



ISSN 1859-3666
E-ISSN 2815-5726

Tạp chí KHOA HỌC THƯƠNG MẠI

TẠP CHÍ CỦA TRƯỜNG ĐẠI HỌC THƯƠNG MẠI



Năm thứ 24 - số 200
4/2025



khoa học thương mại

TẠP CHÍ CỦA TRƯỜNG ĐẠI HỌC THƯƠNG MẠI
BỘ GIÁO DỤC VÀ ĐÀO TẠO

TỔNG BIÊN TẬP:

ĐINH VĂN SƠN

PHÓ TỔNG BIÊN TẬP:

THƯ KÝ TÒA SOẠN

TRƯỞNG BAN TRỊ SỰ

NGUYỄN THỊ QUỲNH TRANG

❑ Tòa soạn

Phòng 202 nhà T

Trường Đại học Thương mại

Số 79 đường Hồ Tùng Mậu

Mai Dịch, Cầu Giấy, Hà Nội

❑ Điện thoại: 024.37643219 máy lẻ 2102

❑ Fax: 024.37643228

❑ Email: tckhtm@tmu.edu.vn

❑ Website: tckhtm.tmu.edu.vn

❑ GP hoạt động báo chí:

Số 195/GP-BTTTT ngày 05/6/2023

❑ Chế bản tại: Tòa soạn

Tạp chí Khoa học Thương mại

❑ In tại: Cty TNHH In & TM Hải Nam

❑ Nộp lưu chiểu: 4/2025

HỘI ĐỒNG KHOA HỌC BIÊN TẬP

Đinh Văn Sơn - Đại học Thương mại (Chủ tịch)

Phạm Vũ Luận - Đại học Thương mại (Phó Chủ tịch)

Nguyễn Bách Khoa - Đại học Thương mại (Phó chủ tịch)

Phạm Minh Đạt - Đại học Thương mại (Ủy viên thư ký)

Các ủy viên

- **Vũ Thành Tự Anh** - ĐH Fulbright Việt Nam (Hoa Kỳ)

- **Lê Xuân Bá** - Viện QLKT TW

- **Hervé B. Boismery** - Đại học Reunion (Pháp)

- **H. Eric Boutin** - Đại học Toulon Var (Pháp)

- **Nguyễn Thị Doan** - Hội Khuyến học Việt Nam

- **Haasis Hans** - Đại học Bremen (Đức)

- **Lê Quốc Hội** - Đại học Kinh tế quốc dân

- **Nguyễn Thị Bích Loan** - Đại học Thương mại

- **Nguyễn Hoàng Long** - Đại học Thương mại

- **Nguyễn Mai** - Chuyên gia kinh tế độc lập

- **Dương Thị Bình Minh** - ĐH Kinh tế Tp Hồ Chí Minh

- **Hee Cheon Moon** - Hội Nghiên cứu TM Hàn Quốc

- **Bùi Xuân Nhàn** - Đại học Thương mại

- **Lương Xuân Quỳ** - Hội Khoa học kinh tế Việt Nam

- **Nguyễn Văn Song** - Học viện Nông nghiệp Việt Nam

- **Nguyễn Thanh Tâm** - Đại học California (Hoa Kỳ)

- **Trương Bá Thanh** - ĐH Kinh tế - Đại học Đà Nẵng

- **Đinh Văn Thành** - Viện Nghiên cứu thương mại

- **Đỗ Minh Thành** - Đại học Thương mại

- **Lê Đình Thắng** - Đại học Québec (Canada)

- **Trần Đình Thiên** - Viện Kinh tế Việt Nam

- **Nguyễn Quang Thuấn** - Viện Hàn lâm KHXH Việt Nam

- **Washio Tomoharu** - ĐH Kwansey Gakuin (Nhật Bản)

- **Lê Như Tuyền** - Grenoble École de Managment (Pháp)

- **Zhang Yujie** - Đại học Tsinghua (Trung Quốc)

MỤC LỤC

KINH TẾ VÀ QUẢN LÝ

- 1. Nguyễn Thành Hưng** - Ảnh hưởng của mức độ công bố thông tin trách nhiệm xã hội đến giá trị doanh nghiệp - nghiên cứu tại các công ty niêm yết trên thị trường chứng khoán Việt Nam. **Mã số: 200.1BAcc.11** 3
- Impact of Social Responsibility Information Disclosure on Firm Value - Case in Listed Companies on Vietnam Stock Market*
- 2. Trần Mạnh Hà và Trần Ngọc Mai** - Ứng dụng mô hình dữ liệu tần suất cao kết hợp học máy nhằm nâng cao hiệu quả dự báo: nghiên cứu với hợp đồng tương lai dầu thô WTI. **Mã số: 200.1HIEM.11** 15
- Enhancing Forecasting Accuracy Through the Use of High-Frequency Data and Machine Learning: A Study on the Crude Oil WTI Futures*

QUẢN TRỊ KINH DOANH

- 3. Đinh Thị Hương, Trần Văn Trang và Nguyễn Thị Thúy Hằng** - Mối quan hệ giữa đổi mới sáng tạo, lợi thế cạnh tranh và hiệu quả kinh doanh của các doanh nghiệp Việt Nam - vai trò tác động của việc thực hiện trách nhiệm xã hội. **Mã số: 200.2BAdm.21** 26
- The Connection Among Innovation, Competitive Advantage, and Business Performance of Vietnamese Enterprises - The Influence of CSR Implementation*
- 4. Nguyễn Thị Thanh Nhân** - Tác động của cấu trúc vốn đến hiệu quả hoạt động của các công ty niêm yết trên thị trường chứng khoán: Một khảo sát ở Việt Nam. **Mã số: 200.2FiBa.21** 45
- The Impact of Capital Structure on the Performance of Listed Companies on the Stock Market: A Survey in Vietnam*

- 5. Lê Xuân Cù** - Hành vi sử dụng chatbot trong bán lẻ điện tử: Vai trò của giá trị thông tin và cảm nhận thông minh. **Mã số: 200.2TrEM.21** 55

Use Behavior Toward Chatbots in E-Retailing: Role of Information Value and Perceived Intelligence

- 6. Phạm Thu Thủy, Nguyễn Thuỳ Linh và Nguyễn Hà Vy** - Tác động của Gamification đến sự gắn kết của khách hàng - nghiên cứu thực nghiệm đối với ngành ngân hàng tại Việt Nam. **Mã số: 200.2FiBa.21** 68

The Impact of Gamification on Customer Engagement: An Experimental Study in the Banking Sector in Vietnam

- 7. Lê Ba Phong và Nguyễn Thị Tuyền** - Tăng cường hành vi chia sẻ tri thức của nhân viên tại các doanh nghiệp vừa và nhỏ: Vai trò của lãnh đạo và an toàn tâm lý. **Mã số: 200.2HRMg.21** 81

Enhancing Knowledge Sharing Behavior Among Employees in Small and Medium-Sized Enterprises: the Role of Leadership and Psychological Safety

