏ hơn tham số phạt số lượng lá γ thì mô hình XGBoost sẽ không chia nhánh mới cho cây f_k. Điều này giúp hạn chế các nhánh vô nghĩa trong mô hình giúp mô hình đơn giản và tập trung vào những nhánh thật sự cần thiết.

W_j là giá trị dự đoán tối ưu tại lá thứ j trong cây f_k mới [42]. Và $\sum_{j=1}^{T} w_j{}^2$ là tổng bình phương tất cả các giá trị dự đoán của tất cả các lá trong cây f_k . Giá trị λ là tham số phạt độ lớn của giá trị dự đoán. Khi tạo ra một cây mới thì mô hình XGBoost luôn tìm ra giá trị tối ưu cho từng lá cây. Tuy nhiên khi các giá trị này quá lớn thì mô hình đang quá "khắt khe" với dữ liệu huấn luyện, điều này dẫn tới sai số lớn trong việc huấn luyện mô hình tại tập dữ liệu thử nghiệm. Giá trị λ được thêm vào để kiểm soát độ mạnh độ mạnh yếu của dự đoán tại từng lá cây bằng việc phạt các giá trị dự đoán quá lớn, điều này làm tăng tính tổng quát của XGBoost, tránh cho việc mô hình quá khắt khe trong việc học tập trên dữ liệu huấn luyện [42].

2.1.3 Mô hình Decision Tree

Cây quyết định (decision tree) là một phương pháp rất mạnh và phổ biến cho cả hai nhiệm vụ của khai phá dữ liệu là phân loại và dự báo [44]. Mặt khác, cây quyết định còn có thể chuyển sang dạng biểu diễn trong đường duy nhất từ một thuật lẽ các luật If-Then.

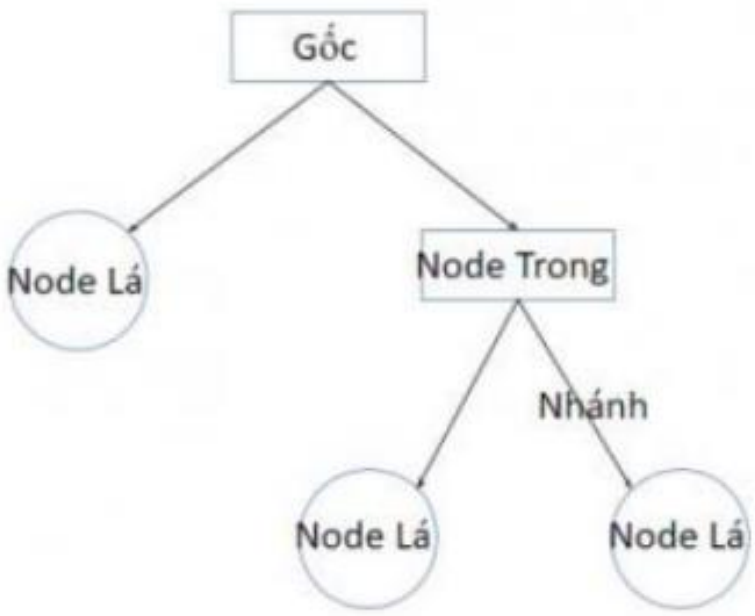
Cây quyết định là cấu trúc biểu diễn dưới dạng cây. Trong đó, nút nội trong (internal node) biểu diễn một thuộc tính, nhánh (branch) biểu diễn giá trị có thể có của thuộc tính, nút lá (leaf node) biểu diễn lớp quyết định (dự đoán) tương ứng của cây quyết định. Cây quyết định có thể được dùng để phân lớp bằng cách xuất phát từ gốc của cây và đi xuyên theo các nhánh cho đến khi gặp nút lá. Trên cơ sở phân lớp này chúng ta có thể chuyển đổi về các luật quyết định.

Cây quyết định được sử dụng để xây dựng một kế hoạch nhắm đạt được mục tiêu mong muốn. Các cây quyết định được dùng để hỗ trợ quá trình ra quyết định. Cây quyết định là một dạng đặc biệt của cấu trúc cây.

Tạo cây quyết định chính là quá trình phân tích cơ sở dữ liệu, phân lớp và đưa ra dự đoán. Cây quyết định được tạo thành bằng cách lần lượt chia (tẻ quy) một tập dữ liệu thành các tập dữ liệu con, mỗi tập con được tạo thành chủ yếu từ các phần tử của cùng một lớp [30]. Lựa chọn thuộc tính để tạo nhánh thông qua Entropy và Gain.

Bởi các cây quyết định mạnh mẽ là một phương pháp thông dụng trong khai phá dữ liệu. Khi đó, cây quyết định mô tả một cấu trúc cây, trong đó, các lá đại diện cho các phân loại, còn cành đại diện cho các kết hợp của các thuộc tính dẫn tới phân loại đó. Một cây quyết định có thể được học bằng cách chia tập hợp nguồn thành các tập con theo thứ tự một kiểm tra giá trị thuộc tính. Quá trình này được lặp lại một cách đệ quy cho mỗi tập con đầu xuất. Quá trình đệ qui hoàn thành khi không thể tiếp tục thực hiện việc chia tách được nữa, hay khi một phân loại đơn có thể áp dụng cho từng phần tử của tập con dẫn xuất.

Cây hồi quy (Regression tree): ước lượng các hàm giá có giá trị là số thực thay vì được sử dụng cho các nhiệm vụ phân loại.



(Nguồn: Kamiński B và cộng sự (2017))

Hình 2.9: Mô tả cây quyết định

Quá trình xây dựng một cây quyết định thường được thực hiện như sau:

(1) Tạo nút gốc cho cây quyết định, nơi biểu diễn tất cả các mẫu của tập dữ liệu.

(2) Tại lớp đang xem xét, nếu tất cả các mẫu thuộc về cùng một lớp đó, nút đang xét sẽ trở thành nút lá và được gán nhãn chính bằng lớp đó.

(3) Ngược lại, dùng độ đo thuộc tính nào đó để chọn thuộc tính sẽ phân tách các mẫu tốt nhất vào các lớp tương ứng.

(4) Một nhánh được tạo ra cho từng giá trị của thuộc tính được chọn.

(5) Tiếp tục quá trình trên để tạo ra các nút mới, nhánh mới.

(6) Tiến trình kết thúc chỉ khi bất kỳ điều kiện nào sau đây là đúng:

- Tất cả các mẫu của một nút cho trước đều thuộc về cùng một lớp.
- Không còn thuộc tính nào mà mẫu có thể dựa vào để phân hoạch xa hơn.
- Không còn mẫu nào cho nhánh.

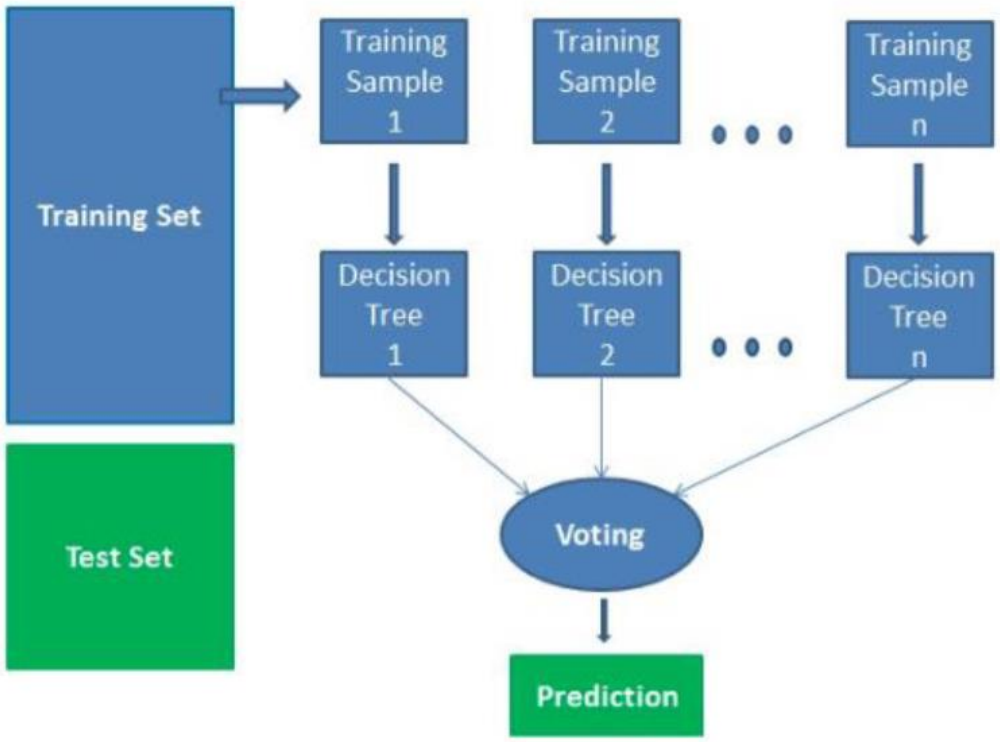
• Tuy nhiên, nếu chúng ta không lựa chọn được thuộc tính nào để phân loại hợp lý tại mỗi nút, cây quyết định sau khi xây dựng có thể rất phức tạp. Vì thế người ta thường sử dụng hai cách sau để xây dựng cây quyết định phù hợp:

• Dừng việc phát triển cây sớm hơn bình thường trước khi phân lớp hoàn toàn tập dữ liệu huấn luyện.

• Sử dụng một số kỹ thuật "cắt", "tỉa" cây phù hợp

2.1.4 Mô hình Random Forest

Random Forest (Rừng ngẫu nhiên) [26] là một thuật toán học có giám sát. Như bạn có thể thấy từ tên của nó, nó tạo ra một khu rừng một cách ngẫu nhiên. "Khu rừng" mà ta có thể tạo ra là một tập hợp các cây quyết định. Ý tưởng chính của phương pháp này là sự kết hợp của các mô hình học tập làm tăng kết quả chung.



(*Nguồn: Internet)*

Hình 2.10: Thuật toán rừng ngẫu nhiên

Rừng ngẫu nhiên được đề xuất vào năm 2001 [45]. Đây là thuật toán phân loại có kiểm định dựa trên cây quyết định và kỹ thuật Bagging and Bootstrapping đã được cải tiến. Bootstrapping là một phương pháp rất nổi tiếng trong thống kê được giới thiệu bởi Efron vào năm 1979. Phương pháp này được thực hiện như sau: từ một quần thể ban đầu lấy ra một mẫu L = (x1, x2,..,xn) gồm n thành phần để tính toán các tham số mong muốn. Trong các bước tiếp theo lặp lại b lần tạo ra mẫu Lb cũng gồm n phần bằng cách lấy lại mẫu với sự thay thế các thành phần trong mẫu ban đầu sau đó tính toán các tham số mong muốn. Phương pháp Bagging được xem như là một phương pháp tổng hợp kết quả có được từ các bootstraping sau đó huấn luyện mô hình từ các mẫu ngẫu nhiên này và cuối cùng đưa ra dự đoán phân loại dựa vào số phiếu bầu cao nhất của lớp phân loại. Cây quyết định là một sơ đồ phát triển có cấu trúc dạng cây phân nhánh đi từ gốc cho đến lá, giá trị các lớp phân loại của mẫu được đưa vào kiểm tra trên cây quyết định. Mỗi mẫu tương ứng có một đường đi từ gốc (tức là dữ liệu đầu vào) đến lá (tức là các kết quả phân loại dự đoán đầu ra), đường đi này biểu diễn sự phân lớp của mẫu đó. Mỗi sơ đồ cây trong tập mẫu được tạo thành từ tập hợp các dữ liệu huấn luyện được lựa chọn ngẫu nhiên để huấn luyện mô hình phân loại Rừng ngẫu nhiên (mỗi tập mẫu bootstrap sẽ cho ra một cây và n cây tương ứng với n bootstrap). Khi một tập mẫu được rút ra từ tập huấn luyện (bootstrap) với sự thay thế có hoàn lại, thì thông thường có khoảng 1/3 các phần tử không nằm trong mẫu này và vì thế chúng không tham gia vào quá trình huấn. Điều n ày có nghĩa là chỉ có khoảng 2/3 các phần tử trong tập huấn luyện tham gia vào trong các tính toán để phân loại và 1/3 các phần tử này dùng để kiểm tra sai số. Dữ liệu kiểm tra được sử dụng để ước lượng sai số tạo ra từ việc kết hợp các kết quả phân loại riêng lẻ sau đó được tổng hợp trong mô hình Rừng ngẫu nhiên cũng như dùng để ước tính các biến quan trọng.

Rừng ngẫu nhiên chứa một lượng lớn các cây, mỗi cây được phát triển từ các tập huấn luyện được lựa chọn ngẫu nhiên. Hai tham số cần được xác định trong thuật toán phân loại này là ntree (số lượng cây được phát triển) và mtry (số lượng biến để phân chia tại mỗi node). Số ntree được lựa chọn phụ thuộc vào khoảng thời gian xử lý ngắn nhất để kết quả đạt được độ sai số thấp nhất và mtry biến động từ số biến độc lập tối thiểu (bằng 1) đến số biến độc lập tối đa được sử dụng trong phân loại. Sau khi mô hình Random Forest được tạo thành, mỗi kết quả của các boostrap trong tập hợp sẽ bỏ phiếu

cho lớp phổ biến nhất và cho ra một kết quả phân loại. Mô hình được tạo thành dựa vào phân loại có số phiếu bầu nhiều nhất của mỗi sơ đồ cây quyết định ntree.

