Microstructure in the Machine Age

David Easley, Marcos López de Prado, Maureen O'Hara, Zhibai Zhang

2019

Lời cảm ơn

Chúng tôi xin cảm ơn Lee Cohn, Michael Lewis, Michael Lock và Yaxiong Zeng vì các cuộc thảo luận hữu ích và hỗ trợ trong việc chuẩn bị một số dữ liệu được sử dụng trong nghiên cứu này. Chúng tôi cũng cảm ơn các thành viên tham dự hội thảo tại Đại học Cornell, bao gồm Pamela Moulton và Shawn Mankad; và tại Hội nghị FinTech của Đại học Bang Georgia, bao gồm các nhà phản biện và người thảo luận của chúng tôi tại hội nghị, Haoxiang Zhu.

Tóm tắt

Chúng tôi chứng minh cách một thuật toán học máy có thể được áp dụng để dự đoán và giải thích các hiện tượng vi cấu trúc thị trường hiện đại. Chúng tôi khảo sát hiệu quả của nhiều thước đo vi cấu trúc khác nhau và cho thấy rằng chúng tiếp tục mang lại những hiểu biết về động lực giá trong các thị trường phức tạp hiện nay. Một số đặc điểm vi cấu trúc với khả năng giải thích cao trong mẫu lại có sức mạnh dự báo thấp, và ngược lại. Chúng tôi cũng nhận thấy rằng một số thước đo dựa trên vi cấu trúc hữu ích cho việc dự báo ngoài mẫu các thống kê thị trường khác nhau, dẫn đến những câu hỏi về tính hiệu quả của thị trường. Kết quả của chúng tôi được rút ra bằng cách sử dụng 87 hợp đồng tương lai có tính thanh khoản cao nhất trên tất cả các loại tài sản.

Từ khóa: Vi cấu trúc thị trường, học máy, tầm quan trọng đặc trưng, MDI, MDA, hợp đồng tương lai, kinh tế lượng.

1 Giới thiệu

Người ta có thể kỳ vọng rằng khi thị trường trở nên nhanh hơn, dữ liệu thị trường ngày càng nhiều hơn, và công nghệ thay thế con người tham gia, thì

vi cấu trúc thị trường sẽ ngày càng ít đóng vai trò trong việc giải thích hành vi thị trường. Thực tế lại ngược lại. Khi thang đo thời gian thu hẹp xuống nano giây, cách thức thị trường được cấu trúc hóa ra lại mang tính quyết định trong việc dự đoán hướng đi của thị trường. Và khi dữ liệu bùng nổ đến kích thước khổng lồ, việc có thể đặc trưng hóa các biến liên quan đến ma sát thị trường có thể và nên quan trọng đối với hành vi thị trường – một trọng tâm đặc biệt của nghiên cứu vi cấu trúc – càng trở nên quan trọng hơn. Tuy nhiên, mặc dù vẫn giữ vai trò quan trọng, nghiên cứu vi cấu trúc đối mặt với những thách thức đáng kể trong kỷ nguyên mới này.

Sự phổ biến của giao dịch bằng máy tính, được thúc đẩy bởi sự trỗi dậy của dữ liệu lớn, đã làm gia tăng sự phức tạp của các chiến lược giao dịch vượt xa những gì được hình dung trong các mô hình vi cấu trúc đơn giản. Tương tự, các thước đo thực nghiệm vốn là công cụ trong "hộp công cụ" vi cấu trúc được xây dựng dựa trên các mối quan hệ đơn giản trong nội bộ thị trường có thể không còn đúng trong thế giới tần suất cao của giao dịch xuyên thị trường. Vấn đề, nói một cách đơn giản, là vi cấu trúc cần phải tiến hóa.

Trong bài báo này, chúng tôi chứng minh cách các kỹ thuật học máy có thể đóng vai trò quan trọng trong quá trình tiến hóa đó. Cũng giống như nghiên cứu vi cấu trúc thường được sử dụng để dự đoán cách giao dịch sẽ ảnh hưởng đến động lực giá và thanh khoản, học máy có thể cải thiện các dự đoán đó khi phải xử lý dữ liệu phức tạp và các ràng buộc tính toán. Bằng cách sử dụng thuật toán học máy *rừng ngẫu nhiên (random forest)*, chúng tôi khảo sát mức độ hiệu quả của một số thước đo vi cấu trúc thực nghiệm tiêu chuẩn (trong ngôn ngữ học máy được gọi là "đặc trưng" hay *features*) trong việc dự báo các biến số mà nhà đầu tư quan tâm. Trọng tâm của chúng tôi là một tập hợp các biến thường được sử dụng trong tạo lập thị trường điện tử, các chiến lược phòng ngừa rủi ro động, và ước lượng biến động. Mục đích của chúng tôi không phải là cung cấp một khảo sát đầy đủ về khả năng dự báo của dữ liệu thị trường, mà là minh họa cách học máy có thể mang lại những góc nhìn mới cho nghiên cứu vi cấu trúc bằng cách cho thấy những đặc trưng nào thực sự có hiệu quả trong dự báo ngoài mẫu (*out-of-sample*). Qua đó, chúng tôi cũng đưa ra bằng chứng rõ ràng về giá trị của một số biến vi cấu trúc hiện hữu trong việc lý giải động lực mới của hành vi thị trường.

Điều đáng nhấn mạnh là, các thuật toán học máy thường mang tính **phi tham số (non-parametric)** và không quy định trước một dạng hàm cụ thể. Việc phi tham số này không nên xem là hạn chế, bởi các thuật toán được thiết kế để thích ứng, nhằm khai thác các mẫu trong dữ liệu mà các mô hình tham số có thể không nhận ra. Kết quả là, các thuật toán học máy thường đem lại sức mạnh dự báo cao hơn và do đó là lựa chọn phù hợp hơn cho việc khảo sát khả năng dự báo dựa trên các biến vi cấu trúc.

Phân tích của chúng tôi dựa trên ba thế hệ mô hình vi cấu trúc thị trường

để cung cấp các thước đo cụ thể làm đầu vào cho khảo sát học máy. Các biến này bao gồm thước đo Roll, Roll Impact, thước đo biến động, hệ số λ của Kyle, thước đo Amihud, và VPIN (xác suất giao dịch có thông tin được đồng bộ theo khối lượng). Chúng tôi tập trung vào việc dự báo sáu kết quả quan trọng của động lực giá thị trường với tầm nhìn dự báo một tuần:

1. Dấu của thay đổi trong chênh lệch mua – bán (bid-ask spread);
2. Dấu của thay đổi trong biến động thực tế;
3. Dấu của thay đổi trong thống kê Jarque-Bera;
4. Dấu của thay đổi trong tương quan tuần tự của lợi suất thực tế;
5. Dấu của thay đổi trong độ lệch tuyệt đối (absolute skewness) của lợi suất;
6. Dấu của thay đổi trong độ nhọn (kurtosis) của lợi suất thực tế.