Ý KIẾN TRAO ĐỔI

- 8. Nguyễn Thị Hồng** - Tác động của thù lao phi tài chính đến hiệu suất làm việc của cán bộ nghiên cứu tại các tổ chức khoa học và công nghệ ngoài công lập ở Việt Nam. **Mã số: 200.3OMIs.31** 93

The Impact of Non - Financial Remuneration on the Performance of Researchers at Non - Public Science and Technology Organizations in Vietnam

- 9. Lê Việt Hà** - Nghiên cứu các yếu tố ảnh hưởng đến hành vi sử dụng ai của giảng viên các trường đại học tại Việt Nam. **Mã số: 200.3OMIs.31** 105

A Study on Factors Affecting University Lecturers' AI Usage Behavior in Vietnam

ỨNG DỤNG MÔ HÌNH DỮ LIỆU TẦN SUẤT CAO KẾT HỢP HỌC MÁY NHẪM NÂNG CAO HIỆU QUẢ DỰ BÁO: NGHIÊN CỨU VỚI HỢP ĐỒNG TƯƠNG LAI DẦU THÔ WTI

Trần Mạnh Hà *

Email: hatm@hvn.edu.vn

Trần Ngọc Mai *

Email: ngocmai@hvn.edu.vn

* Học viên Ngân hàng

Ngày nhận: 04/12/2024

Ngày nhận lại: 20/03/2025

Ngày duyệt đăng: 23/03/2025

Nghiên cứu đánh giá hiệu quả của việc sử dụng dữ liệu tần suất cao và kỹ thuật học máy trong việc nâng cao hiệu quả dự báo biến động. Sử dụng dữ liệu với tần suất 5 phút của hợp đồng tương lai dầu thô WTI với 182,874 quan sát và so sánh kết quả dự báo ba mô hình - GARCH(1,1) sử dụng dữ liệu tần suất thấp, HAR-RV với dữ liệu tần suất cao, và ML-HAR-RV tích hợp học máy, nhóm tác giả chỉ ra rằng việc sử dụng dữ liệu tần suất cao, đặc biệt khi kết hợp với học máy, sẽ mang lại kết quả vượt trội so với các mô hình dự báo thông thường. Bên cạnh đó, kết quả nghiên cứu cũng cho thấy việc sử dụng một năm dữ liệu lịch sử để đào tạo mô hình mang lại kết quả dự báo chính xác hơn so với việc sử dụng khung dữ liệu lịch sử trong hai năm. Kết quả nghiên cứu sẽ có nhiều ý nghĩa đối với các nhà quản lý danh mục đầu tư cũng như các cơ quan quản lý trong việc nghiên cứu và dự báo biến động của các tài sản tài chính.

Từ khóa: Dữ liệu tần suất cao, học máy, biến động thực tế.

JEL Classifications: C21, G17, G32.

DOI: 10.54404/JTS.2025.200V.02

1. Tổng quan lý thuyết

Dự báo biến động là quá trình dự đoán sự thay đổi trong tương lai của giá tài sản tài chính, thường được biểu thị bằng độ lệch chuẩn hoặc phương sai của lợi nhuận trong một khoảng thời gian cụ thể. Biến động là thước đo rủi ro chính trên thị trường tài chính và phản ánh mức độ dao động của giá tài sản. Dự báo chính xác về biến động có ý nghĩa quan trọng để hiểu được động lực thị trường và đưa ra quyết định quản trị trong các lĩnh vực như quản lý rủi ro, tối ưu hóa danh mục

đầu tư, định giá phái sinh cũng như tuân thủ quy định của cơ quan quản lý (Granger & Poon, 2002). Nhằm dự báo mức độ biến động của tài sản tài chính, các nhà nghiên cứu và chuyên gia tài chính đã sử dụng nhiều mô hình và loại dữ liệu khác nhau, từ dữ liệu giá hàng ngày (tần suất thấp) đến dữ liệu trong ngày (tần suất cao), để giúp nâng cao độ chính xác của công tác dự báo (Chortareas et al., 2011; Liu et al., 2021; Zhou, 2017).

Dữ liệu tần suất cao (High frequency data, viết tắt là HFD) đóng vai trò ngày một quan

Tác giả liên hệ: ngocmai@hvn.edu.vn

trọng trong lĩnh vực dự báo biến động thị trường nhờ khả năng cung cấp thông tin chi tiết về động lực giá cả và hoạt động giao dịch trong các khung thời gian ngắn, chẳng hạn như trong ngày. Không giống như dữ liệu tần suất thấp truyền thống, vốn tổng hợp thông tin qua các khoảng thời gian dài hơn như hàng ngày hoặc hàng tuần, HFD nắm bắt được các chi tiết nhỏ về biến động giá và hoạt động giao dịch trong các ngày giao dịch. Tính chi tiết này cho phép các nhà nghiên cứu và thực hành mô hình hóa biến động với độ chính xác cao hơn, đặc biệt trong các kịch bản ngắn hạn nơi các biến động nhanh chóng đóng vai trò quan trọng. Nhiều nghiên cứu đã chứng minh rằng việc sử dụng dữ liệu trong ngày có thể cải thiện đáng kể độ chính xác của các dự đoán biến động so với các mô hình tần suất thấp truyền thống, nhấn mạnh vai trò của HFD trong phân tích tài chính (Chortareas et al., 2011; Liu et al., 2021; Zhou, 2017).

Sự vượt trội của các mô hình dựa trên HFD so với các mô hình tần suất thấp đặc biệt rõ ràng trong dự báo biến động ngắn hạn, nơi chúng có thể nắm bắt các đặc trưng phức tạp như biến động giá trong ngày, các bước nhảy và các mô hình như cụm biến động (Lyócsa et al., 2021). Nhiều phương pháp tiếp cận khác nhau đã được phát triển để khai thác các thông tin phong phú của HFD, qua đó cung cấp thước đo trực tiếp về biến động giá. Một trong các dạng mô hình đó là các bản điều chỉnh của mô hình GARCH truyền thống, đã kết hợp HFD để nắm bắt tốt hơn tính không đồng nhất phương sai trong ngày (Matei, 2011). Bên cạnh việc đa dạng hoá các mô hình, việc ứng dụng các kỹ thuật học máy, chẳng hạn như học sâu và mô hình tổng hợp, đang ngày được áp dụng rộng rãi vì khả năng xử lý tính phức tạp và tính đa chiều của HFD và dữ liệu lớn, qua đó nâng cao hơn nữa hiệu suất dự báo (Lei et al., 2021; Stroud & Johannes, 2014).