2.2 Lý thuyết tối ưu danh mục đầu tư

2.2.1 Mean – Variance Optimization (Harry Markowitz)

Lý thuyết Tối ưu hóa danh mục đầu tư theo phương pháp Trung bình – Phương sai (Mean – Variance Optimization - MVO), do nhà kinh tế học Harry Markowitz công bố lần đầu vào năm 1952 trong bài báo "Portfolio Selection" (Lựa chọn danh mục đầu tư) trên Tạp chí Tài chính, đã đặt nền móng vững chắc cho lĩnh vực tài chính hiện đại và lý thuyết danh mục đầu tư. Mô hình này được xây dựng dựa trên một giả định cốt lõi rằng các nhà đầu tư là những cá nhân duy lý, tìm kiếm sự đánh đổi hiệu quả giữa lợi nhuận kỳ vọng và rủi ro: họ luôn mong muốn tối đa hóa tỷ suất sinh lời mong đợi cho một mức độ rủi ro nhất định mà họ sẵn lòng chấp nhận, hoặc ngược lại, tối thiểu hóa rủi ro cho một mức lợi nhuận kỳ vọng mong muốn [38].

Trong khuôn khổ của lý thuyết MVO, "lợi nhuận kỳ vọng" của một danh mục đầu tư được định nghĩa là tỷ suất sinh lời trung bình có trọng số của các tài sản cấu thành, trong đó trọng số là tỷ trọng đầu tư vào từng tài sản. Mặt khác, "rủi ro danh mục" được đo lường bằng phương sai (variance) hoặc độ lệch chuẩn (standard deviation) của tỷ suất sinh lời danh mục [38]. Phương sai phản ánh mức độ biến động của tỷ suất sinh lời danh mục quanh mức lợi nhuận kỳ vọng của nó, với phương sai lớn cho thấy rủi ro cao hơn và ngược lại. Nhà đầu tư, theo Markowitz, là những người ác cảm với rủi ro (risk-averse) [17] tức là, với hai danh mục đầu tư có cùng mức lợi nhuận kỳ vọng, họ sẽ chọn danh mục có rủi ro thấp hơn; hoặc với hai danh mục có cùng mức rủi ro, họ sẽ chọn danh mục có lợi nhuận kỳ vọng cao hơn.

Bằng việc kết hợp nhiều tài sản có độ tương quan khác nhau, nhà đầu tư có thể giảm thiểu rủi ro danh mục mà không nhất thiết phải giảm return mong muốn. Đây chính là lời ích vượt trội của đa dạng hóa danh mục trong lý thuyết của Markowitz.

2.2.2 Hiểu về Effcient Frontier, Risk-free Asset

Để xây dựng Đường biên hiệu quả, nhà đầu tư cần xác định lợi nhuận kỳ vọng, phương sai (rủi ro), và hiệp phương sai (covariance) hoặc hệ số tương quan giữa các cặp

tài sản trong danh mục [38]. Dựa trên các thông số này, vô số các danh mục đầu tư có thể được hình thành. Tuy nhiên, chỉ một số danh mục nhất định được coi là "hiệu quả". Một danh mục được xem là hiệu quả khi nó thỏa mãn một trong hai tiêu chí sau:

Tối đa hóa lợi nhuận kỳ vọng: Đối với một mức độ rủi ro (phương sai/độ lệch chuẩn) nhất định, danh mục đó mang lại tỷ suất sinh lời kỳ vọng cao nhất có thể.

Tối thiểu hóa rủi ro: Đối với một tỷ suất sinh lời kỳ vọng mong muốn, danh mục đó có mức độ rủi ro (phương sai/độ lệch chuẩn) thấp nhất có thể.

Đồ thị của Đường biên hiệu quả thường có dạng parabol hoặc hình cánh cung quay về phía trên bên trái trong không gian lợi nhuận – rủi ro (trục tung là lợi nhuận kỳ vọng, trục hoành là độ lệch chuẩn rủi ro). Mỗi điểm trên đường cong này đại diện cho một danh mục đầu tư hiệu quả. Các danh mục nằm phía dưới đường biên được coi là không hiệu quả (inefficient) vì chúng mang lại lợi nhuận thấp hơn hoặc rủi ro cao hơn so với một danh mục hiệu quả khác. Các danh mục nằm phía trên đường biên được coi là không thể đạt được (unattainable) với tập hợp tài sản hiện có.

Ý nghĩa quan trọng của Đường biên hiệu quả là nó cho phép nhà đầu tư trực quan hóa sự đánh đổi giữa lợi nhuận và rủi ro. Nhà đầu tư có thể chọn một danh mục trên đường biên tùy thuộc vào mức độ chấp nhận rủi ro cá nhân. Những nhà đầu tư chấp nhận rủi ro thấp sẽ chọn danh mục gần điểm có rủi ro tối thiểu (minimum variance portfolio), trong khi những nhà đầu tư chấp nhận rủi ro cao hơn có thể chọn danh mục có lợi nhuận kỳ vọng cao hơn nhưng cũng đi kèm với rủi ro lớn hơn [15].

2.3 Tối ưu tham số và đánh giá mô hình

Tham số điều chuẩn (Regularization)

- Điều chuẩn L1 (reg_alpha): Thêm hình phạt L1 (tổng giá trị tuyệt đối của các trọng số) vào hàm mục tiêu. L1 có xu hướng làm triệt tiêu một số trọng số bằng cách cố gắng đưa chúng về 0. Điều này dẫn tới việc một số lá cây hoặc nhánh cây sẽ bị loại bỏ, mô hình sẽ chọn lọc được các đặc trưng chính (hiệu ứng lựa chọn đặc trưng implicit), các cây sẽ thưa hơn và hạn chế được hiện tượng overfiting.
- Điều chuẩn L2 (reg_lambda): Thêm hình phạt L2 (tổng bình phương trọng số) vào hàm mục tiêu. Điều chuẩn L2 làm giảm độ lớn của trọng số của tất cả các lá cây mà

không triệt tiêu hoàn toàn một lá nào. Điều này giúp làm mịn mô hình và cũng hạn chế được hiện tượng overfitting.

Chiến lược lấy mẫu

- Lấy mẫu theo hàng (Row Subsampling / Subsample): Xác định tỷ lệ phần trăm các mẫu từ tập huấn luyện được sử dụng ngẫu nhiên cho mỗi cây. Điều này làm tăng tính ngẫu nhiên và tính đa dạng giữa các cây, giúp giảm overfitting – tương tự như kỹ thuật bagging.
- Lấy mẫu theo cột (Column Subsampling / Feature Fraction): Xác định tỷ lệ đặc trưng được lựa chọn ngẫu nhiên khi xây dựng cây. Lấy mẫu theo cột giúp mô hình ít phụ thuộc vào các đặc trưng nhiễu hoặc tương quan cao, đồng thời tăng sự đa dạng trong tập hợp cây.

Cơ chế dừng sớm (Early Stopping Rounds)

- Kỹ thuật này giám sát hiệu suất của mô hình trên tập kiểm định (validation) trong quá trình huấn luyện. Nếu không có cải thiện đáng kể sau một số vòng xác định, quá trình huấn luyện sẽ dừng lại nhằm tránh overfitting và tiết kiệm thời gian huấn luyện.

2.3.1. Kỹ thuật tối ưu siêu tham số

Để các mô hình boosting hay bất kì một mô hình phân loại nào có độ chính xác cao nhất thì việc điều chỉnh các siêu tham số để có mô hình tốt nhất thì là một điều cần thiết. Tuy nhiên việc tính toán và tìm ra các siêu tham số này rất phức tạp và khó khăn đối với các nhà phân tích dữ liệu hay phân tích kinh tế. Có hai phương pháp phổ biến và dễ dàng sử dụng để điều chỉnh siêu tham số là Bayesian Optimization và Grid Search.

Bayesian Optimization

Bayesian Optimization (BO) là một thuật toán tối ưu hóa dựa trên lý thuyết Bayes, đặc biệt hữu ích trong việc lựa chọn siêu tham số cho các mô hình học máy. Thay vì thử nghiệm ngẫu nhiên hoặc sử dụng phương pháp lưới tìm kiếm vốn tốn kém về mặt tính toán, BO xây dựng một mô hình xác suất nhằm dự đoán hiệu suất của các siêu tham số khác nhau và liên tục cập nhật mô hình này để tìm kiếm tối ưu [23]. Thuật toán hoạt động bằng cách xây dựng một mô hình thay thế, thường là Gaussian Process, để ước lượng hàm mục tiêu. Dựa vào mô hình này, BO sử dụng một hàm thu hoạch (acquisition function) nhằm quyết định điểm nào sẽ được đánh giá tiếp theo, giúp tối đa hóa hiệu suất mà không cần thử nghiệm quá nhiều điểm. Trong việc lựa chọn siêu tham số, BO

đặc biệt mạnh mẽ vì nó có thể tìm ra giá trị tối ưu với số lần thử nghiệm ít hơn, giúp tiết kiệm tài nguyên tính toán mà vẫn đảm bảo chất lượng mô hình cao.

Grid Search

Grid Search là một phương pháp tìm kiếm vét cạn trong một tập con xác định trước của không gian siêu tham số [16]. Các siêu tham số được xác định bằng cách chỉ rõ giá trị tối thiểu (cận dưới), giá trị tối đa (cận trên) và số bước chia (steps) trong khoảng này. Thuật toán sẽ huấn luyện mô hình với từng tổ hợp siêu tham số được sinh ra từ lưới tìm kiếm, sau đó đánh giá hiệu suất của từng mô hình bằng các kỹ thuật như cross-validation với các chỉ số đánh giá hiệu suất thường sử dụng độ chính xác [16]. Cuối cùng, tổ hợp siêu tham số mang lại hiệu suất tốt nhất sẽ được chọn là cấu hình tối ưu cho mô hình.

Ưu điểm của Grid Search là chắc chắn sẽ tìm ra được tổ hợp tối ưu, bởi thuật toán này sẽ duyệt toàn bộ tổ hợp có thể, kết quả của GridSearch cũng ổn định và dễ tái lập khi sử dụng cùng dữ liệu và thông số [8]. Tuy nhiên Gird Search lại có chi phí tính toán quá cao. Do sử dụng phương pháp vét cạn để tìm kiếm thuật toán này không phù hợp với không gian tìm kiếm lớn do tổ hợp siêu tham số quá nhiều, Gird Search lại không học hỏi từ kết quả trước để điều chỉnh chiến lược tìm kiếm dẫn tới lãng phí tài nguyên khi tính toán [16].

Trong nghiên cứu này, tác giả sử dụng kỹ thuật Grid Search để tìm kiếm bộ tham số tối ưu. Với ưu điểm đơn giản, dễ triển khai, phù hợp với quy mô dữ liệu của nghiên cứu, đặc biệt là có thể tái sử dụng kết quả thì Grid Search được đánh giá là lựa chọn phù hợp trong cho việc tìm các tham số của mô hình nghiên cứu.

2.3.2 Các chỉ số đánh giá hiệu quả mô hình dự báo

Trong bài toán dự báo giá chứng khoán, việc đánh giá độ chính xác và hiệu quả của mô hình là bước quan trọng nhằm xác định khả năng áp dụng vào thực tiễn đầu tư. Dưới đây là các chỉ số được sử dụng phổ biến trong việc đánh giá các mô hình hồi quy dự báo:

MAE (Mean Absolute Error) - Sai số trung bình tuyệt đối

Trong thống kê, sai số tuyệt đối trung bình (MAE) là một thước đo sai số giữa các quan sát được ghép nối biểu hiện cùng một hiện tượng. Ví dụ về Y so với X bao gồm so sánh dự đoán so với quan sát, thời gian tiếp theo so với thời điểm ban đầu và một kỹ thuật đo lường so với một kỹ thuật đo lường thay thế.

MAE đo lường sai lệch trung bình tuyệt đối giữa giá trị thực tế và giá trị dự báo: MAE dễ tính toán, không bị ảnh hưởng mạnh bởi outlier.[10] MAE được tính như sau:

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^{n} |y_t - \hat{y}_t|$$

RMSE (Root Mean Squared Error) - Sai số trung bình bình phương

Root Mean Square Error (RMSE) là độ lệch chuẩn của các phần dư (sai số dự đoán). Phần dư là thước đo khoảng cách từ các điểm dữ liệu đường hồi quy; RMSE là thước đo mức độ lan tỏa của những phần dư này. Nói cách khác, nó cho bạn biết mức độ tập trung của dữ liệu xung quanh dòng phù hợp nhất. Sai số bình phương trung bình gốc thường được sử dụng trong khí hậu học, dự báo và phân tích hồi quy để xác minh kết quả thực nghiệm.