Chúng tôi đánh giá tầm quan trọng của từng đặc trưng bằng hai phương pháp: *Mean-Decreased Impurity (MDI)* – một thước đo trong mẫu, và *Mean-Decreased Accuracy (MDA)* – một thước đo ngoài mẫu. Dữ liệu sử dụng là 5 năm dữ liệu tick

Nghiên cứu của chúng tôi đưa ra một số kết quả. Như kỳ vọng, chúng tôi nhận thấy rằng các thước đo vi cấu trúc khác nhau cho thấy mức độ quan trọng khác nhau giữa trong mẫu và ngoài mẫu, cho thấy rằng những biến có thể có sức mạnh giải thích trong mẫu không nhất thiết có sức mạnh dự báo ngoài mẫu. Phù hợp với các nghiên cứu trước, tất cả các thước đo đều thể hiện sức mạnh giải thích trong mẫu. Trong sáu biến được dự báo, thước đo Amihud, VIX và VPIN có hiệu suất tốt nhất trong mẫu, trong khi VPIN thể hiện hiệu suất tốt nhất ngoài mẫu.

Ví dụ, trong dự báo dấu của thay đổi chênh lệch mua – bán, kết quả trong mẫu cho thấy Amihud và VPIN luôn có tầm quan trọng cao nhất ở mọi kích thước cửa sổ, trong khi ngoài mẫu thì VPIN chiếm ưu thế. Thực tế, kết quả dự báo ngoài mẫu cho thấy VPIN là đặc trưng quan trọng nhất đối với năm biến, trong khi thước đo Roll chiếm ưu thế ở biến thứ sáu (dự báo dấu thay đổi của tương quan tuần tự). Chúng tôi diễn giải kết quả này rằng những thước đo đơn giản được thiết kế để phản ánh ma sát thị trường vẫn còn hiệu quả trong các thị trường phức tạp hiện đại vốn bị chi phối bởi giao dịch dựa trên máy.

Những kết quả này không chỉ cho thấy tầm quan trọng của các biến vi cấu trúc cụ thể, mà còn cho thấy khả năng dự báo thành công động lực thị trường trong tương lai. Như chúng tôi sẽ thảo luận, những dự báo như vậy

có khả năng ứng dụng rộng rãi trong các lĩnh vực như quản lý rủi ro, chiến lược giao dịch động và tạo lập thị trường điện tử.

Bài báo của chúng tôi đóng góp vào một dòng nghiên cứu ngày càng phát triển về tác động của học máy và dữ liệu lớn trong nghiên cứu kinh tế. Varian (2014), Abadie và Kasy (2017), và Mullainathan và Spiess (2017) đưa ra các thảo luận xuất sắc về cách học máy có thể được áp dụng trong việc phân tích các vấn đề kinh tế liên quan đến dữ liệu lớn, trong khi các ứng dụng gần đây của những kỹ thuật này có thể tìm thấy trong Bajari et al. (2015) và Cavallo và Rigobon (2016).

Trong lĩnh vực tài chính, Chinco, Clarke-Joseph và Ye (2018) áp dụng kỹ thuật LASSO để dự báo lợi suất cổ phiếu trước 1 phút; Rossi (2018) sử dụng cây hồi quy tăng cường để dự báo lợi suất và biến động cổ phiếu; Krauss, Do và Huck (2017) sử dụng học máy để kinh doanh chênh lệch thống kê trên S&P 500; và López de Prado (2018) cung cấp phân tích mở rộng về các kỹ thuật và ứng dụng học máy tài chính. Gu, Kelly và Xiu (2018) áp dụng nhiều thuật toán hồi quy học máy trong định giá tài sản và phát hiện rằng các phương pháp hồi quy phi tuyến có thể tạo ra hệ số R-bình phương tốt hơn so với các mô hình kinh tế lượng.

Nghiên cứu của chúng tôi đóng góp vào dòng tài liệu này bằng cách chỉ ra rằng các kỹ thuật học máy có giám sát, khi kết hợp với các thước đo gợi ý từ lý thuyết vi cấu trúc, có thể giúp xác định những biến thị trường quan trọng mà không cần giả định dạng hàm cụ thể. Chúng tôi tin rằng việc học máy tách biệt việc tìm kiếm biến quan trọng khỏi việc xác định đặc tả mô hình sẽ đóng vai trò then chốt cho sự phát triển của nghiên cứu vi cấu trúc trong kỷ nguyên máy móc.

Bài báo này được tổ chức như sau: Phần tiếp theo trình bày các biến mà chúng tôi quan tâm dự báo và các biến vi cấu trúc được dùng làm đầu vào. Phần 3 giới thiệu ngắn gọn về phương pháp phân loại rừng ngẫu nhiên và các thước đo tầm quan trọng đặc trưng. Phần 4 trình bày dữ liệu, cách chúng tôi biến đổi dữ liệu thành các đơn vị phân tích gọi là "bars", và định nghĩa các biến vi cấu trúc. Phần 5 đưa ra kết quả thực nghiệm và kiểm tra độ vững chắc đối với các cửa sổ quan sát khác nhau, các cấu hình siêu tham số khác nhau, và các loại "bar" khác nhau. Phần 6 kết luận bằng cách thảo luận tác động của kết quả đối với chiến lược giao dịch, vai trò giải thích và dự báo của các biến vi cấu trúc, và đề xuất một chương trình nghiên cứu vi cấu trúc trong kỷ nguyên máy móc.

2 Các biến vi cấu trúc và biến động thị trường

Các mô hình vi cấu trúc cung cấp những biến gián tiếp đo lường các hệ quả quan sát được của ma sát thị trường. Trong chừng mực các thước đo này thành công, chúng nên có khả năng dự báo các giá trị hoặc biến động tương lai của các chỉ số thị trường như chênh lệch mua – bán (bid-ask spread), biến động, và các biến khác liên quan đến hình dạng phân phối lợi suất.

Một số mô hình (mà chúng tôi gọi là "thế hệ thứ nhất") sử dụng dữ liệu giá cho nhiệm vụ này. Ví dụ bao gồm thước đo Roll (1984), thước đo này sử dụng chuỗi giá để ước lượng chênh lệch mua – bán hiệu quả; ước lượng biến động của Beckers (1983) dựa trên giá cao – thấp; và bộ ước lượng chênh lệch mua – bán của Corwin và Schultz (2011).

Các mô hình thế hệ thứ hai tập trung vào dữ liệu giá và khối lượng, tạo ra các thước đo như λ của Kyle (1985), thước đo Amihud (2002), và λ của Hasbrouck (2009).