Tuy nhiên, hiệu quả của việc sử dụng dữ liệu tần suất cao có thể thay đổi đáng kể giữa các thị trường và loại tài sản khác nhau. Noh and Kim (2006) nhận thấy rằng hiệu suất của

các mô hình dựa trên HFD phụ thuộc vào các yếu tố như tính thanh khoản của thị trường, hoạt động giao dịch và đặc điểm cấu trúc. Ví dụ, các mô hình HFD có xu hướng mang lại kết quả tốt hơn ở các thị trường có tính thanh khoản cao, nơi dữ liệu giá dồi dào và ít bị nhiễu hơn, trong khi lợi thế của chúng có thể bị hạn chế ở các thị trường ít giao dịch hoặc không thanh khoản. Nhận thức được những hạn chế này, các nhà nghiên cứu đã tìm cách nâng cao độ chính xác bằng cách bổ sung các biến giải thích phụ trợ, như đã tích hợp dữ liệu về sự chú ý của nhà đầu tư, chẳng hạn như chỉ số tìm kiếm trên web hoặc hoạt động trên mạng xã hội, để phản ánh tâm lý thị trường và các yếu tố hành vi, từ đó cải thiện hiệu suất của mô hình HFD trong việc dự đoán biến động (Lei et al., 2021).

Mặc dù có những lợi thế rõ ràng, nhưng tính ưu việt của HFD giảm đi khi chúng ta kéo dài khung thời gian dự báo (Lyócsa et al., 2021). Cụ thể, trong các khoảng thời gian dài, các thông tin xảy ra với tần suất thấp như biến động kinh tế vĩ mô, vốn không được dữ liệu trong ngày nắm bắt đầy đủ, có xu hướng chi phối động lực biến động và do đó làm giảm lợi thế của HFD. Hạn chế này nhấn mạnh sự cần thiết của một phương pháp tiếp cận cân bằng hơn, trong đó HFD được sử dụng cùng với các loại dữ liệu và mô hình khác để đạt được hiệu suất dự báo tối ưu trong các khung thời gian khác nhau. Các nghiên cứu gần đây đã chỉ ra rằng các mô hình kết hợp sử dụng dữ liệu tần suất cao với học máy và học sâu đã mang lại hiệu quả vượt trội so với các mô hình tuyến tính truyền thống và mô hình chuỗi thời gian (Moreno-Pino & Zohren, 2024; Zhu et al., 2023). Với sự tiên bộ không ngừng của các kỹ thuật thống kê và học máy tiếp tục mở rộng phạm vi ứng dụng của HFD, đảm bảo vị trí trung tâm của nó trong việc phân tích biến động tài chính. Những phát triển này không chỉ củng cố vai trò của HFD trong nghiên cứu mà còn định hình cách các nhà quản lý và nhà đầu tư tiếp cận việc dự báo và quản lý rủi ro trong bối cảnh thị trường ngày càng phức tạp.

Nghiên cứu của nhóm tác giả nhằm đánh giá việc sử dụng kết hợp giữa mô hình dự báo sử dụng dữ liệu tần suất cao và học máy có thể mang lại kết quả vượt trội hơn so với các mô hình dự báo truyền thống. Sử dụng dữ liệu tần suất 5 phút của hợp đồng tương lai dầu thô WTI, nhóm tác giả so sánh kết quả dự báo của các mô hình sử dụng dữ liệu tần suất cao, bao gồm mô hình HAR-RV truyền thống và mô hình ML-HAR-RV kết hợp học máy, với mô hình sử dụng dữ liệu tần suất thấp GARCH (1,1) và với biến động thực tế RV. Kết quả nghiên cứu cho thấy sự vượt trội của mô hình ML-HAR-RV trong dự báo biến động.

Nghiên cứu được trình bày gồm 5 phần. Phần 2 trình bày về các mô hình dự báo, phần 3 trình bày về dữ liệu và phương pháp phân tích. Kết quả nghiên cứu được thảo luận ở phần 4 và nghiên cứu được kết luận trong phần 5.

2. Mô hình sử dụng

2.1. Mô hình dự báo sử dụng dữ liệu theo ngày

Có rất nhiều mô hình được sử dụng để dự báo mức độ biến động của tài sản tài chính, trong đó mô hình GARCH₁ được đề xuất bởi Engle (1982) và phát triển bởi Bollerslev (1986), được sử dụng rộng rãi trong lĩnh vực tài chính. Mô hình GARCH (1,1) được chứng minh là rất hiệu quả trong việc dự báo và cho kết quả vượt trội so với các mô hình GARCH khác (Hansen & Lunde, 2005). Do đó, nhóm tác giả sử dụng GARCH(1,1) là mô hình tham chiếu trong việc dự báo biến động sử dụng dữ liệu tần suất theo ngày. Mô hình GARCH(1,1) được diễn giải như sau:

$$\begin{aligned} r_t &= \varepsilon_t, \\ \varepsilon_t &= \sigma_t z_t, \\ \sigma_{t+1}^2 &= \omega + \alpha \varepsilon_t^2 + \beta \sigma_t^2, \\ z_t &\sim \text{i.i.d. } N(0,1) \end{aligned}$$

trong đó: r_t là mức lợi nhuận ngày t của tài sản tài chính, ε_t là biến ngẫu nhiên có kỳ vọng bằng 0 và phương sai bằng 1, σ_t^2 là phương sai có điều kiện của biến ngẫu nhiên tại thời điểm t và ω, α, β là các tham số của mô hình.

2.2. Mô hình dự báo sử dụng dữ liệu tần suất cao

2.2.1. Đo lường biến động thực tế (Realized volatility)

Biến động thực tế (RV) là thước đo về sự biến động thực tế của lợi nhuận tài sản trong một khoảng thời gian cụ thể, thường được tính toán bằng dữ liệu trong ngày có tần suất cao. Không giống như biến động ngụ ý (implied volatility), phản ánh kỳ vọng của thị trường hoặc biến động lịch sử, sử dụng giá đóng cửa trong quá khứ, biến động thực tế là thước đo sau khi đã tính toán được từ dữ liệu quan sát, khiến nó trở thành ước tính khách quan và chính xác hơn về biến động thực sự của thị trường.

Barndorff-Nielsen and Shephard (2002) đã chỉ ra rằng RV có thể được ước tính bằng cách sử dụng lợi nhuận trong ngày. Giả sử vào ngày t có $M+1$ quan sát giá logarit tại các thời điểm trong ngày cách đều nhau $t_i, i = 1, \dots, M+1$. Lợi nhuận trong ngày giữa hai thời điểm t_{i-1} và t_i là $r_{t_i} = p_{t_i} - p_{t_{i-1}}$ trong đó p_{t_i} là giá logarit tại thời điểm t_i . Biến động thực tế vào ngày t, RV_t , được xác định bằng:

$$RV_t = \sum_{i=1}^M r_{t_i}^2 \xrightarrow{M \rightarrow \infty} \int_{t-1}^t \sigma_s^2 ds$$

Do đó, biến động thực tế rất dễ xác định, nó chỉ đơn giản là tổng của các lợi nhuận trong ngày bình phương, nhưng giúp cung cấp một ước lượng chính xác cho biến động thực tế của thị trường.