RMSE là chỉ số đặc biệt quan trọng vì nhấn mạnh các sai số lớn: Giá trị RMSE cao cho thấy dự báo lệch nhiều so với thực tế. Trong ngành tài chính, RMSE càng thấp chứng tỏ mô hình dự báo ổn định hơn trong dữ liệu thời gian.[10]

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n} e_t^2}{n}}$$

MSE (Mean Squared Error)

MSE là gốc của RMSE, cung cấp giá trị trung bình của chênh lệch bình phương giữa tham số dự đoán và tham số quan sát được.[10]

$$\text{MSE} = \frac{1}{n} \sum_{t=1}^{n} (y_t - \hat{y}_t)^2$$

Trong thống kê, sai số bình phương trung bình (MSE) hoặc độ lệch bình phương trung bình (MSD) của một công cụ ước lượng (của một thủ tục ước tính một đại lượng không được quan sát) đo giá trị trung bình của các bình phương của các lỗi - nghĩa là, sự khác biệt bình phương trung bình giữa các giá trị ước tính và giá trị thực tế. MSE là một hàm rủi ro, tương ứng với giá trị kỳ vọng của tổn thất sai số bình phương. Thực tế là MSE hầu như luôn luôn dương (và không phải bằng 0) là do ngẫu nhiên hoặc do công cụ ước lượng không tính đến thông tin có thể đưa ra ước tính chính xác hơn

2.3.3 Các chỉ số đánh giá danh mục đầu tư

Khi đánh giá hiệu quả của một danh mục đầu tư, các nhà đầu tư, quản lý danh mục và nhà nghiên cứu tài chính thường sử dụng một số chỉ số cố điển như Sharpe Ratio**,** Return và Volatility**.** Đây là những thông số cốt lõi giúp đánh giá mức độ hiệu quả, đồng thời làm căn cứ cho việc ra quyết định đầu tư trong môi trường biến động của thị trường.

a. Return (Tỷ suất sinh lời)

Return là chỉ số phổ biến nhất để đo lường mức sinh lời của danh mục đầu tư trong một khoảng thời gian nhất định. Tỷ suất sinh lời có thể tính dựa trên giá trị ban đầu và giá trị kết thúc của danh mục [5]:

$$\text{Return} = \frac{V_{cuối} - V_{đầu}}{V_{đầu}} \text{ x } 100\%$$

Trong đó:

- $V_{cuối}$ **: giá trị danh mục tại cuối kỳ**;
- $V_{đầu}$ **: giá trị danh mục tại đầu kỳ**

Tỷ suất sinh lời dương cho thấy danh mục tạo ra lợi nhuận; ngược lại, giá trị âm thể hiện lỗ vốn. Đối với danh mục chứa nhiều tài sản, return thường được tính theo phương pháp trung bình gia quyền dựa trên tỷ trọng phân bổ vào từng tài sản. Return có thể tính theo chu kỳ ngày, tháng hoặc năm, tùy thuộc vào yêu cầu của nhà đầu tư hoặc mục tiêu nghiên cứu.

b. **Volatility (Biến động)**

Volatility là thước đo độ rủi ro của danh mục, đặc trưng cho mức dao động của tỷ suất sinh lời quanh giá trị trung bình. Nó được tính dựa trên độ lệch chuẩn (standard deviation) của return [37]:

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (R_i - \bar{R})^2}$$

Trong đó:

- R_i**: tỷ suất sinh lời tại thời điểm y;**
- $\bar{R}$**: lợi nhuận trung bình**

Volatility biến động cao cho thấy danh mục có dao động giá lớn, tương đồng với rủi ro cao hơn. Trong ngành tài chính, volatility được xem là cách tiếp cận định lượng rủi ro giá tốt nhất. Trong nghiên cứu của Markowitz (1952), volatility là yếu tố trung tâm trong mô hình danh mục tối ưu, giúp cân bằng giữa return và rủi ro.

c. Sharpe Ratio (tỷ suất sinh lời điều chỉnh theo rủi ro)

Sharpe Ratio là chỉ số đánh giá hiệu quả điều chỉnh rủi ro của danh mục đầu tư. Chính là tỷ suất sinh lời vượt trội (so với lãi suất phi rủi ro) chia cho độ biến động của return [58]:

$$\text{Sharpe Ratio} = \frac{R_p - R_f}{\sigma_p}$$

Trong đó:

- **R_p: lợi nhuận kỳ vọng của danh mục;**
- **R_f: lãi suất phi rủi ro;**
- **σ_p: độ lệch chuẩn của return (Volatility)**

Chỉ số Sharpe Ratio cho biết mỗi đơn vị rủi ro tương ứng với bao nhiêu đơn vị lợi nhuận vượt trội. Chỉ số Sharpe Ratio cao chứng tỏ danh mục tối ưu hơn về tỷ suất sinh lời theo rủi ro. Các nhà quản lý danh mục chính quyết định duy trị các danh mục với Sharpe Ratio > 1 là đã tốt, > 2 là rất tốt, và > 3 là xuất sắc [18].

TÓM TẮT CHƯƠNG 2

Chương 2 đã trình bày về nguyên lý hoạt động của 3 mô hình XGBoost, LSTM và Decision Tree, đây là các mô hình học máy mạnh mẽ trong các bài toán dự báo và ra quyết định. Mỗi mô hình đều được phân tích rõ nguyên lý hoạt động, cơ chế huấn luyện, cách tối ưu hóa hàm mất mát, và những cải tiến đặc thù nhằm tăng hiệu suất và khả năng tổng quát hóa. Bên cạnh đó, chương này cũng giới thiệu mô hình tối ưu danh mục là Mean Variance Optimization. Việc hiểu rõ bản chất và cách áp dụng của hai kỹ thuật này có ý nghĩa quan trọng trong việc tối ưu danh mục đầu tư và lựa chọn cổ phiếu nhằm nâng cao hiệu quả của hoạt động đầu tư.

CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

3.1 Dữ liệu và quy trình thực nghiệm

3.1.1 Mô tả dữ liệu

Dữ liệu gồm giá đóng, mở cửa và khối lượng của 10 mã cổ phiếu: VCB, FPT, VNM, MWG, GAS, HPG, VJC, SSI, VHM, SAB.

Bộ dữ liệu có: 20201 hàng và 7 cột
Dữ liệu không có n/a, Dữ liệu không có dòng trùng lặp.
Thống kê 10 dòng đầu của bộ dữ liệu (mã FPT):

	time	open	high	low	close	volume	symbol
12273	2017-03-29	12120	12220	12070	12150	605170	FPT
12274	2017-03-30	12150	12240	12050	12050	752470	FPT
12275	2017-03-31	12050	12200	12020	12100	1213360	FPT
12276	2017-04-03	12100	12150	11870	11870	1994720	FPT
12277	2017-04-04	11850	12020	11850	12020	981010	FPT
12278	2017-04-05	12020	12020	11900	11900	598970	FPT
12279	2017-04-07	11850	12030	11850	12020	900280	FPT
12280	2017-04-10	12020	12080	11990	12020	421960	FPT
12281	2017-04-11	12020	12100	11970	12010	405130	FPT
12282	2017-04-12	12020	12210	12010	12160	1296510	FPT

(Nguồn: Dữ liệu thu thập từ tác giả)

Hình 3.1: Mô tả dữ liệu

Từ kết quả mô tả hình ảnh 3.1 cho thấy bộ dữ liệu nghiên cứu không có giá trị thiếu và dữ liệu trùng lặp. Điều này đảm bảo cho tính toàn vẹn và độ tin cậy của dữ liệu đầu vào, giúp cho quá trình xử lý và huấn luyện mô hình không bị sai lệch do các giá trị bất thường hoặc không đầy đủ. Đồng thời, việc không phải thực hiện các bước xử lý thiếu dữ liệu cũng góp phần tiết kiệm thời gian xử lý và hạn chế rủi ro phát sinh từ các kỹ thuật nội suy hoặc loại bỏ quan sát.

3.1.2 Thống kê dữ liệu

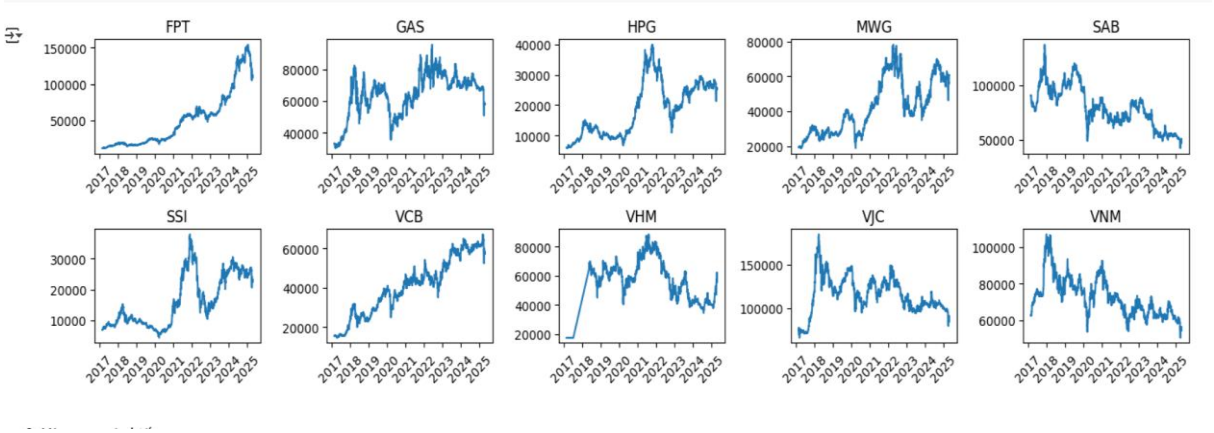
Bảng 3.1: Kết quả mô tả thống kê dữ liệu

Stock	Mean	Std.	Maximum	Minimum
FPT	50,502.06	37,953.73	154,3	11,67
GAS	64,456.70	12,869.63	95,47	30,73
HPG	18,766.74	8,790.85	39,9	6,07
MWG	42,569.82	15,319.15	78,2	18,8
SAB	79,541.99	18,537.66	136,76	42,6
SSI	16,269.84	8,353.40	37,93	4,25
VCB	40,867.32	14,190.08	67,3	14,52
VHM	58,132.90	12,558.42	88,42	34,5
VJC	115,381.23	20,467.03	184,84	70,24
VNM	72,857.98	10,637.00	107,02	50,22

(Nguồn: Tác giả tính toán từ dữ liệu thu thập)

Kết quả thống kê mô tả cho thấy các chỉ số giá cổ phiếu trung bình (Mean), độ lệch chuẩn (Std.), giá cao nhất (Maximum) và giá thấp nhất (Minimum) của 10 mã cổ phiếu thuộc nhóm cổ phiếu lớn trên thị trường chứng khoán Việt Nam được lựa chọn đều có mức giá tốt. Cổ phiếu VJC ghi nhận mức giá trung bình cao nhất là 115.381,23 đồng, đồng thời cũng có độ biến động lớn với độ lệch chuẩn là 20.467,03 đồng và giá

tối đa lên đến 184,84 đồng. Ngược lại, SSI là mã có giá trung bình thấp nhất trong nhóm, ở mức 16.269,84 đồng, cho thấy mức vốn hóa nhỏ hơn so với các cổ phiếu còn lại. SAB và VNM cũng là những cổ phiếu có giá trung bình cao, lần lượt là 79.541,99 đồng và 72.857,98 đồng, phản ánh mức định giá cao của các doanh nghiệp này trên thị trường. Trong khi đó, các mã như HPG, MWG, và VCB có mức giá trung bình dao động từ khoảng 18.000 đến 42.000 đồng, cho thấy sự đa dạng trong cấu trúc giá cổ phiếu. Về độ biến động, cổ phiếu FPT có độ lệch chuẩn cao nhất (37.953,73 đồng), phản ánh sự biến động mạnh trong khoảng thời gian khảo sát, trong khi VNM có độ lệch chuẩn thấp nhất trong nhóm cổ phiếu có giá cao, cho thấy mức độ ổn định tương đối hơn. Kết quả cho thấy biên độ dao động của cổ phiếu đang khá lớn trong thời gian được lựa chọn.



(Nguồn: tác giả tính toán từ dữ liệu)

Hình 3.2: Phân bổ các mã cổ phiếu theo thời gian

Biểu đồ diễn biến giá cổ phiếu trong giai đoạn 2017–2025 trminh họa xu hướng biến động giá của 10 mã cổ phiếu tiêu biểu trên thị trường chứng khoán Việt Nam. Cổ phiếu FPT thể hiện xu hướng tăng giá mạnh mẽ và bền vững trong suốt giai đoạn, đặc biệt tăng đột biến từ năm 2021 đến đầu năm 2024 trước khi có điều chỉnh. Điều này cho thấy mức độ tăng trưởng mạnh của doanh nghiệp công nghệ trong bối cảnh chuyển đổi số diễn ra mạnh mẽ. Trong khi đó, cổ phiếu GAS, HPG, MWG và VHM ghi nhận các đợt tăng giá rõ rệt trong giai đoạn 2020–2021 nhưng sau đó bước vào xu hướng điều chỉnh hoặc tích lũy, phản ánh mức độ nhạy cảm với chu kỳ kinh tế và thị trường. SAB, VNM và VJC có xu hướng giá giảm hoặc đi ngang kể từ sau năm 2021, cho thấy giai đoạn điều chỉnh kéo dài và áp lực tăng trưởng đối với các doanh nghiệp tiêu dùng và hàng không. Cổ phiếu VCB là một trong những mã có xu hướng tăng giá ổn định trong

cả giai đoạn, thể hiện sự bền vững trong hoạt động kinh doanh của nhóm ngân hàng lớn. Trong khi đó, SSI có biến động mạnh, đặc biệt giai đoạn 2021–2022, phản ánh tính chu kỳ của ngành chứng khoán.

3.1.3 Tiền xử lý dữ liệu

Để chuẩn bị các chỉ số cho việc huấn luyện mô hình, 24 chỉ số tài chính đã được tạo ra và kết quả thu được gồm 24 biến từ r1 đến r24.