Các mô hình thế hệ thứ ba sử dụng dữ liệu giao dịch, truyền cảm hứng cho các thước đo như PIN (xác suất giao dịch có thông tin, Easley et al. 1996) và VPIN (xác suất giao dịch có thông tin được đồng bộ theo khối lượng, Easley et al. 2011).

Trong phân tích của chúng tôi, chúng tôi đánh giá sức mạnh dự báo của các thước đo đại diện cho cả ba thế hệ mô hình vi cấu trúc này.

Việc có thể dự báo sự phát triển tương lai của quá trình giá và thanh khoản rõ ràng là quan trọng, nhưng ít rõ ràng hơn là mức độ các thước đo vi cấu trúc tiêu chuẩn này còn hiệu quả trong các thị trường hiện nay. Các mô hình tạo ra những thước đo này khá đơn giản và được thiết kế vào thời điểm mà thị trường kém phức tạp hơn hiện tại. Những mô hình đó không đưa ra nhiều hướng dẫn về dạng hàm mô tả mối quan hệ giữa bất kỳ thước đo nào với động lực giá hay thanh khoản. Do đó, việc áp đặt một dạng hàm cụ thể cho mối quan hệ này, dù linh hoạt đến đâu, và áp dụng các kỹ thuật kinh tế lượng tiêu chuẩn để ước lượng, có thể làm che khuất mối quan hệ thực sự.

Mối quan tâm của chúng tôi là đánh giá khả năng dự báo bằng cách sử dụng nhiều biến vi cấu trúc khác nhau. Chúng tôi bắt đầu với dữ liệu về các biến vi cấu trúc (như tính thanh khoản kém, λ của Kyle hoặc VPIN) và dữ liệu về các thước đo thị trường (như chênh lệch mua – bán, biến động...) mà chúng tôi muốn dự báo.

Tuy nhiên, khác với cách tiếp cận tiêu chuẩn trong kinh tế lượng, chúng tôi không cố gắng quy định trước một quá trình sinh dữ liệu cơ bản, và do đó không ước lượng tham số của một mô hình liên hệ các thước đo vi cấu trúc với các thước đo thị trường. Mối quan tâm chính của chúng tôi là hiểu xem biến vi cấu trúc nào hữu ích cho dự báo và biến nào thì không. Chúng tôi

giữ thái độ trung lập về cơ chế liên hệ các biến trong tập dữ liệu với nhau, bởi việc cố gắng đặc tả một cơ chế, dù cấu trúc hay không gian xác suất cơ bản phức tạp đến đâu, là không cần thiết đối với mục đích khám phá dữ liệu của chúng tôi.

Do đó, chúng tôi sử dụng học máy để khảo sát hiệu quả của một tập hợp các thước đo vi cấu trúc trong việc dự báo một tập hợp các biến được quan tâm rộng rãi trên thị trường. Chúng tôi sẽ thảo luận chi tiết trong Mục 3 về cách thuật toán rừng ngẫu nhiên mà chúng tôi sử dụng hoạt động, nhưng điều quan trọng cần nhấn mạnh là chúng tôi sử dụng thuật toán này để *dự báo dấu của các thay đổi trong biến*, chứ không phải để cung cấp các dự báo điểm (point predictions) thực sự về giá trị của các biến đó.

Mặc dù điều này có thể có vẻ ít quan trọng, bên dưới chúng tôi giải thích vì sao không phải như vậy và thảo luận cách mà, đối với các biến ứng cử viên của chúng tôi, các dự báo như thế có thể được sử dụng trong thực tiễn.

1. **Dấu của thay đổi trong chênh lệch mua–bán (bid–ask spread).** Khi chúng ta kỳ vọng chênh lệch mua–bán sẽ mở rộng, một thuật toán thực thi có thể sử dụng kỳ vọng đó để *tăng* tỷ lệ tham gia khối lượng, qua đó tăng phần lệnh được khớp *trước* khi chi phí giao dịch tăng được hiện thực hóa. Ngược lại, khi chúng ta kỳ vọng chênh lệch mua–bán sẽ thu hẹp, một thuật toán thực thi có thể sử dụng kỳ vọng đó để *giảm* tỷ lệ tham gia khối lượng và như vậy khớp được phần lệnh lớn hơn *sau* khi mức giảm chi phí giao dịch diễn ra. Mức độ thay đổi của tỷ lệ tham gia khối lượng sẽ là một hàm của mức độ tin cậy vào độ chính xác của dự báo.

2. **Dấu của thay đổi trong biến động thực tế (realized volatility).** Khi chúng ta kỳ vọng biến động thực tế sẽ tăng, một thuật toán thực thi có thể sử dụng kỳ vọng đó để *tăng* tỷ lệ tham gia khối lượng, nhằm giảm bất định của giá khớp trung bình (rủi ro thị trường). Không nhất thiết rằng chúng ta sẽ muốn *giảm* tỷ lệ tham gia nếu kỳ vọng biến động giảm, bởi đến khi biến động đã giảm thì giá có thể đã trôi xa khỏi mục tiêu của chúng ta. Nói chung, chúng ta muốn *tăng* tỷ lệ tham gia khi dự báo biến động tăng, và *giảm* tỷ lệ tham gia sau khi mức giảm biến động đã thực sự xảy ra.

3. **Dấu của thay đổi trong thống kê Jarque–Bera.** Thống kê Jarque–Bera kiểm định giả thuyết gốc rằng các quan sát được rút ra từ một phân phối Chuẩn. Điều này có liên quan đến mục đích quản trị rủi ro, vì nhiều mô hình rủi ro giả định chuẩn tính của lợi suất. Xác suất cao hơn về lợi suất không-Chuẩn sẽ làm giảm mức độ tin cậy của chúng ta vào các mô hình đó. Ví dụ, một nhà quản trị rủi

ro có thể muốn *giảm* mức ý nghĩa (tỉ lệ dương tính giả, xác suất sai lầm loại I) của các mô hình Gauss khi kỳ vọng lợi suất không-Chuẩn.