Phương trình RV cũng cho thấy khi số lượng quan sát trong ngày, M tăng lên, biến động thực tế sẽ tiến tới giới hạn xác suất của nó, tức là phương sai tích hợp. Điều này cho thấy giá trong ngày nên được quan sát ở tần suất cao nhất có thể, chẳng hạn như từng giao dịch. Tuy nhiên, giá giao dịch bị ảnh hưởng bởi nhiều cấu trúc vi mô, ví dụ nổi bật nhất là sự bật giá thầu - giá chào bán. Điều này đề cập đến xu hướng các giao dịch thay đổi giữa giá thầu và giá chào bán theo thời gian. Do đó, ngay cả khi không có thay đổi nào về giá thực, sự thay đổi về giá giao dịch vẫn sẽ được quan sát thấy. Có những hiện tượng cấu trúc vi mô khác cũng làm ảnh hưởng đến giá giao dịch. Ví dụ bao gồm quy mô lệnh, thời gian

lệnh đến, tác động của giá, khả năng phục hồi giá và hiệu quả thị trường (Kyle & Obizhaeva, 2016). Cuối cùng, nhiều do các vấn đề về cấu trúc vì mô gây ra có xu hướng dẫn đến ước tính RV bị thổi phồng (Andersen et al., 2003). Để giảm thiểu nhiều cấu trúc vì mô, chúng tôi tuân theo tài liệu và sử dụng kỹ thuật phổ biến là sử dụng tần suất lấy mẫu trong ngày thô. Chuẩn mực đã phát triển là sử dụng lưới giá trong ngày cách nhau 5 phút. Do đó, chúng tôi cũng sử dụng dữ liệu với tần suất 5 phút để tính RV.

2.2.2. Mô hình tự hồi quy không đồng nhất HAR-RV

Mô hình tự hồi quy không đồng nhất (Heterogeneous Autoregressive Realized Volatility - viết tắt là HAR-RV), được giới thiệu bởi Corsi (2009), là một trong các mô hình phổ biến nhất được sử dụng trong việc mô hình và dự báo biến động thực tế của tài sản tài chính. Mô hình HAR-RV dựa trên giả thuyết rằng những người tham gia thị trường hoạt động trong các khoảng thời gian khác nhau (như hàng ngày, hàng tuần, hàng tháng) và chính sự không đồng nhất này góp phần tạo nên mức độ biến động của tài sản tài chính. Mô hình HAR-RV có dạng:

$$r_t = \varepsilon_t,$$

$$\varepsilon_t = \sigma_t z_t,$$

$$z_t \sim \text{i.i.d. } N(0,1)$$

$$RV_t = \gamma_0 + \gamma_1 RV_{t-1,1} + \gamma_2 RV_{t-5,5} + \gamma_3 RV_{t-22,22} + \eta_t$$

$$v_t \sim \text{NID}(0,1),$$

$$\text{trong đó } RV_{t-h,h} = \frac{1}{h} \sum_{i=0}^{h-1} RV_{t-h+i}$$

Để dự báo RV_t trước hết cần tính toán được độ trễ RV theo ngày ($RV_{t-1,1}$), theo tuần ($RV_{t-5,5}$) và theo tháng ($RV_{t-22,22}$), qua đó sử dụng mô hình hồi quy để xác định được các tham số γ_j , qua đó dự báo được biến động thực tế RV_t .

Nhờ sự cân bằng giữa tính đơn giản và hiệu quả, mô hình HAR-RV vẫn là nền tảng trong việc dự báo biến động của tài sản tài chính và là lựa chọn phổ biến trong cả lĩnh vực học thuật lẫn thực tế ứng dụng trong ngành tài chính dự báo và tính sẵn có ngày

càng tăng của dữ liệu tần suất cao đã thúc đẩy việc ứng dụng HAR-RV trong thực tế.

2.2.3. Ứng dụng học máy trong mô hình HAR-RV

Mô hình HAR-RV dựa trên độ trễ của RV trong các khoảng thời gian khác nhau (hàng ngày, hàng tuần và hàng tháng). Các độ trễ này có thể tồn tại hiện tượng đa cộng tuyến (như độ trễ hàng ngày có thể có tương quan mạnh với giá trị trung bình hàng tuần), điều này có thể dẫn đến ước tính không ổn định của các hệ số trong hồi quy bình phương tối thiểu thông thường (OLS). Hồi quy Ridge giúp giảm thiểu vấn đề này.

Hồi quy Ridge, được giới thiệu bởi Hoerl and Kennard (1970), là một kỹ thuật học máy cơ bản áp dụng phương pháp điều chuẩn L2 (L2 regulation) vào mô hình hồi quy tuyến tính. Nó mở rộng khả năng của mô hình hồi quy tuyến tính truyền thống để xử lý hiện tượng đa cộng tuyến, quá khớp (overfitting) và các tập dữ liệu có chiều cao, khiến nó trở thành một phương pháp được sử dụng rộng rãi trong mô hình dự đoán. Hồi quy Ridge đặc biệt hữu ích khi các biến dự báo có tương quan với nhau, thường xảy ra trong các mô hình chuỗi thời gian tài chính như HAR-RV.

Trong mô hình HAR-RV kết hợp học máy (sau đây viết tắt là ML-HAR-RV), hồi quy Ridge được thiết kế để tối ưu hàm chi phí như sau:

$$J(\beta) = \frac{1}{2n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \frac{\lambda}{2} \sum_j \beta_j^2$$

$$J(\beta) = \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p \beta_j^2$$

Trong đó:

γ_j là biến phụ thuộc, trong trường hợp của mô hình HAR-RV thì biến phụ thuộc là RV

x_{ij} là biến dự báo, bao gồm các chuỗi RV với độ trễ theo ngày, theo tuần và theo tháng.

β_0, β_j là hệ số chặn và hệ số trượt của mô hình hồi quy

λ là tham số điều chuẩn. λ càng lớn thì điều chuẩn càng mạnh.

3. Dữ liệu sử dụng và phương pháp phân tích

Nghiên cứu sử dụng dữ liệu tần suất cao của hợp đồng tương lai dầu thô WTI từ ngày 01/01/2014 đến 27/6/2016, với tần suất dữ

liệu 5 phút. Lý do nhóm tác giả sử dụng tần suất 5 phút với dữ liệu trên, bởi nó hài hoà được độ mảnh của dữ liệu cũng như giảm được độ nhiễu do cấu trúc vĩ mô của thị trường. Đồng thời, tần suất dữ liệu 5 phút cũng được chứng minh có hiệu quả vượt trội so với các tần suất khác (Liu et al., 2015).

Chuỗi dữ liệu trong nghiên cứu này bao gồm 799 ngày với 182,874 quan sát (trung bình một ngày có 229 quan sát), thể hiện biến động của giá hợp đồng tương lai dầu thô WTI mỗi 5 phút. Đây là dữ liệu có tính phí và được thu thập từ trang tickdata.com.