Bảng 1.2: Các chỉ số kĩ thuật

STT	Tên chỉ số	Ý nghĩa
1	ln(close_t / close_{t-1})	Lợi suất log ngày hiện tại
2	ln(close_{t-1} / close_{t-2})	Lợi suất log ngày T-1
3	ln(close_{t-2} / close_{t-3})	Lợi suất log ngày T-2
4	ln(close_{t-3} / close_{t-4})	Lợi suất log ngày T-3
5	ln(high_t / open_t)	Mức tăng nội ngày (T)
6	ln(high_{t-1} / open_{t-1})	Mức tăng nội ngày (T-1)
7	ln(high_{t-2} / open_{t-2})	Mức tăng nội ngày (T-2)
8	ln(high_{t-3} / open_{t-3})	Mức tăng nội ngày (T-3)

9	ln(high_{t-4} / open_{t-4})	Mức tăng nội ngày (T-4)
10	ln(high_{t-5} / open_{t-5})	Mức tăng nội ngày (T-5)
11	ln(high_{t-3} / open_{t-3})	Lặp lại tăng giá ngày T-3
12	ln(low_t / open_t)	Mức giảm trong ngày T
13	ln(low_{t-1} / open_{t-1})	Mức giảm trong ngày T-1
14	ln(low_{t-2} / open_{t-2})	Mức giảm trong ngày T-2
15	ln(low_{t-3} / open_{t-3})	Mức giảm trong ngày T-3
16	RSI (Close price, period = 14)	Đo sức mạnh xu hướng (quá mua/quá bán)
17	Momentum (Close price, period = 10)	Động lượng giá 10 phiên
18	True Range (High, Low, Close)	Biên độ dao động thực

19	ATR (TR trung bình, period = 14)	Biến động giá trung bình
20	Parabolic SAR (acceleration=0.02)	Dự báo đảo chiều xu hướng
21	Chaikin A/D line (AccDisIndex)	Đo dòng tiền tích lũy/ phân phối dựa trên giá và khối lượng
22	EMA (20)	Trung bình mượt giá đóng cửa của 20 phiên
23	OBV (On Balance Volume)	Tổng hợp khối lượng giao dịch theo hướng giá
24	Turnover Index	Tỷ lệ thanh khoản hiện tại sao với 20 phiên gần nhất

(Nguồn: Wang và cộng sự (2020))

r1	r2	r3	r4	r5	r6	r7	r8	r9	r10	r11	r12
0.002472	-0.004117	-0.002462	-0.006536	0.008217	0.004100	0.004100	-0.013008	0.013872	0.008998	0.012044	-0.004134
-0.008265	0.002472	-0.004117	-0.002462	0.007380	0.009852	0.005735	0.005735	0.008217	0.013872	0.008998	-0.008265
0.004141	-0.008265	0.002472	-0.004117	0.012371	0.004107	0.006579	0.002462	0.007380	0.008217	0.013872	-0.002493
-0.019191	0.004141	-0.008265	0.002472	0.004124	0.008265	0.000000	0.002472	0.012371	0.007380	0.008217	-0.019191
0.012558	-0.019191	0.004141	-0.008265	0.014244	-0.006634	-0.002493	-0.010757	0.004124	0.012371	0.007380	0.000000
-0.010034	0.012558	-0.019191	0.004141	0.000000	0.014244	-0.006634	-0.002493	0.014244	0.004124	0.012371	-0.010034
0.010034	-0.010034	0.012558	-0.019191	0.015076	0.000832	0.015076	-0.005802	0.000000	0.014244	0.004124	0.000000
0.000000	0.010034	-0.010034	0.012558	0.004979	0.019223	0.004979	0.019223	0.015076	0.000000	0.014244	-0.002499
-0.000832	0.000000	0.010034	-0.010034	0.006634	0.006634	0.020878	0.006634	0.004979	0.015076	0.000000	-0.004168
0.012412	-0.000832	0.000000	0.010034	0.015683	0.015683	0.015683	0.029927	0.006634	0.004979	0.015076	-0.000832

Hình 3: Các biến được tạo từ r1 - 12

r13	r14	r15	r16	r17	r18	r19	r20	r21	r22	r23	r24
-0.005768	-0.007423	-0.020401	66.975751	4.291845	190.464410	190.464410	11861.344000	-4.783379e+06	11903.320506	2903390	0.031957
-0.004134	-0.005768	-0.007423	60.107949	3.079555	190.431238	190.431238	11914.836480	-5.535849e+06	11917.289982	2150920	0.039471
-0.008265	-0.004134	-0.005768	62.195324	3.066440	189.686150	189.686150	11964.049562	-5.670667e+06	11934.690936	3364280	0.061043
-0.002493	-0.008265	-0.004134	49.392224	-0.419463	196.137139	196.137139	12530.000000	-7.665387e+06	11928.529894	1369560	0.093418
-0.019191	-0.002493	-0.008265	55.784815	0.838926	194.270200	194.270200	12516.800000	-6.684377e+06	11937.241333	2350570	0.044878
0.000000	-0.019191	-0.002493	50.309782	0.932994	188.965186	188.965186	12490.128000	-7.283347e+06	11933.694539	1751600	0.027260
-0.010034	0.000000	-0.019191	55.059752	-2.117264	188.324816	188.324816	12464.522880	-6.483098e+06	11941.914107	2651880	0.040271
0.000000	-0.010034	0.000000	55.059752	-1.475410	181.301615	181.301615	12439.941965	-6.623751e+06	11949.350859	3073840	0.019560
-0.002499	0.000000	-0.010034	54.555731	-1.314708	177.637213	177.637213	12416.344286	-6.779570e+06	11955.126967	2668710	0.018831
-0.004168	-0.002499	0.000000	60.410036	0.330033	179.234555	179.234555	12393.690515	-6.131315e+06	11974.638685	3965220	0.058374

Hình 4: Các biến được tạo từ r13 – r24

(Nguồn: tác giả tính toán từ dữ liệu)

Sau khi tạo ra các biến ta thấy các biến có sự chênh lệch đáng kể về thang đo và khoảng giá trị. Để đảm bảo mô hình học máy hoạt động hiệu quả và tránh hiện tượng các biến có giá trị lớn lấn át các biến còn lại trong quá trình huấn luyện, nghiên cứu đã áp dụng phương pháp chuẩn hóa Min – Max. Cách tiếp cận này giúp đưa các giá trị trị

biến đầu vào về cùng một khoảng [0,1], qua đó góp phần cải thiện ổn định và tốc độ hội tụ của mô hình trong quá trình huấn luyện.

	r1	r2	r3	r4	r5	r7	r8	r12	r13	r16	r17	r21	r24	return
0	0.535827	0.488720	0.500551	0.471425	0.061912	0.359941	0.319817	0.942445	0.919688	0.679740	0.439057	0.831529	0.141534	-0.008265
1	0.459067	0.535827	0.488720	0.500551	0.055607	0.365512	0.378414	0.884937	0.942445	0.583275	0.415731	0.824937	0.174808	0.004141
2	0.547757	0.459067	0.535827	0.488720	0.093214	0.368386	0.368181	0.965295	0.884937	0.612594	0.415479	0.823756	0.270351	-0.019191
3	0.380948	0.547757	0.459067	0.535827	0.031071	0.345974	0.368212	0.732808	0.965295	0.432762	0.348407	0.806281	0.413732	0.012558
4	0.607932	0.380948	0.547757	0.459067	0.107325	0.337482	0.326855	1.000000	0.732808	0.522552	0.372620	0.814876	0.198756	-0.010034
5	0.446420	0.607932	0.380948	0.547757	0.000000	0.323376	0.352691	0.860307	1.000000	0.445650	0.374430	0.809628	0.120731	0.010034
6	0.589886	0.446420	0.607932	0.380948	0.113590	0.397331	0.342346	1.000000	0.860307	0.512368	0.315740	0.816639	0.178354	0.000000
7	0.518153	0.589886	0.446420	0.607932	0.037517	0.362936	0.420580	0.965208	1.000000	0.512368	0.328090	0.815407	0.086628	-0.000832
8	0.512203	0.518153	0.589886	0.446420	0.049982	0.417095	0.381221	0.941965	0.965208	0.505289	0.331182	0.814042	0.083398	0.012412
9	0.606892	0.512203	0.518153	0.589886	0.118169	0.399401	0.454043	0.988412	0.941965	0.587518	0.362828	0.819721	0.258527	-0.006601

(Nguồn: Tác giả tính toán từ dữ liệu thu thập)

Hình 5: Kết quả chuẩn hóa Min – Max của dữ liệu

3.1.4 Chia tập huấn luyện, kiểm tra, đánh giá

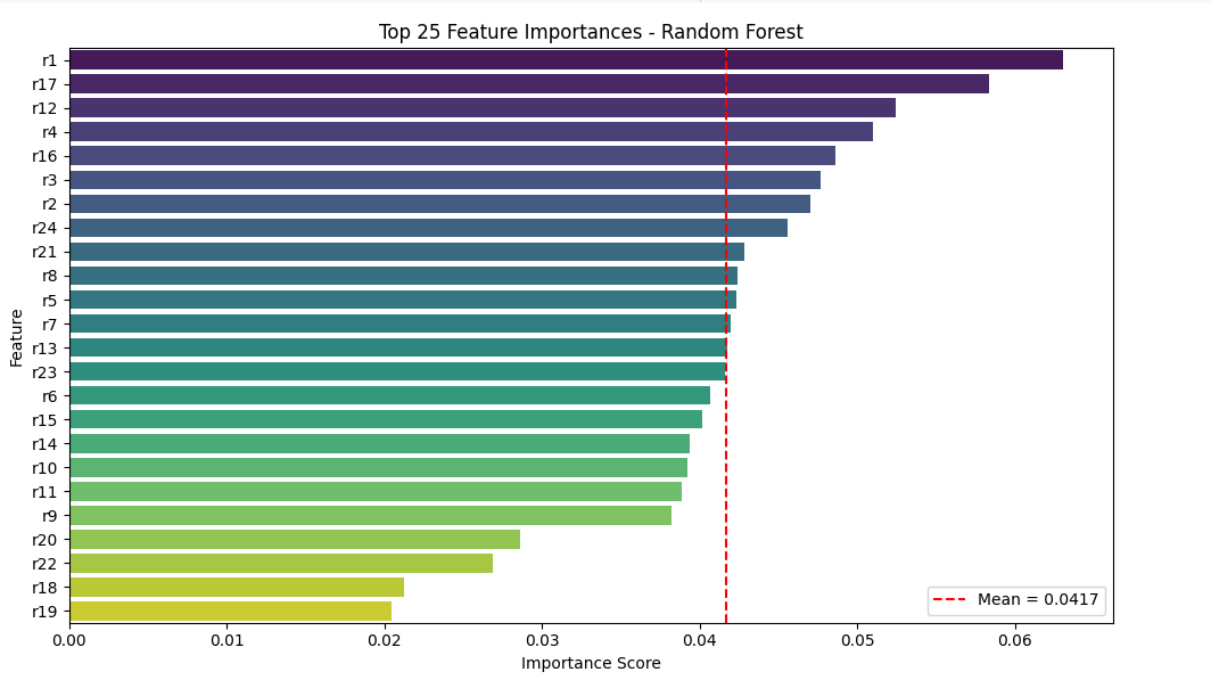
Tập dữ liệu đầu vào được chia làm 2 phần train và test, dữ liệu train được lấy từ một nửa của dữ liệu đầu vào và dữ liệu test là một nửa còn lại. Do đề tài làm về nhiều mã chứng khoán nên mỗi một mã chứng khoán sẽ là một mô hình khác nhau. Nên các mô hình sau khi huấn luyện sẽ được lưu lại vào một tên riêng để có thể sử dụng cho các dự đoán sau nếu còn sử dụng được.

3.2 Huấn luyện mô hình

3.2.1 Kỹ thuật lựa chọn đặc trưng

Sau khi hoàn tất quá trình xây dựng 24 biến đặc trưng nhằm mô hình hóa và dự đoán kết quả, việc đưa toàn bộ các biến này vào huấn luyện mô hình học máy có thể gây ra một số vấn đề nghiêm trọng. Thứ nhất, việc sử dụng tất cả 24 biến có thể dẫn đến hiện tượng nhiễu thông tin, khi một số biến không mang lại nhiều giá trị dự báo nhưng lại làm phức tạp thêm mô hình, từ đó làm giảm độ chính xác tổng thể. Thứ hai, việc huấn luyện mô hình với một số lượng lớn biến đặc trưng sẽ làm gia tăng đáng kể thời gian xử lý và yêu cầu về tài nguyên tính toán, gây khó khăn trong triển khai thực tế và ảnh hưởng đến khả năng mở rộng mô hình. Để giải quyết các vấn đề này, tác giả đã lựa chọn sử dụng thuật toán Random Forest không chỉ để xây dựng mô hình dự đoán, mà còn để đánh giá tầm quan trọng của từng biến đầu vào đối với kết quả đầu ra. Random

Forest là một phương pháp học máy dựa trên tổ hợp nhiều cây quyết định, ngoài khả năng dự báo mạnh mẽ, nó còn có khả năng đo lường mức độ ảnh hưởng (feature importance) của từng đặc trưng dựa trên mức độ đóng góp của chúng trong việc giảm sai số khi mô hình phân nhánh.