4. **Dấu của thay đổi trong độ nhọn (kurtosis) / Dấu của thay đổi trong độ lệch tuyệt đối (absolute skewness) của lợi suất.**
   Thống kê Jarque–Bera sử dụng skewness và kurtosis để kiểm định chuẩn tính của các quan sát. Kiểm định này hàm ý một sự *đánh đổi* giữa skewness và kurtosis, theo nghĩa là kiểm định có thể không bác bỏ giả thuyết gốc về chuẩn tính khi một mức tăng trong skewness được bù trừ bởi một mức giảm trong kurtosis. Tuy nhiên, việc bù trừ skewness bằng kurtosis *không* phải là vô nghĩa về mặt kinh tế. Do skewness là một moment bậc lẻ, nó làm biến dạng phân phối Chuẩn bằng cách dồn xác suất về một phía. Một lý do khả dĩ cho biến dạng này là sự hiện diện của các nhà giao dịch có thông tin, những người đẩy giá nhằm lấp đầy lệnh trước khi một tin tức được biết rộng rãi. Trái lại, do kurtosis là một moment bậc chẵn, nó làm biến dạng phân phối Chuẩn bằng cách dồn xác suất *đối xứng* về phía các sự kiện cực đoan. Một lời giải thích khả dĩ cho biến dạng này là sự *suy giảm thanh khoản*, khi các nhà tạo lập thị trường giảm kích thước báo giá để đón đầu việc công bố tin tức, từ đó làm tăng khả năng xảy ra các kết cục cực đoan ở cả hai phía. Từ góc độ thực thi và quản lý danh mục, việc phân biệt giữa hai nguyên nhân của sự không-Chuẩn này là quan trọng, và cần được dự báo *riêng biệt*.

5. **Dấu của thay đổi trong tương quan tuần tự của lợi suất thực tế.**
   Một giả định phổ biến khác của các mô hình rủi ro (ví dụ trong cách tiếp cận VaR) là lợi suất không có tương quan tuần tự. Khi lợi suất có tương quan tuần tự, các xu hướng xuất hiện với tần suất cao hơn so với kỳ vọng thông thường. Điều này dẫn tới khả năng xảy ra mức sụt giảm (drawdown) lớn hơn và thời gian "ngụp lặn" kéo dài hơn. Tương tự trường hợp không-Chuẩn, xác suất cao hơn về lợi suất có tương quan tuần tự làm giảm mức độ tin cậy của chúng ta vào các mô hình giả định cấu trúc phi tương quan. Do đó, sẽ là hợp lý nếu giảm mức ý nghĩa của loại mô hình rủi ro này khi kỳ vọng lợi suất có tương quan tuần tự.

Ghi nhãn dự báo và các công thức liên quan. Trong phân loại, chúng tôi quan tâm đến việc dự báo *dấu* của thay đổi ở một số biến quan trọng. Các nhãn này là nhị phân, nhận giá trị $+1$ nếu thay đổi dương và -1 nếu thay đổi âm, phản ánh việc chúng tôi dự báo *dấu* của thay đổi trong biến

liên quan.

- *Dấu của thay đổi chênh lệch mua–bán:* chênh lệch S_τ được ước lượng bằng bộ ước lượng Corwin–Schultz,

$$S_\tau \;=\; \frac{2\big(e^{\alpha_\tau}-1\big)}{1+e^{\alpha_\tau}}, \qquad \alpha_\tau \;=\; \sqrt{2\,\beta_\tau} \;-\; \sqrt{\frac{\beta_\tau}{3-2\sqrt{2}}} \;-\; \sqrt{\frac{\gamma_\tau}{3-2\sqrt{2}}},$$

$$\beta_\tau \;=\; \mathbb{E}\left[\sum_{j=0}^{21} \log\!\left(\frac{H_{\tau-j}}{L_{\tau-j}}\right)\right], \qquad \gamma_\tau \;=\; \left[\log\!\left(\frac{H_{\tau-1,\tau}}{L_{\tau-1,\tau}}\right)\right]^2,$$

  trong đó $H_{\tau-j}, L_{\tau-j}$ là giá cao/thấp tại $\tau-j$, và $H_{\tau-1,\tau}, L_{\tau-1,\tau}$ là giá cao/thấp trong 2 bar $(\tau-1,\tau)$. Với chân trời dự báo h, nhãn là

$$\operatorname{sign}\big[S_{\tau+h}-S_\tau\big],$$

  tức chúng tôi dự báo liệu chênh lệch ước lượng sẽ mở rộng hay thu hẹp.

- *Dấu của thay đổi biến động thực tế:*

$$\operatorname{sign}\big[\sigma_{\tau+h}-\sigma_\tau\big],$$

  trong đó σ_τ là biến động thực tế của lợi suất 1-bar trên cửa sổ hồi cứu kích thước W.

- *Dấu của thay đổi thống kê Jarque–Bera của lợi suất thực tế:*

$$\operatorname{sign}\big[JB[r_{\tau+h}]-JB[r_\tau]\big], \qquad JB[r_\tau] \;=\; \frac{W}{6}\left(\mathrm{Skew}_\tau^2 \;+\; \frac{1}{4}\,(\mathrm{Kurt}_\tau-3)^2\right),$$

  trong đó Skew_τ là độ lệch và Kurt_τ là độ nhọn của lợi suất trên cửa sổ W.

- *Dấu của thay đổi trong tương quan tuần tự của lợi suất thực tế:*

$$AR_\tau \;=\; \operatorname{corr}(r_\tau, r_{\tau-1}), \qquad \operatorname{sign}\big[AR_{\tau+h}-AR_\tau\big],$$

  trong đó tương quan được tính trên chuỗi lợi suất của W bar gần nhất.

- *Dấu của thay đổi trong độ lệch tuyệt đối và trong độ nhọn của lợi suất thực tế:*

$$\operatorname{sign}\big[\mathrm{Skew}_{\tau+h}-\mathrm{Skew}_\tau\big], \qquad \operatorname{sign}\big[\mathrm{Kurt}_{\tau+h}-\mathrm{Kurt}_\tau\big].$$

Ghi chú. Trong phân tích hiện tại, chúng tôi cố định chân trời dự báo h bằng 250 bar (xấp xỉ một tuần giao dịch). Chúng tôi cũng xem xét $h = 50$ bar và thấy kết quả tương tự. Ngoài ra, ở đây chúng tôi chỉ xét thay đổi dương và âm; trong bối cảnh đầu tư/giao dịch, người ta có thể (và có lẽ nên) xét phân hoạch tinh hơn, tối thiểu là thêm một hạng mục "thay đổi nhỏ".

3 Thuật toán phân loại rừng ngẫu nhiên và thước đo tầm quan trọng đặc trưng

Trong phần này, chúng tôi giới thiệu thuật toán phân loại rừng ngẫu nhiên (random forest classification) và cách sử dụng nó để đánh giá sức mạnh dự báo của một tập biến giải thích. Trong học máy, phân loại là việc sử dụng các biến giải thích để dự đoán biến mục tiêu rời rạc/định tính. Nó tương tự như hồi quy ở chỗ cả hai đều được ước lượng bằng cách tối thiểu hóa một hàm lỗi, nhưng biến mục tiêu trong phân loại là rời rạc (ví dụ: "có" hoặc "không"), nên các hàm lỗi quen thuộc trong hồi quy (như sai số bình phương trung bình) không thích hợp. Thay vào đó, các thước đo lỗi như entropy chéo (cross-entropy) hoặc thông tin thu được (information gain) thường được sử dụng.