Để đánh giá hiệu quả dự báo biến động với việc sử dụng dữ liệu tần suất cao và phương pháp học máy, nghiên cứu đánh giá kết quả dự báo với ba mô hình: (1) Mô hình GARCH(1,1) sử dụng dữ liệu thông thường, (2) mô hình HAR-RV với dữ liệu tần suất cao và (3) mô hình ML-HAR-RV sử dụng dữ liệu tần suất cao kết hợp với học máy. Để so sánh hiệu quả dự báo, nghiên cứu sử dụng phương pháp sai số toàn phương trung bình (Root Mean Square Error - RMSE) để đánh giá độ lệch giữa giá trị ước lượng của mô hình và số liệu thực tế. Mô hình có MSE càng nhỏ thì hiệu quả dự báo càng cao.

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

Trong đó, y_i là giá trị thực tế của biến động, được thể hiện qua RV

$\hat{y}_i$ là giá trị ước lượng của biến động do mô hình dự báo đưa ra.

4. Kết quả nghiên cứu

Nhóm tác giả sử dụng kỹ thuật dự báo dựa trên cửa sổ trượt và vòng lặp nhằm tăng tính chính xác và cập nhật các thông tin có trong dữ liệu mới phát sinh trong ngày làm việc tiếp theo. Cụ thể, nhóm sử dụng dữ liệu lịch sử trong 2 năm, tương ứng với 503 ngày làm việc để ước lượng tham số cho các mô hình GARCH, HAR-RV và ML-HAR-RV và dự báo mức độ biến động cho ngày thứ 504 trong chuỗi thời gian. Tiếp theo, cửa sổ trượt sẽ di chuyển tịnh tiến, sử dụng dữ liệu từ ngày thứ 2 cho đến ngày 504 để xác định lại tham số và dự báo mức độ biến động tại ngày 505. Sử dụng cửa sổ trượt và vòng lặp như vậy, nhóm

tác giả thu được dữ liệu dự báo của 276 ngày tiếp theo và so sánh với biến động thực tế của ngày đó RV_t để đánh giá độ chính xác của từng mô hình. Để minh họa, nhóm tác giả trình bày kết quả ước lượng sử dụng dữ liệu của sổ trượt đầu tiên với các mô hình

$$RV_t = \gamma_0 + \gamma_1 RV_{t-1,1} + \gamma_2 RV_{t-5,5} + \gamma_3 RV_{t-22,22} + \eta_t$$

GARCH(1,1) và HAR-RV trong bảng dưới.

Có thể thấy rằng kết quả ước lượng các tham số của mô hình GARCH(1,1) cho thấy giá trị tham số GARCH tiệm cận 1 và tham số ARCH tiệm cận 0, điều này phù hợp với kết quả ước lượng chung với các mô hình GARCH, chỉ ra rằng phương sai có điều kiện của ngày hôm nay phụ thuộc phần lớn vào phương sai có điều kiện của ngày trước đó hơn là vào các thông tin mới phát sinh trong ngày. Với mô hình HAR-RV, kết quả ước lượng cho thấy tham số γ_1 và γ_3 có giá trị dương và có ý nghĩa thống kê, nhưng γ_2 không có ý nghĩa thống kê, hàm ý rằng phương sai có điều kiện của ngày hôm nay phụ thuộc nhiều vào phương sai có điều kiện của ngày trước đó và trung bình của tháng trước đó, chứ không phụ thuộc vào phương sai trung bình của tuần trước đó. Đặc biệt, giá trị ước lượng của γ_1 và γ_3 cho thấy mức độ phụ thuộc vào giá trị trung bình tháng nhiều hơn gấp đôi so với mức độ phụ thuộc vào ngày liền kề trước đó.

Nhóm tác giả trình bày kết quả dự báo của 03 mô hình và so sánh với biến động thực tế RV trong Biểu đồ 1. Từ kết quả hiển thị trên biểu đồ, có thể thấy rằng các giá trị dự báo được tạo ra bởi mô hình HAR-RV và ML-HAR-RV đều theo sát biến động thực tế, trong đó mức độ sai lệch của HAR-RV lớn hơn ML-HAR-RV. Ngược lại, mô hình GARCH(1,1) đưa ra các giá trị dự báo có sai lệch tương đối lớn và không bám sát được xu hướng biến động của biến dự báo, đặc biệt sau giai đoạn thị trường có biến động mạnh. Đây cũng là một trong các đặc trưng và điểm yếu của mô hình GARCH, đặc trưng bởi hiện tượng biến động cụm (volatility clustering) khi biến động của thời điểm hiện tại chịu ảnh hưởng lớn bởi các ngày trước đó.

Bảng 1: Kết quả ước lượng tham số của mô hình GARCH(1,1)

Tham số	Giá trị ước lượng	Sai số chuẩn	Thống kê t	Giá trị p
ω	1.0356e-06	1.2729e-06	0.81363	0.19059
α	0.069204	0.015997	4.3261	1.5179e-05
β	0.9308	0.014803	62.878	0

Ghi chú: Bảng 1 trình bày kết quả ước lượng của mô hình GARCH(1,1) sử dụng dữ liệu cửa sổ trượt đầu tiên, bằng phương pháp ước lượng hợp lý cực đại (Maximum likelihood estimation).

(Nguồn: tác giả ước lượng trên phần mềm Matlab)

Bảng 2: Kết quả ước lượng tham số của mô hình HAR-RV

Tham số	Giá trị ước lượng	Sai số chuẩn	Thống kê t	Giá trị p
γ_0	0.00096756	0.00073824	1.3106	0.19059
γ_1	0.31895	0.050528	6.3122	6.0908e-10
γ_2	0.0027254	0.10784	0.025272	0.97985
γ_3	0.64202	0.10824	5.9316	5.6181e-09

Ghi chú: Bảng 2 trình bày kết quả ước lượng của mô hình HAR-RV sử dụng dữ liệu cửa sổ trượt đầu tiên, bằng phương pháp ước lượng bình phương nhỏ nhất (Ordinary Least Square).