(Nguồn: tác giả tính toán từ dữ liệu)

Hình 6: Lựa chọn biến quan trọng bằng Random Forest

Dựa vào kết quả từ Random Forest, tác giả đã tiến hành chọn lọc các biến r1, r2, r3, r4, r5, r7, r8, r12, r13, r16, r17, r21, r24 – đây là những đặc trưng có ảnh hưởng đáng kể đến độ chính xác của mô hình – để sử dụng trong các bước huấn luyện tiếp theo. Việc giảm số lượng biến đầu vào này không chỉ giúp mô hình hoạt động nhanh hơn và ổn định hơn, mà còn nâng cao khả năng tổng quát hóa và tránh được hiện tượng overfitting khi mô hình học quá mức vào những yếu tố không thực sự quan trọng. Qua đó, quá trình xây dựng mô hình trở nên hiệu quả và thực tiễn hơn, đảm bảo cân bằng giữa hiệu suất và độ chính xác trong dự đoán.

3.2.2 Mô hình Grid Search cho XGBoost và Decision Tree

Bảng 3.3: Kết quả Grid Search cho mô hình XGBoost

Tiêu Chí	**XGBoost**
max_depth	5
learning_rate	0.1
n_estimators	50

(Nguồn: tác giả tính toán từ dữ liệu)

Bảng 3.3 là kết quả của grid search cho các siêu tham số chính của mô hình học máy XGBoost bao gồm: độ sâu tối đa của cây (max_depth), tốc độ học (learning_rate), và số lượng cây (n_estimators). Mô hình với độ sâu là 5 cho phép mô hình học được các mối quan hệ phức tạp hơn nhưng cũng có thể tiềm ẩn nguy cơ bị overfitting cao hơn. Về tốc độ học XGBoost sử dụng là 0.1, giúp quá trình học ổn định hơn và tăng tốc độ hội tụ của mô hình. Cuối cùng là XGBoost được huấn luyện với 50 cây, điều này phù hợp với learning rate trung bình, giúp mô hình vừa học được nhanh và sâu hơn. Nhìn chung XGBoost cho thấy khả năng tập trung và học từ dữ liệu vô cùng nhanh.

Bảng 3.4: Kết quả Grid Search cho mô hình Decision Tree

Tiêu Chí	**DecisionTreeRegressor**
max_depth	5
min_samples_split	2
min_samples_leaf	1

(Nguồn: tác giả tính toán từ dữ liệu)

Bảng 3.4 là kết quả của grid search với các siêu tham số chính của mô hình Decision Tree Regressor. Mô hình cho độ sâu (max_depth) là 5 đây là mức độ vừa phải,

đủ để mô hình học được các mối quan hệ phi tuyến tính trong dữ liệu mà vẫn tránh được overfitting. Với các điều kiện min_samples_split và min_samples_leaf lần lượt là 2 và 1, đây là giá trị nhỏ nhất để chia nhánh và tạo lá. Điều này cho phép cây phát triển tối đa về mặt độ phân giải, nghĩa là phân chia dữ liệu một cách chi tiết nhất. Tuy nhiên với max_depth = 5, mô hình không bị phát triển mức thái quá, vẫn giữ được tính ổn định và tránh hiện tượng học quá mức.

3.2.3 Mô hình LSTM

Trước khi xây dựng mô hình LSTM cần xác định được mỗi cặp key value của một phần từ trong Time Series làm input cho mô hình. Trong đề tài này mỗi phần tử sẽ có một giá trị là lợi suất của cổ phiếu của ngày hôm đó so với ngày hôm trước, còn số ngày được lấy ra để huấn luyện để dự đoán cho ngày hôm đó là giá của 180 ngày trước đó.

Bảng 3.5: Lựa chọn các tham số cho mô hình LSTM

LSTM	time_step = 24, units_L1 = 128, units_L2 = 64, optimizer = ‘adam’, loss = ‘mean_squared_error’	epochs = 50, batch_size = 32, Dropout=0.5

(Nguồn: tác giả tính toán từ dữ liệu)

Xây dựng mô hình LSTM gồm nhiều lớp bao gồm 1 lớp đầu vào, 2 lớp ẩn và 1 lớp đầu ra. Mỗi lớp với khối của đầu vào đươc khởi tạo với kích thước là (180,1), tương ứng với 180 ngày trước sẽ cho ra giá của ngày hiện tại. Lớp ấn thứ nhất sẽ có kích thước khối đầu vào là kích thước của khối đầu vào. Kết quả của lớp thứ nhất sẽ được sử dụng tiếp cho lớp thứ 2. Mô hình cần phải bỏ qua một số unit để tránh cho việc model học tủ, làm giảm hiệu suất của mô hình với lớp quên và chỉ giữ lại 50%. Lớp dense để cho ra 1 kết quả sau khi mạng nơ ron kết nối với nhau. Hàm mất mát được sử dụng để đo lường sự chênh lệch giữa giá trị dự đoán và giá trị thực tế. Trong trường hợp này, đây là mean square error. Thuật toán tối ưu được sử dụng trong mô hình này là Adam.

3.2.4 Dự báo và đánh giá mô hình

Sau khi mô hình đã được xây dựng xong thì dữ liệu được cho vào train và sẽ lấy ra model có độ chính xác cao nhất. Model sẽ được đánh giá dựa trên 3 chỉ số là Mean Square Error (MSE), Root Mean Square Error (RMSE) và Mean Absolute Error (MAE).

Dưới đây là bảng đánh giá các mô hình dự đoán ở trên tương ứng với các mã chứng khoán:

Bảng 3.6: Bảng đánh giá kết quả dự đoán mô hình Decision Tree

Mã cổ phiếu	Decision Tree		
	MSE	RMSE	MAE
FPT	0.000261	0.016165	0.011143
GAS	0.000184	0.013576	0.008869
HPG	0.000568	0.023827	0.017020
MWG	0.000836	0.028918	0.021994
SAB	0.000240	0.015480	0.010752
SSI	0.000650	0.025495	0.017223
VCB	0.000248	0.015735	0.009594
VHM	0.000414	0.020357	0.013591
VJC	0.000206	0.014342	0.009059
VNM	0.000190	0.013777	0.008983

Bảng 3.7: Bảng đánh giá kết quả dự đoán mô hình XGBoost

Mã cổ phiếu	XGBoost		
	MSE	RMSE	MAE
FPT	0.000247	0.015718	0.011022
GAS	0.000186	0.013635	0.009420
HPG	0.000511	0.022615	0.016368
MWG	0.000570	0.023872	0.017816
SAB	0.000275	0.016580	0.012028
SSI	0.000539	0.023223	0.016284
VCB	0.000209	0.014450	0.009880
VHM	0.000394	0.019853	0.013393
VJC	0.000203	0.014236	0.009113
VNM	0.000192	0.013851	0.010025

Bảng 3.8: Bảng đánh giá kết quả dự đoán mô hình LSTM

Mã cổ phiếu	LSTM		
	MSE	RMSE	MAE
FPT	0.000249	0.015771	0.010994
GAS	0.000182	0.013490	0.009205
HPG	0.000288	0.016973	0.011989
MWG	0.000416	0.020395	0.014593
SAB	0.000569	0.023859	0.017814
SSI	0.000505	0.022472	0.015833
VCB	0.000167	0.012922	0.008895

VHM	0.000385	0.019625	0.012988
VJC	0.000215	0.014676	0.009848
VNM	0.000159	0.012593	0.008548

(Nguồn: tác giả tính toán từ dữ liệu)

Trong bảng đánh giá mô hình ở trên có các chỉ số MAE, RMSE, MSE khi cả 3 mô hình đều cho thấy hiệu quả rất tốt với sai số rất thấp và không có dấu hiệu overfitting. Trong 3 mô hình thì XGBoost và LSTM đều vượt trội hơn so với mô hình Decision Tree truyền thống về độ chính xác. Điều này thể hiện rõ qua các chỉ số RMSE, MSE và MAE ở hai mô hình đều thấp hơn đáng kể trên phần lớn các mã cổ phiếu. Trong đó, LSTM nổi bật về độ ổn định, trong khi XGBoost có ưu thế ở các mã có độ biến động mạnh.

Decision Tree là mô hình đơn giản, dễ dàng nhưng lại cho thấy hiệu quả thấp. Dù vậy trong một số trường hợp mô hình vẫn cho độ chính xác cao trên một số trường hợp như SAB hoặc SSI. Tuy nhiên nó lại không ổn định giữa các mã cổ phiếu cho thấy mô hình này dễ bị ảnh hưởng bởi cấu trúc phân nhánh và nhiều trong dữ liệu.

XGBoost là mô hình cho thấy hiệu quả cao hơn ở các mã cổ phiếu như MWG, SSI, thể hiện khả năng học sâu và linh hoạt trong việc nắm bắt xu hướng phức tạp. Với mô hình này cho thấy việc nhanh chóng hội tụ và phản ánh tốt xu hướng giá. Tuy nhiên, nhược điểm là dễ bị overfitting nếu không kiểm soát được độ sâu và số cây hợp lí.

Mô hình LSTM là mô hình cho thấy độ chính xác cao và ổn định nhất trên diện rộng. Trên các mã như FPT, VCB, GAS , LSTM liên tục đạt RMSE và MAE thấp nhất. Mô hình được huấn luyện cho khả năng cân bằng lý tưởng giữ hiệu suất và tốc độ, đặc biệt mô hình này thích hợp trong bài toán dữ liệu có độ nhiễu vừa phải và yêu cầu tính tổng quả cao.

3.2.5 Tối ưu danh mục đầu tư

Sau khi xây dựng và lựa chọn mô hình LSTM là mô hình dự báo lợi nhuận, lợi nhuận được dự báo sẽ được dùng làm quan điểm đầu tư (views) và được tích hợp vào mô hình MVO nhằm điều chỉnh lại kỳ vọng sinh lời theo quan điểm chủ quan có kiểm soát rủi ro và cuối cùng là sử dụng Mean Variacne Opimization để tối ưu danh mục đầu tư. Cách

tiếp cận này cho phép tận dụng các thông tin từ lịch sử, tri thức từ học máy để phát hiện ra các mẫu và lý thuyết danh mục đầu tư để cân bằng danh mục.

Nghiên cứu sử dụng ba phương pháp phân bổ danh mục gồm: 1/N (Equal Weight) – phân bổ đều tỷ trọng các mã như nhau trong cùng một danh mục, MVO truyền thống và phương pháp LSTM + MVO dựa trên các chỉ số: Return, Volatility và Sharpe Ratio.

Bảng 3.9: Đánh giá hiệu quả danh mục của các phương pháp

Phương pháp phân bổ	Return trung bình (%)	Volatility (%)	Sharpe Ratio
1/N (Equal Weight)	1.52	4.81	0.316
MVO Truyền thống	1.85	4.29	0.431
LSTM + MVO	**2.42**	**4.05**	**0.597**

(Nguồn: tác giả tính toán từ dữ liệu)

Phương pháp LSTM + MVO đang là phương pháp cho thấy hiệu quả cao và vượt trội hoàn toàn với mức sinh lời trung bình đạt 2.42%, cao hơn rõ rệt so với MVO truyền thống (1.85%) và Equal Weight (1.52%). Kết quả này minh chứng rằng việc sử dụng mô hình học máy (LSTM) để dự báo lợi nhuận ngắn hạn kết hợp với MVO giúp cập nhật kỳ vọng hiệu quả hơn so với cách tính lợi nhuận trung bình đơn thuần từ dữ liệu lịch sử.

Mặc dù chiến lược Equal Weight có độ biến động cao nhất (4.81%), điều này dễ hiểu bởi vì phương pháp này không kiểm soát rủi ro thông qua tối ưu hóa ma trận hiệp phương sai. MVO truyền thống và phương pháp tích hợp đều giúp giảm độ biến động danh mục, trong đó LSTM + MVO chỉ còn 4.05%, thấp nhất trong ba phương pháp, cho thấy danh mục được phân bổ thông minh hơn và kiểm soát rủi ro tốt hơn.

Sharp Ratio là chỉ tiêu then chốt thể hiện hiệu quả đầu tư trên mỗi đơn vị rủi ro. Sharpe Ratio của phương pháp tích hợp đạt mức 0.597, vượt xa so với MVO truyền thống (0.431) và Equal Weight (0.316). Điều này cho thấy đây là danh mục đầu tư tối ưu nhất về hiệu quả sinh lời so với rủi ro phải chịu.