Thuật toán rừng ngẫu nhiên

Trong các phương pháp phân loại học máy, rừng ngẫu nhiên là một trong những thuật toán bền vững và được sử dụng rộng rãi nhất. Nó bao gồm một số lượng lớn các bộ phân loại con gọi là *cây quyết định* (decision trees), và sử dụng trung bình kết quả phân loại của các cây này để đưa ra dự đoán cuối cùng. Khi số lượng cây tăng và ít tương quan với nhau, phương sai của lỗi dự báo giảm, từ đó giảm khả năng quá khớp dữ liệu (overfitting).

Với mỗi hợp đồng tương lai, ta xây dựng một tập dữ liệu $\{(x_t, y_t)\}_{t=1}^{T}$, trong đó x_t là vector đặc trưng và y_t là nhãn (label). Cây quyết định được xây dựng bằng cách chia tập mẫu thành hai tập con, sau đó tiếp tục chia các tập con này, lặp lại quá trình.

Hàm thông tin thu được (Information Gain)

Để tạo ra phép chia, ta tính toán cho mỗi đặc trưng mức thông tin thu được (information gain) nếu chia mẫu theo đặc trưng đó. Với tập mẫu S tại nút n, khi chia thành hai tập con L và R, thông tin thu được được xác định như sau:

$$IG(S, n) = I(S) - \frac{N_L}{N_S} I(L) - \frac{N_R}{N_S} I(R),$$

trong đó $I(S)$ là độ thuần khiết (purity) của tập S. Chúng tôi sử dụng chỉ số Gini:

$$I(S) = \sum_i p_i(1 - p_i),$$

với p_i là tỉ lệ nhãn thuộc lớp i trong tập S. N_S, N_L, N_R lần lượt là số điểm dữ liệu trong tập gốc, tập con trái, và tập con phải.

Phép chia tối ưu là phép chia cho giá trị IG lớn nhất. Quá trình này được lặp lại cho đến khi đạt tiêu chí dừng định trước (ví dụ: số điểm dữ liệu trong một nút nhỏ hơn một ngưỡng tối thiểu, hoặc không còn phép chia nào tạo ra IG dương).

Khi xây dựng một cây quyết định, có thể xảy ra hiện tượng "quá khớp" (overfitting) nếu cây phát triển quá sâu và bắt đầu ghi nhớ nhiễu trong dữ liệu huấn luyện. Một ưu điểm then chốt của rừng ngẫu nhiên là thay vì chỉ dựa vào một cây, một tập hợp lớn các cây được tạo ra, mỗi cây được huấn luyện trên một mẫu bootstrap khác nhau của dữ liệu, và tại mỗi nút chỉ xét một tập con ngẫu nhiên của các đặc trưng. Sự ngẫu nhiên kép này (ở cả quan sát và đặc trưng) giúp giảm tương quan giữa các cây và làm cho kết quả bỏ phiếu đa số của toàn rừng ổn định hơn, ít nhạy cảm với nhiễu hơn.

Trong quá trình dự báo, mỗi cây trong rừng đưa ra một nhãn cho quan sát mới, và nhãn cuối cùng được xác định bằng cách bỏ phiếu đa số từ tất cả các cây. Nhờ vậy, rừng ngẫu nhiên có xu hướng đạt được độ chính xác cao ngoài mẫu và giảm đáng kể nguy cơ quá khớp so với một cây đơn lẻ.

Các thước đo tầm quan trọng đặc trưng

Trong phân tích, chúng tôi sử dụng hai thước đo chuẩn: **Mean Decreased Impurity (MDI)** và **Mean Decreased Accuracy (MDA)**.

(1) Mean Decreased Impurity (MDI). MDI đánh giá tầm quan trọng của đặc trưng dựa trên tổng mức độ cải thiện thuần khiết trong tất cả các cây:

$$MDI(i) \;=\; \frac{1}{100} \sum_{N} \sum_{n \in N : v(s_n) = i} p(t)\, IG(s_n, n),$$

trong đó $v(s_n)$ là đặc trưng được dùng để chia tại nút s_n, với s_0 là tập dữ liệu ban đầu.

(2) Mean Decreased Accuracy (MDA). Cần lưu ý rằng MDI là một phương pháp *trong mẫu* (in-sample), bởi nó được suy ra từ cùng tập thông tin được sử dụng để huấn luyện các cây. Điều này làm cho MDI tương tự như một giá trị p trong phân tích hồi quy. Ngược lại, MDA đánh giá tầm quan trọng của đặc trưng *ngoài mẫu* (out-of-sample), và khác với MDI, nó có thể được sử dụng với *bất kỳ bộ phân loại nào*.

Thủ tục MDA tính toán tầm quan trọng của đặc trưng như sau:

1. Chia tập dữ liệu thành hai phần không chồng lấn: tập huấn luyện và tập kiểm định.

2. Huấn luyện một bộ phân loại trên tập huấn luyện sử dụng tất cả các đặc trưng.

3. Thực hiện dự báo trên tập kiểm định và ghi nhận một thước đo hiệu suất (ví dụ: độ chính xác), gọi là p_0.

4. Với từng đặc trưng i, xáo trộn ngẫu nhiên giá trị của nó trong tập kiểm định, trong khi giữ nguyên tất cả các đặc trưng khác.

5. Thực hiện lại dự báo trên tập kiểm định đã bị xáo trộn và ghi nhận thước đo hiệu suất, gọi là p_i.

6. Độ quan trọng của đặc trưng i được tính là:

$$MDA(i) \;=\; \frac{p_0 - p_i}{p_0}.$$

Do đó, tầm quan trọng của đặc trưng theo MDA được xác định bởi mức độ mà dự báo ngoài mẫu trở nên tệ hơn do việc xáo trộn giá trị của một đặc trưng cụ thể. Sự suy giảm hiệu suất càng lớn thì đặc trưng đó càng quan trọng.

Độ chính xác dự báo. Cuối cùng, chúng tôi chuyển sang vấn đề về độ chính xác dự báo. Chúng tôi định nghĩa **độ chính xác** là số dự báo đúng chia cho tổng số dự báo được tạo ra từ một tập dữ liệu nhất định được chia thành tập huấn luyện và tập kiểm định.

Nếu chúng ta chỉ áp dụng ý tưởng này một lần cho tập dữ liệu hợp đồng tương lai của mình, thì có thể dẫn đến kết quả thiên lệch do tính chất phụ thuộc theo thời gian của dữ liệu tài chính. Để có được một ước lượng vững chắc hơn về độ chính xác ngoài mẫu, chúng tôi sử dụng thủ tục **Purged K-Fold Cross-Validation**.