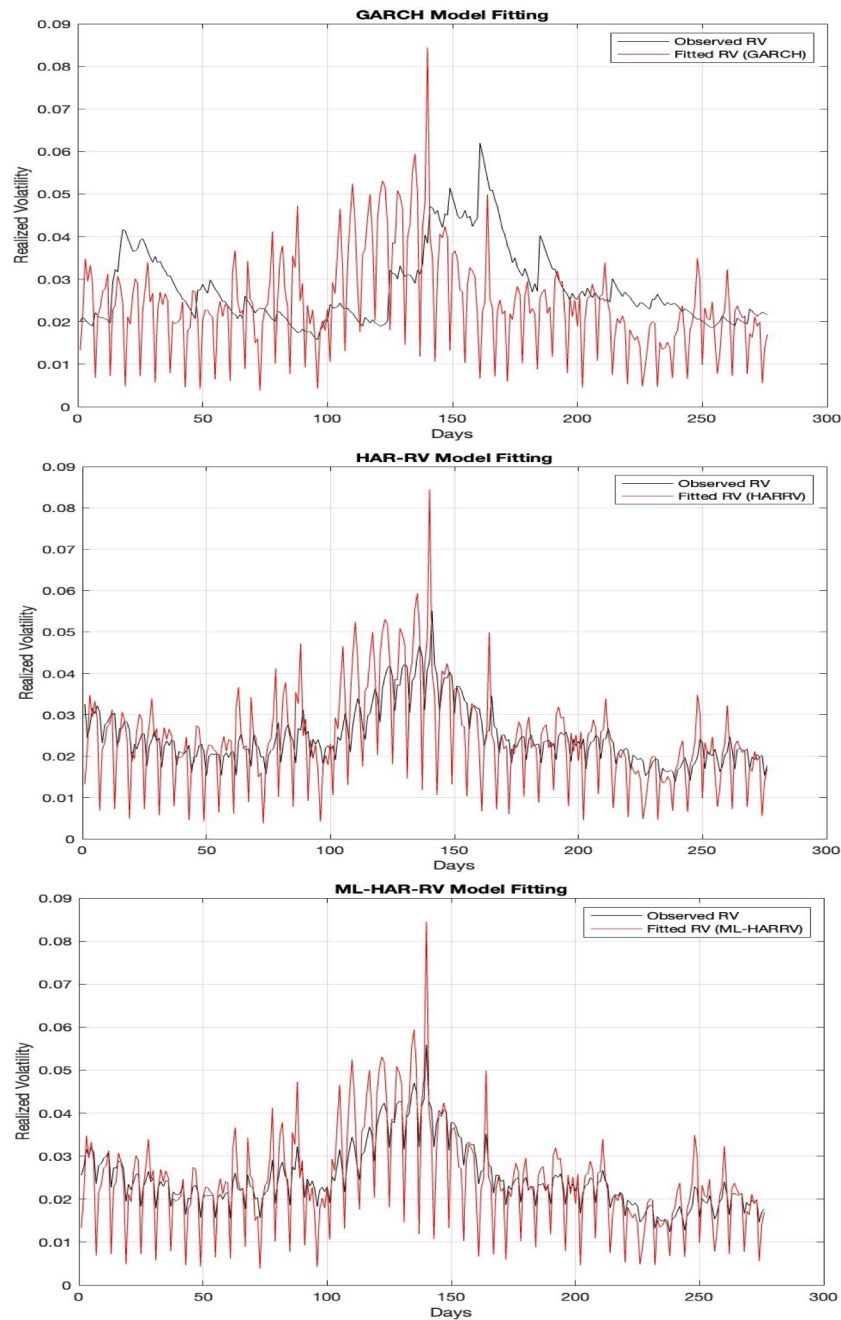
(Nguồn: Tác giả ước lượng trên phần mềm Matlab)

Đánh giá kết quả dự báo thông qua việc so sánh sai lệch giữa giá trị ước lượng và biến động thực tế (đại diện bởi RV) của giá hợp đồng tương lai dầu WTI, có thể thấy rằng trong 03 mô hình, mô hình kết hợp giữa HAR-RV và học máy (ML-HAR-RV) mang lại kết quả vượt trội so với hai mô hình còn lại và năng lực dự báo rất sát với biến động thực tế RV. Điều này được thể hiện qua giá trị sai số toàn phương trung bình RSME. Mô hình GARCH(1,1) cho ra sai số RSME lớn nhất, thể hiện rằng đây là mô hình kém chính xác nhất trong 03 mô hình dự báo. Mô hình HAR-RV, mặc dù sử dụng dữ liệu tần suất cao và có RSME thấp hơn so với GARCH(1,1), nhưng xét về độ lớn sai số của HAR-RV vẫn tương ứng khoảng 68,75% so với GARCH(1,1). Tuy nhiên, khi ứng dụng cùng với học máy, mô hình HAR-RV cho ra kết quả cải thiện đáng kể. Mô hình ML-HAR-RV cho giá trị RMSE bằng 0.0068, mức độ sai số giảm đi 53% so với GARCH(1,1) và nhỏ hơn 32% so với HAR-RV. Điều đó thể hiện tính ưu việt

của kỹ thuật học máy khi áp dụng vào các mô hình dự báo truyền thống nhằm nâng cao tính chính xác của mô hình.

Kiểm tra tính ổn định của kết quả nghiên cứu

Để đánh giá liệu kết quả nghiên cứu có đảm bảo tính vững hay không, nhóm tác giả thay đổi kích thước của cửa sổ trượt và so sánh kết quả dự báo giữa các mô hình. Thay vì dùng dữ liệu lịch sử của 503 ngày (hay 2 năm), nhóm sử dụng dữ liệu trong một năm (hay 252 ngày làm việc) để ước lượng tham số của các mô hình và dự báo biến động cho ngày $t+1$. Phương pháp cửa sổ trượt tiếp tục được áp dụng để có kết quả dự báo cho 547 ngày tiếp theo và so sánh kết quả với biến động thực tế RV. Kết quả được trình bày ở Bảng 4 dưới đây. So sánh giá trị RMSE giữa các mô hình cho thấy kết quả nghiên cứu một lần nữa khẳng định tính ưu việt của mô hình ML-HAR-RV so với HAR-RV và GARCH(1,1) trong việc đưa ra giá trị dự báo tiệm cận mức độ biến động thực tế RV. Một



(Nguồn: Tác giả biểu đồ hóa trên phần mềm Matlab)

Biểu đồ 1: Kết quả dự báo của các mô hình và so sánh với biến động thực tế sử dụng
khung dữ liệu lịch sử trong 02 năm liền trước

Ghi chú: Biểu đồ 1 thể hiện kết quả dự báo của 03 mô hình và so sánh với biến động thực tế RV qua từng ngày trong 276 ngày dự báo. Đường màu đỏ thể hiện kết quả dự báo của mô hình dựa vào dữ liệu quá khứ, đường màu đen thể hiện biến động thực tế của ngày đó được thể hiện qua giá trị RV.

Bảng 3: So sánh kết quả dự báo của các mô hình với biến động thực tế

	GARCH(1,1)	HAR-RV	ML-HAR-RV	RV
Giá trị nhỏ nhất	0,0157	0,0138	0,0124	0,0038
Giá trị lớn nhất	0,0620	0,0551	0,0559	0,0844
Giá trị trung bình	0,0279	0,0247	0,0248	0,0242
Giá trị RMSE	0,0144	0,0099	0,0068	0

Ghi chú: Bảng 3 trình bày kết quả đánh giá độ chính xác của các mô hình dự báo sử dụng của số trượt với dữ liệu lịch sử bao gồm 503 ngày. Các mô hình dự báo bao gồm GARCH(1,1), HAR-RV và ML-HAR-RV và so sánh với giá trị thực tế được đại diện bởi RV. Với hàng RMSE, giá trị càng nhỏ thể hiện mô hình dự báo càng chính xác.