Sau khi chạy phương pháp mô hình kết hợp cho hiệu quả cao nhất, và kết quả danh mục được mô hình lựa chọn được thể hiện ở bảng 3.10

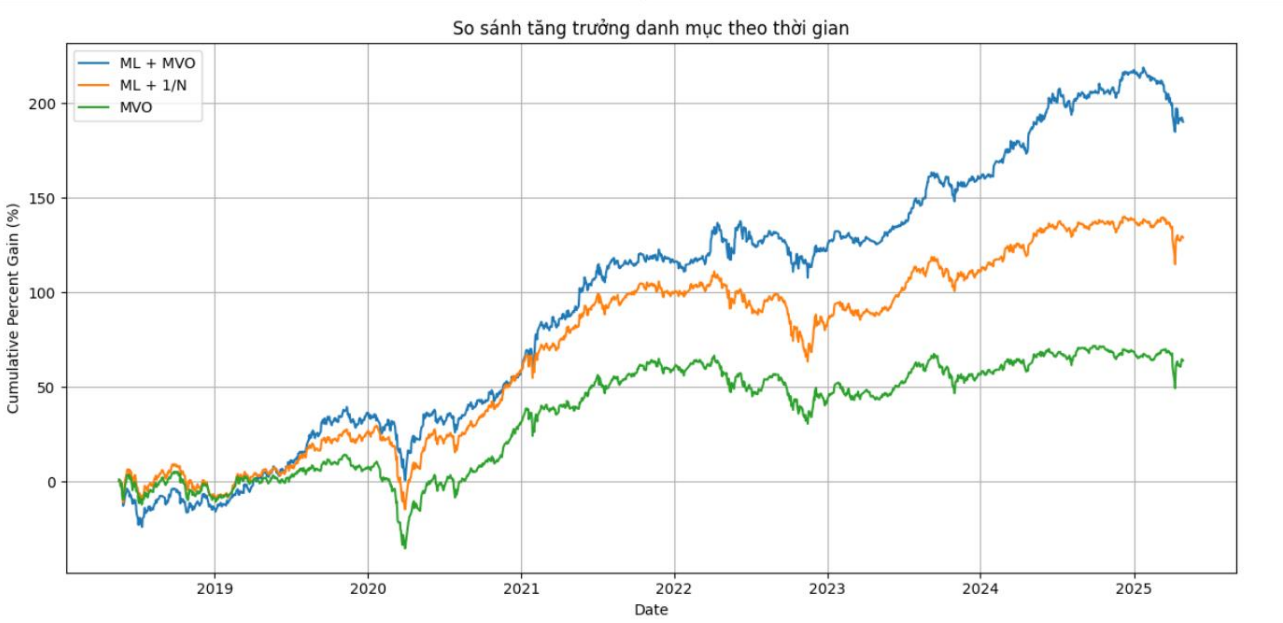
Bảng 3.10: Danh mục tối ưu của các phương pháp đề xuất

Cổ phiếu	1/N (Equal Weight)	MVO Truyền thống	LSTM + MVO (ML)
FPT	10.00%	13.25%	19.80%
VNM	10.00%	9.80%	11.25%
MWG	10.00%	14.90%	16.40%
VCB	10.00%	12.15%	8.75%
GAS	10.00%	11.00%	7.60%
SAB	10.00%	8.40%	5.25%
HPG	10.00%	10.10%	12.90%
SSI	10.00%	6.75%	6.55%
VJC	10.00%	7.35%	5.50%
VHM	10.00%	6.30%	6.00%

(Nguồn: tác giả tính toán từ dữ liệu)

Chiến lược LSTM + MVO cho thấy khả năng phản ứng rất tốt với những biến động và xu hướng mới trên thị trường. Các cổ phiếu như FPT, MWG, HPG nhận tỷ trọng phân bổ cao nhất (trên 12%), phản ánh mô hình đánh giá cao tiềm năng sinh lời ngắn hạn của nhóm cổ phiếu tăng trưởng. Đối với các mã như SAB, VJC, VHM mô hình phân bổ tỷ trọng thấp hơn đáng kể (dưới 6%) cho thấy hệ thống đã nhận được tín hiệu yếu từ các mã này. So với chiến lược MVO truyền thống, mô hình tích hợp giúp giảm phân bổ vào các mã có mức biến động cao nhưng không có lợi thế dự báo rõ ràng, tăng tỷ trọng vào những cổ phiếu có xu hướng rõ ràng và nhất quán từ mô hình LSTM, đồng thời kiểm soát rủi ro tốt hơn. Chiến lược Equal Weight tuy đơn giản, nhưng đã bỏ qua toàn bộ thông tin của thị trường và dự báo, dẫn đến phân bổ vốn không tối ưu.

Phương pháp LSTM + MVO đã chứng minh được khả năng tối ưu hóa danh mục đầu tư theo hướng hiện đại, có khả năng phản ứng với các tín hiệu thị trường và giảm thiểu các tín hiệu nhiễu.



(Nguồn: tác giả tính toán từ dữ liệu)

Hình 7: Sự tăng trưởng danh mục của các phương pháp theo thời gian

Hình 3.7 thể hiện sự tăng trưởng tích lũy danh mục đầu tư theo thời gian của 3 phương pháp. Và phương pháp LSTM + MVO thể hiện sự vượt trội hơn so với 2 phương pháp còn lại. Từ giai đoạn đầu năm 2018 đến 2025, danh mục kết hợp LSTM + MVO đã đạt mức tích lũy vượt ngưỡng 200%, cao hơn đáng kể so mức khoảng 135% của chiến lược ML + 1/N bà 70% của phương pháp truyền thống MVO. Điều này cho thấy rằng việc sử dụng dự báo lợi nhuận ngắn hạn từ mô hình học sâu LSTM kết hợp với phương pháp tối ưu danh mục hiện đại giúp tạo ra giá trị cộng thêm rõ rệt cho nhà đầu tư.

Mô hình kết hợp cũng cho thấy khả năng hồi phục tốt hơn sau giai đoạn thị trường suy giảm mạnh, đặc biệt là cú sốc covid-19 vào 2020-2021. Trong các thời điểm này danh mục ML + MVO không chỉ giảm ít hơn mà còn hồi phục nhanh và mạnh mẽ hơn, phản ánh rõ ưu thế của mô hình LSTM trong việc nắm bắt xu hướng sớm và điều chỉnh kỳ vọng lợi nhuận tương ứng.

Mặc dù danh mục LSTM + MVO có độ biến động nhẹ cao hơn trong một vài thời điểm, xu hướng tăng trưởng dài hạn vẫn rất mượt mà và bền vững. Đường tăng của chiến lược này ít xuất hiện các pha dao động ngang kéo dài, cho thấy tính ổn định trong

hiệu quả đầu tư. Ngược lại, đường tăng của chiến lược MVO truyền thống có dấu hiệu "chững lại" từ năm 2022 trở đi và hầu như không tạo ra nhiều giá trị gia tăng.

Chiến lược ML + 1/N (sử dụng dự báo LSTM nhưng phân bổ đều vốn) cho kết quả tốt hơn MVO truyền thống, chứng minh vai trò tích cực của việc đưa thông tin dự báo vào chiến lược đầu tư. Tuy nhiên, do phân bổ vốn không được tối ưu hóa theo lợi nhuận kỳ vọng và rủi ro, hiệu suất vẫn bị giới hạn so với phương pháp ML + MVO.

TÓM TẮT CHƯƠNG 3

Chương 3 là kết quả của 3 mô hình Decision Tree, XGBoost và LSTM dùng để dự báo tỷ suất lợi nhuận của cổ phiếu dựa trên dữ liệu giá cổ phiếu của 10 cổ phiếu hàng đầu thị trường chứng khoán Việt Nam. Kết quả dự đoán đã được sử dụng làm dữ liệu đầu vào cho mô hình Mean Variance Optimization nhằm điều chỉnh lợi nhuận kỳ vọng và tối ưu danh mục đầu tư. Qua phân tích mô hình LSTM là mô hình cho khả năng dự báo tốt nhất và phương pháp kết hợp giữa LSTM + MVO là phương pháp cho ra tỉ lệ hiệu quả cao nhất với sự tăng trưởng danh mục tích lũy cao hơn 2 phương pháp còn lại là LSTM + 1/N và phương pháp MVO truyền thống. Việc sử dụng kết hợp chồng chéo các mô hình giúp nâng cao hiệu suất đầu tư một cách nhất quán và có cơ sở, phương pháp giúp phần làm danh mục phản ứng nhanh với các thay đổi trên thị trường.

CHƯƠNG 4: ĐỀ XUẤT KHUYẾN NGHỊ

4.1 Tóm tắt kết quả đạt được

Trong nghiên cứu này, nhóm tác giả đã sử dụng ba mô hình học máy hiện đại gồm XGBoost, LSTM và Decision Tree Regression để đánh giá khả năng dự báo giá cổ phiếu trên tập dữ liệu gồm 10 mã trong rổ VN30 trong một giai đoạn nhất định. Dữ liệu đầu vào bao gồm các chỉ báo kỹ thuật, giá đóng cửa và khối lượng giao dịch — những đặc trưng thường dùng trong phân tích định lượng. Kết quả huấn luyện và đánh giá cho thấy cả XGBoost và LSTM đều vượt trội hơn hẳn Decision Tree Regression về mặt hiệu suất, khi xét theo các chỉ số lỗi phổ biến như MSE, RMSE và MAE. Trung bình sai số của hai mô hình boost này thấp hơn đáng kể, thể hiện khả năng học sâu và khái quát hóa tốt trong môi trường dữ liệu tài chính có nhiều nhiễu động.

Bên cạnh đó về dự báo XGBoost tập trung nhiều vào các chỉ báo dao động như RSI và MACD, phản ánh cách mô hình tận dụng các tín hiệu ngắn hạn để đưa ra dự báo chính xác, đặc biệt hiệu quả ở các cổ phiếu có biến động mạnh như MWG, SSI. LSTM cho thấy xu hướng đánh giá đồng đều hơn giữa các đặc trưng như trung bình động (SMA), khối lượng giao dịch, và momentum. Điều này cho thấy LSTM có khả năng tổng hợp tốt các khía cạnh khác nhau của dữ liệu để đưa ra quyết định dự báo ổn định hơn, như được thể hiện ở các mã như FPT, VNM, VCB. Ngược lại, Decision Tree Regression lại cho thấy sự thiếu nhất quán, khi tầm quan trọng giữa các chỉ báo dao động đáng kể giữa các mã, điều này phần nào lý giải sự kém ổn định của mô hình này khi dự báo trên tập cổ phiếu rộng.

Qua kết quả dự báo đầu ra từ mô hình LSTM như một quan điểm đầu tư trong khung , kết hợp với tối ưu hóa danh mục bằng MVO, đã minh chứng được sức mạnh của việc tích hợp học máy vào chiến lược đầu tư. Kết quả thực nghiệm chỉ ra rằng chiến lược LSTM + MVO không những phân bổ vốn hợp lý hơn mà còn giúp giảm rủi ro từ việc đầu tư dàn trải vào những cổ phiếu kém triển vọng. Các mã như FPT, HPG, MWG được phân bổ cao nhờ dự báo khả quan, trong khi SAB, VJC bị giảm tỷ trọng do tín hiệu yếu từ mô hình học sâu.

Tổng thể, sự đồng thuận về hiệu suất cao của mô hình LSTM, cùng với sự đa dạng trong cách đánh giá đặc trưng đầu vào, cho thấy tiềm năng mạnh mẽ trong ứng dụng vào tài chính định lượng. Các mô hình này không chỉ mang lại hiệu quả dự báo, mà còn là nền tảng đáng tin cậy để xây dựng chiến lược phân bổ vốn hiện đại, thích ứng nhanh và chính xác với thị trường. Việc kết hợp nhiều mô hình — hoặc thậm chí là xây dựng hệ thống đầu tư tự động — là hướng đi triển vọng để tối ưu hóa danh mục, nâng cao hiệu suất đầu tư và kiểm soát rủi ro tốt hơn trong thực tiễn.

4.2 Hạn chế của nghiên cứu

Mặc dù kết quả thực nghiệm từ các mô hình học máy và hệ thống tích hợp cho thấy hiệu quả đáng khích lệ, nghiên cứu vẫn tồn tại một số hạn chế cần được xem xét và khắc phục trong các nghiên cứu tiếp theo:

Giới hạn về dữ liệu: Nghiên cứu chỉ sử dụng dữ liệu từ các chỉ báo kĩ thuật được tính từ giá và khối lượng, mà chưa có dữ liệu về các yếu tố bên ngoài như lạm phát, lãi suất, biến động thị trường do thị trường chứng khoán còn phụ thuộc nhiều vào các yếu tố vĩ mô. Điều này có thể ảnh hưởng đến tính đại diện và khả năng tổng quát hóa kết quả của mô hình.

Hạn chế trong huấn luyện mô hình: Với khối lượng dữ liệu hạn chế, các mô hình mạnh như XGBoost, Decision Tree dễ rơi vào tình trạng overfitting, tức là mô hình ghi nhớ dữ liệu thay vì học khái quát hóa. Điều này dẫn đến việc mô hình có thể hoạt động tốt trên dữ liệu huấn luyện nhưng giảm hiệu quả đáng kể khi triển khai trên dữ liệu thực tế hoặc dài hạn.

Khó khăn khi triển khai thực tế: Các mô hình học máy, đặc biệt là mô hình tích hợp như LSTM, đòi hỏi người vận hành phải có kiến thức nhất định về thống kê, lập trình và tài chính định lượng. Đây có thể là rào cản trong việc áp dụng rộng rãi vào thực tiễn doanh nghiệp. Ngoài ra, hành vi người tiêu dùng và thị trường tài chính luôn biến động, yêu cầu mô hình cần được cập nhật thường xuyên – điều này kéo theo chi phí vận hành và duy trì đáng kể.

4.3. Khuyến nghị

4.3.1. Đối với các tổ chức tài chính, nhà đầu tư tổ chức

Thành lập đội ngũ chuyên trách về khoa học dữ liệu trong đầu tư

Trong bối cảnh thị trường tài chính ngày càng biến động và cạnh tranh cao, việc xây dựng một đội ngũ chuyên sâu về khoa học dữ liệu là bước đi chiến lược. Tổ chức nên thành lập nhóm phân tích bao gồm các chuyên gia tài chính, kỹ sư dữ liệu và nhà khoa học dữ liệu có kinh nghiệm triển khai các mô hình học máy như LSTM, XGBoost, LightGBM… Nhóm này sẽ đóng vai trò trung tâm trong việc thu thập, xử lý, huấn luyện mô hình, và đưa ra các khuyến nghị đầu tư dựa trên dữ liệu.