Trong thủ tục này, tập dữ liệu được chia thành K phần liên tiếp theo thời gian. Ở mỗi vòng lặp, một phần được giữ lại làm tập kiểm định, và các quan sát nằm trong một khoảng thời gian nhất định xung quanh phần đó bị loại bỏ ("purged") để ngăn chặn rò rỉ thông tin (look-ahead bias). Mô hình được huấn luyện trên phần dữ liệu còn lại và sau đó đánh giá trên tập kiểm định. Độ chính xác ngoài mẫu cuối cùng được tính bằng trung bình kết quả từ tất cả các vòng lặp.

Trong nghiên cứu này, chúng tôi chọn $K = 10$ (tức là 10-fold), với khoảng thời gian purge là một tuần giao dịch. Điều này cho phép tận dụng dữ liệu hiệu quả trong khi vẫn đảm bảo tính độc lập ngoài mẫu, và kết quả thu được là một thước đo đáng tin cậy về sức mạnh dự báo thực sự của các biến vi cấu trúc.

4 Dữ liệu

Trong phần này, chúng tôi chuyển sang dữ liệu và các định nghĩa cụ thể về các *nhãn* và *đặc trưng* mà chúng tôi sử dụng trong phân tích. Chúng tôi cũng đề cập đến một loạt các vấn đề triển khai. Phân tích của chúng tôi sử dụng *dollar-volume bars*, vì vậy chúng tôi trình bày cách chúng tôi dùng dữ liệu tick để tạo các bars cho các hợp đồng khác nhau trong mẫu. Vì chúng tôi dùng dữ liệu hợp đồng tương lai, dữ liệu của chúng tôi phải "cuộn" (roll) qua các kỳ đáo hạn của hợp đồng để tạo ra một chuỗi giá liên tục. Chúng tôi mô tả cách thực hiện việc chuyển tiếp đó bằng một quy trình tương tự như tạo ra một ETF trên hợp đồng. Cuối cùng, chúng tôi thảo luận các vấn đề đo lường liên quan đến việc xem các biến vi cấu trúc trong các bars theo khối lượng so với các đơn vị dựa trên thời gian. :contentReference[oaicite:0]index=0

Phân tích của chúng tôi được thực hiện trên 87 hợp đồng tương lai có tính thanh khoản cao nhất được giao dịch trên toàn cầu, với các chi tiết cụ thể của từng hợp đồng được nêu trong Bảng A.1 ở Phụ lục. Chúng tôi sử dụng các hợp đồng tương lai này vì hai lý do. Thứ nhất, chúng tôi có thể xem xét toàn bộ vũ trụ các hợp đồng tương lai đang hoạt động, nên không có vấn đề chọn mẫu từ một tập hợp lớn hơn các tài sản tài chính. Thứ hai, chúng tôi có dữ liệu giao dịch đầy đủ về giao dịch của các tài sản này. Giai đoạn mẫu của chúng tôi bắt đầu vào ngày 2 tháng 7 năm 2012 và kết thúc vào ngày 2 tháng 10 năm 2017. Dữ liệu cấp tick sẵn có cho phần lớn các hợp đồng này trong một giai đoạn dài hơn, nhưng chúng tôi quan tâm đến VIX như một đặc trưng và hợp đồng tương lai trên VIX (mã UX1) chỉ bắt đầu được giao dịch vào tháng 7 năm 2012. Chúng tôi lưu ý rằng hai hợp đồng hàng hoá trong mẫu (IK1 và BTS1) có giai đoạn mẫu ngắn hơn, bắt đầu vào tháng 10 năm 2015.

A. Tạo các dollar-volume bars

Chúng tôi thu thập dữ liệu giao dịch cấp tick cho mỗi hợp đồng tương lai và gộp dữ liệu này thành các khoảng, hay *bars*, dựa trên khối lượng tính theo đô-la. Việc gộp dữ liệu thành các bars được xác định theo thời gian hoặc theo gia số khối lượng là thông lệ chuẩn trong ngành và trong nghiên cứu học thuật (xem, chẳng hạn, Engle và Lange [2001]; Easley et al. [2012]; Chakrabarty et al. [2012]; Easley et al. [2016]; Low et al. [2018]). Barardehi, Bernhardt và Davies [2019] cũng đề xuất một cách tiếp cận theo *trade time* trong đo lường thanh khoản và cho thấy nó hoạt động tốt hơn so với cách tiếp cận theo *clock time*.

Easley và O'Hara [1992] chỉ ra rằng thời gian giữa các giao dịch nên tương quan với sự tồn tại của thông tin mới, cung cấp cơ sở cho việc chúng tôi xem xét *trade time* (khối lượng) thay vì *clock time*. Sự đi tới của thông tin tạo ra các mẫu hình trong khối lượng, về cơ bản tương tự như các tính mùa vụ trong ngày.[1]

Bằng cách lấy mẫu bất cứ khi nào thị trường trao đổi một khối lượng cố định, chúng tôi tìm cách mô phỏng việc thị trường tiếp nhận các tin tức có mức độ liên quan so sánh được. Chúng tôi sử dụng *dollar-volume* để cho phép khả năng so sánh giữa 87 hợp đồng trong mẫu. Ngoài ra, López de Prado [2018] đưa ra bằng chứng rằng tần suất lấy mẫu của các *dollar-volume bars* ổn định hơn tần suất lấy mẫu của *time bars* hoặc *volume bars*. Một lý do cho sự ổn định này là *dollar-volume bars* tính đến các dao động giá, do đó chuẩn hoá giá trị giao dịch theo đô-la giữa các khoảng thời gian khác nhau.

Bar thứ τ được hình thành tại tick t khi

$$\sum_{j=t_0^\tau}^{t} p_j V_j \;\geq\; L,$$

trong đó t_0^τ là chỉ số của tick đầu tiên trong bar thứ τ, p_j là giá giao dịch tại tick j, V_j là khối lượng giao dịch tại tick j, và L là một ngưỡng được ấn định trước sao cho (trong năm 2016) mỗi ngày có xấp xỉ 50 bars.[2]

Lưu ý rằng do khối lượng giao dịch trung bình hàng ngày khác nhau giữa các hợp đồng, nên *dollar-volume* trong mỗi bar sẽ khác nhau giữa các hợp đồng tương lai cụ thể, nhưng số lượng bar trung bình mỗi ngày thì không (trong năm 2016). Đối với từng hợp đồng riêng lẻ, vào một ngày sôi động, các bars sẽ được lấp đầy nhanh hơn và có thể có nhiều hơn 50 bars trong

[1] Hợp đồng tương lai thường giao dịch trong một ngày 23.75 giờ và các mẫu hình khối lượng rất rõ rệt.