(Nguồn: Tác giả tính toán trên phần mềm Matlab)

Bảng 4: Kiểm tra tính ổn định của mô hình

	GARCH(1,1)	HAR-RV	ML-HAR-RV	RV
Giá trị nhỏ nhất	0,0143	0,0124	0,0122	0,0035
Giá trị lớn nhất	0,0614	0,0553	0,0522	0,0844
Giá trị trung bình	0,0251	0,0240	0,0233	0,0237
Giá trị RMSE	0,0130	0,0096	0,0074	0

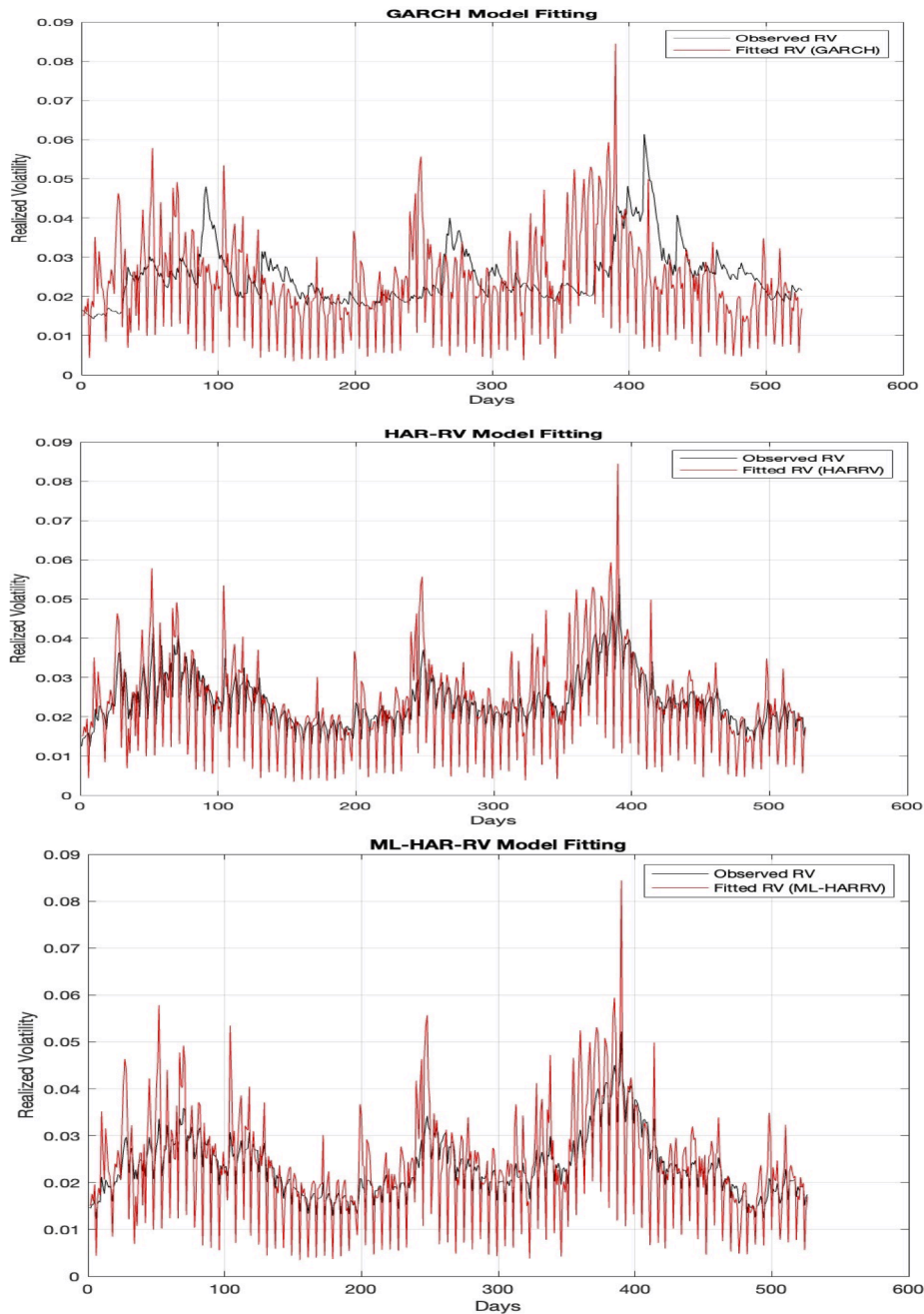
Ghi chú: Bảng 4 trình bày kết quả đánh giá độ chính xác của các mô hình dự báo sử dụng của số trượt với dữ liệu lịch sử bao gồm 252 ngày. Các mô hình dự báo bao gồm GARCH(1,1), HAR-RV và ML-HAR-RV và so sánh với giá trị thực tế được đại diện bởi RV. Với hàng RMSE, giá trị càng nhỏ thể hiện mô hình dự báo càng chính xác.

(Nguồn: Tác giả tính toán trên phần mềm Matlab)

điều thú vị nữa khi so sánh kết quả dự báo giữa 02 kích cỡ cửa sổ trượt (503 ngày và 252 ngày), cho thấy rằng nếu chỉ sử dụng 01 năm dữ liệu lịch sử, các mô hình GARCH(1,1) và HAR-RV và ML-HAR-RV cho ra kết quả dự báo chính xác hơn so với việc sử dụng dữ liệu 02 năm để dự báo giai đoạn tiếp theo. Điều này cũng phù hợp do nếu dữ liệu lịch sử càng dài, các biến cố đã xảy ra lâu trong quá khứ vẫn theo lý thuyết vẫn có khả năng tác động đến kết quả dự báo của ngày hiện tại, điều đó làm giảm hiệu quả dự báo. Đây cũng là điểm gợi mở cho các nhà nghiên cứu cũng như nhà quản trị danh mục trong việc lựa chọn khung dữ liệu lịch sử phù hợp để cho ra kết quả dự báo tốt nhất.

Kết quả nghiên cứu có một số hàm ý chính sách. Dưới góc độ vĩ mô, các cơ quan quản lý cần đầu tư vào hệ thống công nghệ để thu thập và cung cấp dữ liệu tần suất cao, tạo

điều kiện cho việc giám sát thị trường và dự báo biến động ở cấp độ vĩ mô, đồng thời nâng cao khả năng tiếp cận dữ liệu tần suất cao, đảm bảo tính minh bạch và công bằng cho tất cả các bên tham gia thị trường. Bên cạnh đó, các cơ quan giám sát tài chính có thể sử dụng các mô hình dự báo cải tiến từ dữ liệu tần suất cao và học máy để phát hiện sớm các rủi ro hệ thống, từ đó đưa ra các biện pháp can thiệp kịp thời nhằm ổn định thị trường, đặc biệt trong các thời điểm chịu tác động bởi các thông tin bất lợi. Dưới góc độ vĩ mô, các nhà đầu tư và nhà quản trị danh mục có thể áp dụng các mô hình học máy kết hợp với dữ liệu tần suất cao để tối ưu hóa danh mục đầu tư và dự đoán biến động thị trường chính xác hơn để có thể đưa ra các quyết định đầu tư và quản lý rủi ro nhanh chóng, đặc biệt trong giai đoạn thị trường có những biến động ngoài dự kiến.