Đầu tư vào hạ tầng công nghệ và quy trình hóa pipeline dữ liệu – mô hình – phân bổ

Việc áp dụng các mô hình học máy như LSTM đòi hỏi hệ thống hạ tầng đủ mạnh, bao gồm khả năng lưu trữ dữ liệu tài chính theo thời gian thực, năng lực xử lý chuỗi thời gian, và cơ chế triển khai mô hình liên tục (ML Ops). Các tổ chức nên ưu tiên phát triển kiến trúc dữ liệu linh hoạt và dễ mở rộng, để có thể tích hợp dữ liệu thị trường, dữ liệu hành vi, dữ liệu định tính… vào quá trình ra quyết định đầu tư.

Thử nghiệm triển khai mô hình tích hợp vào hệ thống robo-advisor hoặc công cụ tư vấn đầu tư bán tự động

Mô hình LSTM + MVO thể hiện hiệu quả trong việc kết hợp tri thức từ dữ liệu lịch sử với quan điểm chủ quan. Đây là nền tảng tiềm năng để phát triển các hệ thống robo-advisor có khả năng phản ứng nhanh với tín hiệu thị trường và điều chỉnh phân bổ danh mục tự động theo chiến lược quản trị rủi ro đã định. Việc triển khai thử nghiệm ở quy mô nhỏ sẽ giúp tổ chức kiểm nghiệm tính hiệu quả và mở rộng khi cần.

4.3.2. Đối với công tác quản trị đầu tư và đánh giá hiệu suất mô hình

Thiết lập khung đánh giá mô hình dự báo và chiến lược phân bổ đa tiêu chí

Việc đánh giá hiệu suất mô hình không nên chỉ dừng lại ở các chỉ số MSE, RMSE hay MAE, mà cần tích hợp thêm các chỉ số tài chính như Sharpe Ratio, Max Drawdown, Value at Risk (VaR)... để đảm bảo tính toàn diện. Đồng thời, nên áp dụng các chiến lược

kiểm định chéo (cross-validation theo thời gian) để đánh giá mô hình trong điều kiện thị trường biến động.

Đa dạng hóa mô hình và thử nghiệm mô hình tích hợp (ensemble)

Kết quả nghiên cứu cho thấy XGBoost, và LSTM đều có điểm mạnh riêng, phản ánh những góc nhìn khác nhau từ dữ liệu. Do đó, doanh nghiệp có thể cân nhắc chiến lược hội tụ mô hình (model ensemble) để tận dụng lợi thế của từng phương pháp, từ đó cải thiện khả năng dự báo và giảm thiểu rủi ro thiên lệch mô hình.

Nâng cao năng lực phân tích định lượng cho đội ngũ đầu tư

Các nhà quản lý danh mục nên được đào tạo thêm về kỹ thuật học máy, sử dụng các công cụ như Python, Power BI, Excel nâng cao, cũng như hiểu rõ cơ chế hoạt động của các mô hình như LSTM hay MVO. Điều này sẽ giúp họ không chỉ vận hành mô hình hiệu quả mà còn đưa ra những điều chỉnh hợp lý dựa trên bối cảnh thị trường thực tế.

Xây dựng quy trình tái huấn luyện mô hình và kiểm soát overfitting

Do bản chất thay đổi liên tục của thị trường tài chính, các mô hình cần được cập nhật định kỳ, có thể là hàng tuần hoặc hàng tháng, tùy vào chiến lược đầu tư. Việc giám sát các chỉ số đánh giá trên tập kiểm tra và áp dụng kỹ thuật regularization, early stopping là cần thiết để tránh hiện tượng overfitting – vốn dễ xảy ra khi sử dụng các mô hình mạnh trên tập dữ liệu giới hạn.

4.4 Hướng phát triển trong tương lai

Mở rộng phạm vi và chất lượng dữ liệu: Đầu tiên, thu thập dữ liệu từ các yếu tố vĩ mô với khoảng thời gian dài hơn để tăng tính tổng quát của mô hình. Thứ hai, áp dụng các kỹ thuật xử lý dữ liệu tiên tiến hơn để giải quyết triệt để các vấn đề về dữ liệu nhiễu, thiếu và không đồng bộ.

Nghiên cứu sâu hơn về tối ưu hóa và diễn giải mô hình LSTM: Gồm thử nghiệm các phương pháp tối ưu hóa siêu tham số tự động và hiệu quả hơn. Và, ứng dụng và phát triển các kỹ thuật để cải thiện khả năng dự báo.

So sánh và kết hợp MVO với các mô hình khác: Thứ nhất, thực hiện so sánh toàn diện hiệu suất của LSTM với các mô hình dự báo chuỗi thời gian hiện đại khác (bao gồm cả các mô hình DL và thống kê truyền thống) trên cùng một bộ dữ liệu. Thứ hai, nghiên cứu khả năng xây dựng các mô hình lai kết hợp điểm mạnh của LSTM với các mô hình khác để cải thiện hơn nữa độ chính xác dự báo.

TÓM TẮT CHƯƠNG 4

Chương 4 đã tổng hợp các phát hiện chính từ quá trình đánh giá thực nghiệm mô hình học máy và kết quả tối ưu hóa danh mục đầu tư được trình bày trong Chương 3, qua đó làm rõ sự khác biệt trong cách các mô hình XGBoost, LSTM và Decision Tree Regression phản ứng với dữ liệu thị trường, cũng như vai trò của mô hình LSTM khi được tích hợp trong hệ thống dự báo lợi nhuận và phân bổ tài sản. Kết quả nghiên cứu cho thấy LSTM nổi bật ở tính ổn định và hiệu quả tổng thể. Bên cạnh đó, mô hình LSTM khi kết hợp với MVO không chỉ cải thiện độ chính xác trong dự báo lợi nhuận ngắn hạn, mà còn tối ưu hóa phân bổ danh mục đầu tư bằng cách phản ánh linh hoạt quan điểm thị trường dựa trên dữ liệu thời gian thực. Trên cơ sở đó, các hướng nghiên cứu tiếp theo được đề xuất bao gồm: mở rộng thử nghiệm sang các nhóm cổ phiếu khác (mid-cap, penny), và kết hợp các mô hình theo phương pháp ensemble nhằm tận dụng ưu điểm từng phương pháp.

KẾT LUẬN

Nghiên cứu này đã tập trung đánh giá hiệu quả của các mô hình học máy và học sâu trong việc dự báo lợi nhuận và tối ưu hóa phân bổ danh mục đầu tư, với trọng tâm là so sánh các mô hình XGBoost, Decision Tree Regression và LSTM. Thông qua phân tích thực nghiệm trên dữ liệu cổ phiếu thuộc rổ VN30, nghiên cứu không chỉ làm rõ đặc điểm, ưu – nhược điểm của từng mô hình mà còn cho thấy tiềm năng tích hợp các kỹ thuật hiện đại vào quy trình ra quyết định đầu tư.

Kết quả cho thấy: Các mô hình học máy như XGBoost và học sâu LSTM vượt trội trong việc xử lý dữ liệu thị trường có tính phi tuyến và biến động cao, đồng thời mang lại hiệu suất ổn định trong phân tích lợi nhuận kỳ vọng. Mô hình Decision Tree Regression tuy dễ diễn giải nhưng kém bền vững trước nhiễu và sự thay đổi của thị trường. Đặc biệt, khi kết hợp LSTM với phương pháp Mean-Variance Optimization (MVO), hệ thống không chỉ nâng cao độ chính xác trong dự báo ngắn hạn, mà còn góp phần tạo nên phân bổ danh mục đầu tư phản ánh được quan điểm thị trường một cách linh hoạt và thích ứng.

Dù mang lại nhiều giá trị thực tiễn, nghiên cứu vẫn đối mặt với một số hạn chế như: phạm vi dữ liệu còn hẹp, thời gian phân tích ngắn hạn, và rủi ro quá khớp (overfitting) khi sử dụng mô hình phức tạp với tập dữ liệu hạn chế. Điều này mở ra nhiều hướng phát triển tiềm năng, bao gồm: mở rộng quy mô thử nghiệm sang các nhóm cổ phiếu khác, kiểm định mô hình qua nhiều chu kỳ thị trường, ứng dụng mô hình ensemble để kết hợp điểm mạnh của các phương pháp, và đặc biệt là tích hợp mô hình vào hệ thống hỗ trợ đầu tư tự động (robo-advisor) nhằm phục vụ các nhà đầu tư trong bối cảnh thị trường ngày càng biến động và dữ liệu ngày càng phức tạp.

Nghiên cứu này đã góp phần khẳng định vai trò ngày càng quan trọng của các mô hình học máy và học sâu trong lĩnh vực phân tích tài chính, mở đường cho những ứng dụng tiên tiến và bền vững hơn trong tương lai.

TÀI LIỆU THAM KHẢO

TÀI LIỆU TIẾNG VIỆT

1. Bùi Thị K. (2021). *Vai trò của thị trường chứng khoán trong chính sách kinh tế vĩ mô*. Nhà xuất bản Kinh tế.
2. Cao Văn L. (2020). *Quản lý và đánh giá hiệu quả doanh nghiệp thông qua thị trường chứng khoán*. Tạp chí Kinh tế và Phát triển, 15(3), 45-58.
3. Dương Văn N. (2020). *Chính sách tiền tệ và thị trường chứng khoán*. Nhà xuất bản Lao động.
4. Đinh Thị M. (2021). *Huy động vốn thông qua thị trường trái phiếu chính phủ*. Nhà xuất bản Tài chính.
5. Đỗ Thị F. (2020). *Tính thanh khoản trên thị trường chứng khoán Việt Nam*. Tạp chí Ngân hàng, 8(2), 23-35.
6. Đỗ Thị G. (2021). *Đặc điểm và điều kiện chuyển nhượng cổ phiếu ưu đãi*. Tạp chí Tài chính Doanh nghiệp, 19(4), 23-37.
7. Đỗ Thị Q. (2022). *Phân tán rủi ro trong đầu tư danh mục*. Tạp chí Đầu tư Tài chính, 18(4), 56-71.
8. Hồ, T. Q. (2014). *Ứng dụng một số mô hình tài chính vào công tác quản lý danh mục đầu tư tại Tổng Công ty Tài chính Cổ phần Dầu khí Việt Nam – Chi nhánh Đà Nẵng*.
9. Hoàng Thị E. (2021). *Quyền của cổ đông và quản trị công ty*. Nhà xuất bản Lao động Xã hội.
10. Hoàng Thị O. (2022). *Các yếu tố ảnh hưởng đến cơ cấu danh mục đầu tư*. Nhà xuất bản Thống kê.
11. Hoàng Thị W. (2021). *Các nguyên tắc hoạt động của thị trường chứng khoán*. Tạp chí Khoa học Kinh tế, 12(4), 78-92.
12. Hoàng Văn E. (2021). *Thị trường sơ cấp và thứ cấp: Đặc điểm và vai trò*. Nhà xuất bản Đại học Kinh tế Quốc dân.
13. Lê Thị C. (2021). *Cổ phiếu thường: Đặc điểm và vai trò trong huy động vốn*. Tạp chí Kinh tế và Phát triển, 16(5), 67-81.
14. Lê Thị T. (2020). *Các tổ chức trung gian trên thị trường chứng khoán*. Tạp chí Tài chính, 7(1), 12-25.
15. Lê Văn C. (2019). *Sở giao dịch chứng khoán và vai trò trong thị trường tài chính*. Nhà xuất bản Thống kê.

16. Lê Văn M. (2022). *Đa dạng hóa tài sản trong danh mục đầu tư hiện đại*. Tạp chí Kinh tế và Phát triển, 19(3), 45-62.

17. Lưu Thị O. (2021). *Thu hút đầu tư nước ngoài thông qua thị trường chứng khoán*. Tạp chí Đầu tư, 9(3), 56-68.

18. Lý Thị J. (2021). *Cổ tức tích lũy và không tích lũy: Phân tích so sánh*. Tạp chí Kế toán - Kiểm toán, 8(3), 45-58.

19. Lý Văn I. (2021). *Môi trường đầu tư và quản lý rủi ro*. Nhà xuất bản Khoa học và Kỹ thuật.

20. Ngô Văn G. (2021). *Thị trường phái sinh tài chính: Lý thuyết và thực tiễn*. Nhà xuất bản Đại học Quốc gia.

21. Ngô Văn H. (2021). *Phân chia lợi nhuận và quyền lợi cổ đông*. Nhà xuất bản Thống kê.

22. Ngô Văn R. (2021). *Rủi ro hệ thống và phi hệ thống trong đầu tư chứng khoán*. Nhà xuất bản Đại học Kinh tế TP.HCM.

23. Ngô, T. Thoa. (2018). *Đa dạng hóa danh mục đầu tư sử dụng độ đo rủi ro đa trị* [Luận văn thạc sĩ, Viện Toán học – Viện Hàn lâm Khoa học & Công nghệ Việt Nam].

24. Nguyễn Minh A. (2021). *Lý thuyết cổ phiếu và định giá chứng khoán*. Nhà xuất bản Đại học Kinh tế Quốc dân.

25. Nguyễn Thị S. (2021). *Các chủ thể tham gia thị trường chứng khoán*. Tạp chí Kinh tế và Dự báo, 18(2), 34-47.

26. Nguyễn Văn A. (2020). *Lý thuyết thị trường tài chính*. Nhà xuất bản Đại học Kinh tế TP.HCM.

27. Nguyễn Văn K. (2022). *Lý thuyết và thực hành danh mục đầu tư*. Nhà xuất bản Đại học Kinh tế Quốc dân.

28. Phạm Thị D. (2020). *Thị trường OTC và xu hướng phát triển*. Nhà xuất bản Chính trị Quốc gia.

29. Phạm Văn D. (2020). *Quyền biểu quyết và tham gia quản trị công ty*. Tạp chí Quản lý Nhà nước, 14(7), 89-102.

30. Phạm Văn N. (2021). *Quản lý rủi ro trong đầu tư tài chính*. Tạp chí Ngân hàng, 15(8), 23-38.

31. Phạm Văn V. (2020). *Minh bạch thông tin trên thị trường chứng khoán*. Tạp chí Kế toán và Kiểm toán, 14(6), 67-79.

32. Phan Văn P. (2020). *Huy động vốn doanh nghiệp thông qua thị trường chứng khoán*. Tạp chí Quản lý Kinh tế, 11(4), 89-103.