[2] Chúng tôi chọn năm 2016 vì đó là năm đầy đủ cuối cùng trước khi kết thúc mẫu.

một ngày; vào một ngày kém sôi động, các bars sẽ được lấp đầy chậm hơn và có thể có ít hơn 50 bars trong một ngày.

Chúng tôi tính toán mỗi biến vi cấu trúc trong phân tích tại từng bar τ bằng cách áp dụng một "cửa sổ nhìn lại" trượt có kích thước W. Ví dụ, tại bar τ, chúng tôi sử dụng các bars thuộc tập $\{\tau-W+1,\ \tau-W+2,\ \ldots,\ \tau-1,\ \tau\}$ để tính các biến vi cấu trúc và các nhãn. Trong phân tích của mình, chúng tôi xem xét các cửa sổ nhìn lại từ 25 bars đến 2000 bars.

B. Cuộn hợp đồng (Rolling contracts)

Vì các hợp đồng tương lai có ngày đáo hạn, chúng ta cần "cuộn" các hợp đồng (tức là bán hợp đồng sắp đáo hạn và tham gia hợp đồng mới) để tạo ra một chuỗi giá như thể đó là một công cụ liên tục. Để làm như vậy, chúng tôi biến đổi giá của hợp đồng tương lai thành giá trị của một ETF bám sát hoàn hảo hợp đồng tương lai với vốn ban đầu 1 đô la.[3] Để hiểu quy trình này, hãy xét ví dụ sau.

Giả sử chúng ta muốn nắm giữ vị thế mua ở hợp đồng gần nhất của E-mini S&P 500 futures (mã Bloomberg: ES1<Index>), bắt đầu từ 01/02/2015. Vào ngày 03/20/2015, hợp đồng tương lai tháng gần nhất sắp đáo hạn và chúng ta phải bán nó rồi mua hợp đồng tháng kế tiếp, do đó "cuộn sang hợp đồng kế". Trong quá trình cuộn này, không có thay đổi nào về giá trị khoản đầu tư của chúng ta ngoại trừ chi phí giao dịch rất nhỏ. Tuy nhiên, thường tồn tại chênh lệch về mức giá thô giữa hợp đồng tháng gần nhất và hợp đồng tháng kế. Nếu hợp đồng tháng gần nhất giao dịch ở mức \$2000 trong khi hợp đồng tháng kế ở mức \$2020, thì nếu ta đơn giản chuyển chuỗi thời gian giá từ tháng gần nhất sang tháng kế, lúc này sẽ có một khác biệt 1%. Thuật toán học máy sẽ nghĩ sai rằng có một cú nhảy giá đột ngột và coi đó là một dạng tín hiệu.

Để tránh vấn đề này, chúng tôi tạo ra một chuỗi thời gian mới mà chúng tôi gọi là *giá ETF* của chuỗi hợp đồng tương lai, phản ánh giá trị của 1 đô la được đầu tư vào hợp đồng tương lai giả định rằng có thể nắm giữ các phần lẻ của hợp đồng. Chuỗi này bắt đầu từ mức 1, và giá trị hiện tại của nó bằng với lợi nhuận lũy kế của khoản đầu tư (xem Bảng A.2 để biết ví dụ). Khi hợp đồng tương lai được cuộn, người ta bán hợp đồng cũ và đầu tư toàn bộ tiền vào hợp đồng mới. Trong sự kiện này, không có thay đổi nào đối với khoản đầu tư nếu giả định chi phí giao dịch bằng 0, vì vậy giá ETF không bị ảnh hưởng bởi sự thay đổi nhân tạo trong mức giá thô. Hình 1 đưa ra đồ thị cho lợi nhuận lũy kế và chuỗi giá ETF của chỉ số ES1.

Trong Phụ lục A.2, chúng tôi cung cấp các chi tiết tính toán cho quy

[3] Để thảo luận chi tiết về kỹ thuật này, xem López de Prado [2018].

trình này. Trong các phân tích tiếp theo, đối với mỗi hợp đồng tương lai, chúng tôi sử dụng giá dựa trên ETF và khối lượng tương ứng thay vì giá và khối lượng thô, trừ khi có nêu khác đi.

C. Đặc trưng và Nhãn (Features and Labels)

Như đã thảo luận ở phần trước, chúng tôi tập trung vào một số biến vi cấu trúc thị trường quen thuộc. Các đặc trưng này đều được xây dựng từ dữ liệu *bar* đã mô tả ở trên. Một vấn đề phát sinh trong quá trình xây dựng các thước đo vi cấu trúc này là ban đầu chúng không được định nghĩa dựa trên cùng khái niệm về khoảng thời gian hoặc *bar*, hoặc dựa trên các cửa sổ nhìn lại. Do đó, với mỗi thước đo chúng tôi phải điều chỉnh định nghĩa gốc sang bối cảnh của chúng tôi. Chúng tôi vẫn gọi các thước đo này bằng tên gốc của chúng, nhưng lưu ý rằng đó thực chất là phiên bản được "dịch" sang thiết lập của chúng tôi.

Cụ thể hơn, chúng tôi có:

- **Thước đo Roll**, được cho bởi

$$R_\tau \;=\; 2\sqrt{\big|\operatorname{cov}(\Delta P_\tau,\, \Delta P_{\tau-1})\big|},$$

  trong đó

$$\Delta P_\tau = [\,\Delta p_{\tau-W},\, \Delta p_{\tau-W+1},\, \ldots,\, \Delta p_\tau\,], \qquad \Delta P_{\tau-1} = [\,\Delta p_{\tau-W-1},\, \Delta p_{\tau-W},\, \ldots,\, \Delta p_{\tau-1}\,],$$

  và Δp_τ là thay đổi của giá đóng cửa giữa các *bar* $\tau-1$ và τ, còn W là kích thước cửa sổ nhìn lại.

- **Roll impact**, là thước đo Roll chia cho giá trị đô-la giao dịch trong một giai đoạn nhất định:

$$\widetilde{R}_\tau \;=\; \frac{2\sqrt{\big|\operatorname{cov}(\Delta P_\tau,\, \Delta P_{\tau-1})\big|}}{p_\tau\, V_\tau}.$$

  Mẫu số được đánh giá tại từng *bar*; do cách hình thành *bar*, mẫu số thay đổi rất ít giữa các *bar* liên tiếp.

- **λ của Kyle**, được cho bởi

$$\lambda_\tau \;=\; \frac{p_\tau - p_{\tau-W}}{\displaystyle\sum_{i=\tau-W}^{\tau} b_i\, V_i},$$

  trong đó $b_i = \operatorname{sign}(p_i - p_{i-1})$ (tính ở cấp *bar*) và W là kích thước cửa sổ nhìn lại.