(Nguồn: tác giả biểu đồ hóa trên phần mềm Matlab)

Biểu đồ 2: Kết quả dự báo của các mô hình và so sánh với biến động thực tế sử dụng
khung dữ liệu lịch sử trong 01 năm liền trước

Ghi chú: Biểu đồ 2 thể hiện kết quả dự báo của 03 mô hình và so sánh với biến động thực tế RV qua từng ngày trong 547 ngày dự báo. Đường màu đỏ thể hiện kết quả dự báo của mô hình dựa vào dữ liệu quá khứ, đường màu đen thể hiện biến động thực tế của ngày đó được thể hiện qua giá trị RV.

5. Kết luận

Nghiên cứu này đã khẳng định vai trò quan trọng của dữ liệu tần suất cao (HFD) trong việc nâng cao độ chính xác của dự báo biến động tài chính, đặc biệt khi kết hợp với các kỹ thuật học máy tiên tiến. Bằng cách so sánh hiệu suất của ba mô hình - GARCH(1,1) sử dụng dữ liệu tần suất thấp, HAR-RV với dữ liệu tần suất cao và ML-HAR-RV tích hợp học máy - nghiên cứu cho thấy mô hình kết hợp giữa việc sử dụng dữ liệu tần suất cao và học máy ML-HAR-RV vượt trội so với mô hình dữ liệu tần suất cao HAR-RV và mô hình sử dụng dữ liệu theo ngày GARCH(1,1). Kết quả này nhấn mạnh rằng việc khai thác thông tin chi tiết từ HFD, kết hợp với khả năng xử lý đa cộng tuyến và tính phức tạp của hồi quy Ridge, không chỉ cải thiện đáng kể hiệu suất dự báo biến động ngắn hạn mà còn theo sát biến động thực tế (RV) của hợp đồng tương lai dầu thô WTI. Tuy nhiên, nghiên cứu cũng chỉ ra rằng hiệu quả của các mô hình phụ thuộc vào khung dữ liệu lịch sử, với cửa sổ 252 ngày mang lại kết quả chính xác hơn so với 503 ngày, gợi ý rằng dữ liệu quá dài có thể làm giảm hiệu quả dự báo do ảnh hưởng của các biến cố xa trong quá khứ.

Những phát hiện này mang lại nhiều đóng góp cả về mặt lý thuyết lẫn thực tiễn. Về mặt lý thuyết, nghiên cứu củng cố vai trò của HFD trong việc nắm bắt các đặc trưng phức tạp như cụm biến động và các bước nhảy, đồng thời chứng minh tiềm năng của học máy trong việc nâng cao các mô hình truyền thống như HAR-RV. Về mặt thực tiễn, kết quả nghiên cứu cung cấp cơ sở cho các nhà đầu tư và quản trị rủi ro để tối ưu hóa danh mục đầu tư và ra quyết định nhanh chóng trong các thị trường biến động, đồng thời khuyến khích các cơ quan quản lý đầu tư vào cơ sở hạ tầng dữ liệu tần suất cao và ứng dụng các mô hình tiên tiến nhằm giám sát và ổn định thị trường tài chính. Dẫu vậy, hạn chế của HFD trong dự báo dài hạn, nơi các yếu tố kinh tế vĩ mô

chiếm ưu thế, cho thấy cần có một cách tiếp cận tích hợp, kết hợp HFD với dữ liệu tần suất thấp để đạt hiệu quả tối ưu trên các khung thời gian khác nhau. Các nghiên cứu trong tương lai có thể mở rộng bằng cách thử nghiệm các kỹ thuật học máy khác hoặc khám phá tác động của các biến hành vi bổ sung nhằm nâng cao hơn nữa độ chính xác dự báo, đặc biệt trong bối cảnh thị trường tài chính ngày càng phức tạp và biến động. ♦

Tài liệu tham khảo:

- Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2), 579-625.
- Barndorff-Nielsen, O. E., & Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(2), 253-280.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3), 307-327.
- Chortareas, G., Jiang, Y., & Nankervis, J. C. (2011). Forecasting exchange rate volatility using high-frequency data: Is the euro different? *International Journal of Forecasting*, 27(4), 1089-1107.
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2), 174-196.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the econometric society*, 987-1007.
- Granger, C. W., & Poon, S.-H. (2002). Forecasting volatility in financial markets: A review (revised edition). *Available at SSRN 331800*.

Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: does anything beat a GARCH (1, 1)? *Journal of applied econometrics*, 20(7), 873-889.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.

Kyle, A. S., & Obizhaeva, A. A. (2016). Market microstructure invariance: Empirical hypotheses. *Econometrica*, 84(4), 1345-1404.

Lei, B., Zhang, B., & Song, Y. (2021). Volatility forecasting for high-frequency financial data based on web search index and deep learning model. *Mathematics*, 9(4), 320.

Liu, L. Y., Patton, A. J., & Sheppard, K. (2015). Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes. *Journal of econometrics*, 187(1), 293-311.

Liu, M., Lee, C.-C., & Choo, W.-C. (2021). The role of high-frequency data in volatility forecasting: evidence from the China stock market. *Applied Economics*, 53(22), 2500-2526.

Lyócsa, Š., Molnár, P., & Výrost, T. (2021). Stock market volatility forecasting: Do we need high-frequency data? *International Journal of Forecasting*, 37(3), 1092-1110.

Matei, M. (2011). Non-Linear volatility modeling of economic and financial time series using high frequency data. *Romanian Journal of Economic Forecasting*, 2, 2011.

Moreno-Pino, F., & Zohren, S. (2024). Deepvol: Volatility forecasting from high-frequency data with dilated causal convolutions. *Quantitative Finance*, 24(8), 1105-1127.

Noh, J., & Kim, T.-H. (2006). Forecasting volatility of futures market: the S&P 500 and FTSE 100 futures using high frequency returns and implied volatility. *Applied Economics*, 38(4), 395-413.

Stroud, J. R., & Johannes, M. S. (2014). Bayesian modeling and forecasting of 24-hour high-frequency volatility. *Journal of the American Statistical Association*, 109(508), 1368-1384.

Zhou, J. (2017). Forecasting REIT volatility with high-frequency data: a comparison of alternative methods. *Applied Economics*, 49(26), 2590-2605.

Zhu, H., Bai, L., He, L., & Liu, Z. (2023). Forecasting realized volatility with machine learning: Panel data perspective. *Journal of Empirical Finance*, 73, 251-271.

Summary

The study evaluates the effectiveness of using high-frequency data and machine learning techniques in improving forecasting accuracy. Utilizing 5-minute interval data from Crude oil WTI futures contracts with 182,874 observations, we compare the forecasting power of the GARCH(1,1) model with daily data, the HAR-RV model with intraday data and the hybrid ML-HAR-RV model integrated with machine learning, we find that the use of high-frequency data, especially when combined with machine learning, yields superior results compared to traditional forecasting models. Besides, this study shows that the performance of the models depends on the historical data window, with a 252-day window yielding more accurate results than a 503-day one, suggesting that the longer pre-sample data may reduce forecast accuracy. These findings of this study hold significant implications for portfolio managers and regulatory authorities in researching and predicting the volatility of financial assets.