33. Tô Thị Q. (2021). *Thương hiệu doanh nghiệp và niêm yết chứng khoán*. Nhà xuất bản Tri thức.

34. Trần Thị B. (2021). *Thị trường vốn và tăng trưởng kinh tế*. Nhà xuất bản Khoa học Xã hội.

35. Trần Thị L. (2021). *Quản lý danh mục đầu tư: Từ lý thuyết đến thực tiễn*. Nhà xuất bản Lao động.

36. Trần Văn B. (2020). *Phân loại và đặc điểm các loại cổ phiếu*. Nhà xuất bản Tài chính.

37. Trần Văn U. (2021). *Cạnh tranh và hiệu quả thị trường chứng khoán*. Tạp chí Kinh tế Thế giới, 16(5), 123-138.

38. Trịnh Văn I. (2020). *Quyền biểu quyết có điều kiện của cổ phiếu ưu đãi*. Tạp chí Luật Kinh tế, 12(2), 34-47.

39. Trịnh Văn J. (2020). *Thanh khoản và rủi ro đầu tư chứng khoán*. Nhà xuất bản Tài chính - Marketing.

40. Trịnh Văn S. (2022). *Đường biên hiệu quả và tối ưu hóa danh mục đầu tư*. Tạp chí Kinh tế Định lượng, 14(2), 89-105.

41. Võ Thị H. (2020). *Huy động vốn cho phát triển kinh tế*. Nhà xuất bản Kinh tế TP.HCM.

42. Võ Văn F. (2020). *Cổ phiếu ưu đãi: Lý thuyết và thực tiễn tại Việt Nam*. Nhà xuất bản Khoa học Xã hội.

43. Võ Văn P. (2021). *Hiệu ứng đa dạng hóa trong lý thuyết danh mục hiện đại*. Nhà xuất bản Tài chính.

44. Vũ Văn R. (2020). *Đa dạng hóa danh mục đầu tư*. Tạp chí Đầu tư Tài chính, 13(1), 45-59.

45. Đoàn, N. T. T. H., & Trung, M. (2018). Áp dụng thuật toán phân loại Random Forest để xây dựng bản đồ sử dụng đất/thảm phủ tỉnh Đắk Lắk dựa vào ảnh vệ tinh Landsat 8 OLI. Tạp chí Nông nghiệp vàPhát triển nông thôn, số, 13, 122-129.

TÀI LIỆU TIẾNG ANH

1. Ash Lei. (2025, May 14). *What is the risk of time series forecasting?* BytePlus.

2. Benchaji, I., Douzi, S., El Ouahidi, B., & Jaafari, J. (2021). *Enhanced credit card fraud detection based on attention mechanism and LSTM deep model*. Journal of Big Data, 8(151).
3. Bodie, Z., Kane, A., & Marcus, A. J. (2014). *Investments* (10th ed.). McGraw-Hill Education.
4. Chen, J. (2025, May 31). *Portfolio investment: Definition and asset classes*. Investopedia.
5. Elton, E. J., Gruber, M. J., Brown, S. J., & Goetzmann, W. N. (current edition). *Modern Portfolio Theory and Investment Analysis*. John Wiley & Sons.
6. Espiga-Fernández, F., García-Sánchez, Á., & Ordieres-Meré, J. (2024). *A Systematic Approach to Portfolio Optimization: A Comparative Study of Reinforcement Learning Agents, Market Signals, and Investment Horizons*. Algorithms, 17(12), 570.
7. Evans, R. B. (2021). *Risk, Return, and Modern Portfolio Theory*.
8. Felix A. Gers et al. *Learning to Forget: Continual Prediction with LSTM*.
9. Garcia, J. (n.d.). *What is diversification? Definition & examples*. Investopedia.
10. Hochreiter, S., & Schmidhuber, J. (1997). *Long Short-Term Memory*. Neural Computation, 9(8), 1735–1780.
11. Investopedia. (2024). *Modern Portfolio Theory: What MPT Is and How Investors Use It*.
12. Kamiński B., Jakubczyk M., Szufel P. (2017). *A framework for sensitivity analysis of decision trees*. Central European Journal of Operations Research, 26(1), 135–159.
13. Markowitz, H. (1952). *Portfolio Selection*. The Journal of Finance, 7(1), 77–91.
14. Nadana Ravishankar, T., & Komarasamy, G. (2023). *Application of Time Series Analysis for Better Decision-Making in Business*. Technoarete Transactions on Intelligent Data Mining and Knowledge Discovery, 2(4).
15. Pillai, R. P., & Latha, D. P. (2025). *A deep learning based hybrid model using LSTM and CNN techniques for automated internal fraud detection in banking systems*. Journal of Information Systems Engineering and Management, 10(40s), 98.4% accuracy.
16. Raza, A. (2025, May 16). *How JPMorgan uses AI to save 360,000 legal hours a year*.

17. Risal, N. (2013). *A study on risk-return relationship: The effect of diversification on unsystematic risk*. PRAVAHA Journal, 20(1), 159–168.

18. Sharpe, W. F. (1966). *Mutual Fund Performance*. Journal of Business, 39(1), 119–138.

19. Sima Siami-Namini, & Akbar Siami Namin. (2018). *Forecasting Economics and Financial Time Series: ARIMA vs. LSTM*. ArXiv.

20. Superior Data Science. (2024). *J.P. Morgan – COiN – a case study of AI in finance*.

21. Tulsi, K., Dutta, A., Singh, N., & Jain, D. (2024). *Transforming financial services: The impact of AI on JP Morgan Chase's operational efficiency and decision-making*. International Journal of Scientific Research & Engineering Trends, 10(1), 207–212.

22. Wang, W., Li, W., Zhang, N., & Liu, K. (2020). *Portfolio formation with preselection using deep learning from long-term financial data*. Expert Systems with Applications, 143, 113042.

23. Xiao, R., Feng, Y., Yan, L., & Ma, Y. (2022). *Predict stock prices with ARIMA and LSTM*.

24. Zhang, A., Lipton, Z., Li, M., & Smola, A. J. (2024). *Dive into Deep Learning*. Cambridge University Press.

25. Wang, W., Li, W., Zhang, N., & Liu, K. (2020). Portfolio formation with preselection using deep learning from long-term financial data. *Expert Systems with Applications*, 143, 113042.

26. Gilles Louppe, “Understanding Random Forest from theory to pratic”, University of Liège, Faculty of Applied Sciences, Department of Electrical Engineering & Computer Science, pp. 55-115

PHỤ LỤC

Phụ lục A: Dữ liệu

Dữ liệu giá 10 dòng đầu của 10 mã cổ phiếu từ 2017 – 2025

	time	open	high	low	close	volume	symbol
461	2017-03-02	4750	4750	4660	4730	356057	ACB
462	2017-03-03	4730	4790	4620	4790	658085	ACB
463	2017-03-06	4810	4850	4750	4750	767074	ACB
464	2017-03-07	4790	4790	4700	4730	342874	ACB
465	2017-03-08	4730	5040	4680	4890	3300268	ACB
...	...	...	...	...	...	...	...
64227	2017-11-06	32150	32150	32150	32150	800	VRE
64228	2017-11-07	32150	32150	32150	32150	0	VRE
64229	2017-11-08	32150	32150	32150	32150	0	VRE
64230	2017-11-09	32150	32150	32150	32150	0	VRE
64231	2017-11-10	34370	34370	34370	34370	31030	VRE

150 rows × 7 columns

Phụ lục B: Code Python

Hình 1: Code tạo các chỉ số

```python
def create_addtional_features(group_ticker):
  """
  Create addtional features for the stock data, for prediction returns
  """
  group = group_ticker.copy()

  # tính log-price
  group['return'] = np.log(group['close'].shift(-1) / group['close'])

  # tính r1-> r4: ln(close / close-1)
  for i in range(1,5):
    group[f'r{i}'] = np.log(group['close'].shift(i-1) / group['close'].shift(i))

  # tính r5: ln(high / open)
  group['r5'] = np.log(group['high'] / group['open'])

  # tính r6 - r8: ln(high_t / open_t-n)
  for i in range(1, 4):
    group[f'r{5+i}'] = np.log(group['high'] / group['open'].shift(i))

  # tính r9-r11: ln(high_t-i / open_t-i)
  for i in range(1, 4):
    group[f'r{8+i}'] = np.log(group['high'].shift(i) / group['open'].shift(i))

  # tính r12-r15: ln(low_t-i / open_t-i)
  for i in range(0, 4):
    group[f'r{12+i}'] = np.log(group['low'].shift(i) / group['open'].shift(i))

  # tính Relative Strength Index (RSI- close price, period=14)
  group['r16'] = ta.momentum.RSIIndicator(close=group['close'], window=14).rsi()

  # tính Momentum Index (close price, period-10)
  group['r17'] = ta.momentum.ROCIndicator(close=group['close'], window=10).roc()

  # tính True Range (high, low and close price)
  group['r18'] = ta.volatility.AverageTrueRange(high close: Any '], low=group['low'], close=group['close']).average_true_range()

  # tính Average True Range (high, low, close and period=14)
  group['r19'] = ta.volatility.AverageTrueRange(high=group['high'], low=group['low'], close=group['close'], window=14).average_true_rang

  # tính Parabolic SAR (high, low, acceleration=0.02, maximum=0)
  group['r20'] = ta.trend.PSARIndicator(high=group['high'], low=group['low'], close=group['close'], step=0.02, max_step=0.2).psar()

  # tính Chaikin Oscillator
  group['r21'] = ta.volume.AccDistIndexIndicator(  high=group['high'],
                                                   low=group['low'],
                                                   close=group['close'],
                                                   volume=group['volume']
                                                ).acc_dist_index()
  # tính EMA - tính cho chu kỳ 20
  group['r22'] = ta.trend.EMAIndicator(close=group['close'], window=20).ema_indicator()
```

Hình 2: Code chọn feature

```python
from sklearn.feature_selection import RFE
from sklearn.ensemble import RandomForestRegressor

rf_model = RandomForestRegressor(random_state=42)
rf_model.fit(train_X, train_y)
```

```
RandomForestRegressor
RandomForestRegressor(random_state=42)
```

```python
feature_scores = pd.Series(rf_model.feature_importances_, index=train_X.columns).sort_values(ascending=False)
```

```python
plt.figure(figsize=(10, 6))

# Vẽ thanh
sns.barplot(x=feature_scores.values[:25], y=feature_scores.index[:25], palette='viridis')

# Tính giá trị trung bình
mean_importance = feature_scores.values[:25].mean()

# Vẽ đường trung bình
plt.axvline(mean_importance, color='red', linestyle='--', label=f'Mean = {mean_importance:.4f}')

plt.title("Top 25 Feature Importances - Random Forest")
plt.xlabel("Importance Score")
plt.ylabel("Feature")
plt.legend()
plt.tight_layout()
plt.show()
```

Hình 3: Code huấn luyện mô hình

```
# Hàm dự báo nhiều bước bằng LSTM cho từng mã
# Kết hợp dự báo + smoothing bằng trọng số gần (recent bias)
def forecast_multi_day_lstm(model, recent_sequence, n_days=7):
    preds = []
    current_seq = recent_sequence.copy()
    for step in range(n_days):
        pred = model.predict(current_seq.reshape(1, *current_seq.shape), verbose=0).flatten()[0]
        weight = 1.0 - (step / n_days)  # giảm trọng số cho bước xa hơn
        preds.append(pred * weight)
        next_step = current_seq[-1].copy()
        next_step[-1] = pred
        current_seq = np.append(current_seq[1:], [next_step], axis=0)
    return np.sum(preds) / n_days

# --- Dự báo return từ mô hình LSTM cho từng mã ---
predicted_returns = {}
look_back = 32

for symbol in selected_symbols:
    if symbol not in models_dict:
        continue

    data = symbol_dfs[symbol].copy()
    data = data[selected_features + ['return']].dropna()
    scaler = MinMaxScaler()
    X_scaled = scaler.fit_transform(data[selected_features])
    y = data['return'].values

    def create_sequences(X, y, seq_length=look_back):
        Xs, ys = [], []
        for i in range(len(X) - seq_length):
            Xs.append(X[i:i+seq_length])
            ys.append(y[i+seq_length])
        return np.array(Xs), np.array(ys)

    X_seq, y_seq = create_sequences(X_scaled, y)
    if len(X_seq) == 0:
        continue

    recent_seq = X_seq[-1]
    model = models_dict[symbol]
    y_pred_mean = forecast_multi_day_lstm(model, recent_seq, n_days=7)
    predicted_returns[symbol] = y_pred_mean

# --- Dữ liệu lịch sử để tính MVO truyền thống ---
price_data = pd.DataFrame()

for symbol in selected_symbols:
    df = symbol_dfs[symbol].copy()
    df = df.sort_values('time')
    df = df[['time', 'close']].dropna().set_index('time')
    price_data[symbol] = df['close']
```