- λ **của Amihud**, được cho bởi

$$\lambda_\tau^A \;=\; \frac{1}{W} \sum_{i=\tau-W+1}^{\tau} \frac{|r_i|}{p_i\, V_i},$$

trong đó r_i, p_i, V_i lần lượt là lợi suất, giá và khối lượng tại *bar* i, còn W là kích thước cửa sổ nhìn lại. Việc đo lường này khi sử dụng *dollar-volume bars* có liên hệ gần gũi với biến thể theo *trade time* của thước đo Amihud trong Barardehi, Bernhardt và Davies (2019).

- **VPIN (Volume-synchronized Probability of Informed Trading)**, được ước lượng như sau:

$$\mathrm{VPIN}_\tau \;=\; \frac{1}{W} \sum_{i=\tau-W+1}^{\tau} \frac{\left|\, V_i^S - V_i^B \,\right|}{V_i},$$

trong đó khối lượng được gán dấu theo phương pháp BVC:

$$V_i^B \;=\; V_i\, \Phi\!\left(\frac{\Delta p_i}{\sigma_{\Delta p_i}}\right), \qquad V_i^S \;=\; V_i - V_i^B,$$

và W là kích thước cửa sổ nhìn lại. Xem Easley et al. (2016) để biết thêm chi tiết.

- **Chỉ số VIX**. Chúng tôi sử dụng dữ liệu giao dịch cấp tick của hợp đồng tương lai VIX kỳ gần nhất (mã Bloomberg: UX1<Index>) để đại diện cho VIX. Với mỗi *bar*, chúng tôi lấy giá của tick UX1 gần nhất trước dấu thời gian của *bar* đó làm giá trị của VIX.

Nhãn (Labels). Đối với bài toán phân loại, chúng tôi quan tâm dự báo *dấu* của thay đổi ở một số biến quan trọng. Lưu ý rằng các nhãn là nhị phân, nhận giá trị dương $+1$ hoặc âm -1, phản ánh việc chúng tôi dự báo dấu của thay đổi. Cụ thể, chúng tôi gán nhãn theo:

- Bid-Ask Spread (Corwin-Schultz Estimator)

  Bid-ask spread được tính toán thông qua **Corwin-Schultz estimator**:

$$S_\tau = \frac{2\,(e^{\alpha_\tau} - 1)}{1 + e^{\alpha_\tau}}$$

  Trong đó:

$$\alpha_\tau = \frac{\sqrt{2\beta_\tau} - \sqrt{\beta_\tau}}{3 - 2\sqrt{2}} - \frac{\sqrt{\gamma_\tau}}{\sqrt{3 - 2\sqrt{2}}}$$

Và:

$$\beta_\tau = \mathbb{E}\left[\sum_{j=0}^{1}\left(\log\left(\frac{H_{\tau-j}}{L_{\tau-j}}\right)\right)^2\right]$$

Với $H_{\tau-j}$ và $L_{\tau-j}$ là giá cao và giá thấp tại thời điểm $\tau - j$, và xích ma chạy từ **0 đến 1**.

$$\gamma_\tau = \left(\log\left(\frac{H_{\tau-1,\tau}}{L_{\tau-1,\tau}}\right)\right)^2$$

Với $H_{\tau-1,\tau}$ và $L_{\tau-1,\tau}$ là giá cao và giá thấp trong 2 bars liên tiếp $(\tau - 1, \tau)$.

Để dự báo sự thay đổi của bid-ask spread, chúng ta tính:

$$\text{sign}[S_{\tau+h} - S_\tau]$$

Mục tiêu là dự báo liệu spread sẽ **widen** (mở rộng) hay **narrow** (co lại) trong khoảng thời gian h.

- **Dấu của thay đổi trong biến động thực tế (realized volatility),**

$$\text{sign}\big[\sigma_{\tau+h} - \sigma_\tau\big],$$

trong đó σ_τ là biến động thực tế của lợi suất 1-*bar* trên một cửa sổ nhìn lại kích thước W; ở đây ta dự báo biến động sẽ tăng hay giảm.

- **Dấu của thay đổi trong thống kê Jarque–Bera của lợi suất thực tế,**

$$\text{sign}\big[JB[r_{\tau+h}] - JB[r_\tau]\big], \qquad JB[r_\tau] \;=\; \frac{W}{6}\left(\text{Skew}_\tau^2 + \frac{1}{4}\big(\text{Kurt}_\tau - 3\big)^2\right),$$

trong đó Skew_τ là độ lệch và Kurt_τ là độ nhọn của lợi suất trên cửa sổ kích thước W; nhãn này có thể được xem là khái quát hoá bậc cao hơn của biến động thực tế.

- **Dấu của thay đổi trong tự tương quan bậc một (first-order autocorrelation) của lợi suất thực tế,**

$$\text{sign}\big[AR_{\tau+h} - AR_{\tau}\big], \qquad AR_{\tau} \;=\; \text{corr}(r_{\tau},\, r_{\tau-1}),$$

  được tính trên chuỗi lợi suất của W *bar* gần nhất.

- **Dấu của thay đổi trong độ lệch (skewness) và trong độ nhọn (kurtosis) của lợi suất thực tế,**

$$\text{sign}\big[\text{Skew}_{\tau+h} - \text{Skew}_{\tau}\big], \qquad \text{sign}\big[\text{Kurt}_{\tau+h} - \text{Kurt}_{\tau}\big].$$

Trong phân tích hiện tại, chúng tôi cố định chân trời dự báo h là 250 *bars* phía trước, tương ứng xấp xỉ một tuần giao dịch. Trong Mục 5.4, chúng tôi phân tích một chân trời dự báo là 50 *bars* và nhận thấy kết quả tương tự.

5 Kết quả và phân tích

Trong phần này, trước hết chúng tôi trình bày các tham số của phương pháp phân loại rừng ngẫu nhiên. Sau đó, chúng tôi trình bày các kết quả chính của bài báo, cụ thể là tầm quan trọng đặc trưng của các biến vi cấu trúc. Tiếp theo là một phân tích độ nhạy trong đó chúng tôi tinh chỉnh các tham số của rừng ngẫu nhiên, và các kiểm tra độ vững khác nhau, bao gồm so sánh giữa kết quả dựa trên *dollar-volume-bar* và *time-bar* cũng như so sánh kết quả học máy của chúng tôi với kết quả của hồi quy logistic. :contentReference[oaicite:0]index=0

Phân tích của chúng tôi sử dụng một gói phần mềm học máy mã nguồn mở tiêu chuẩn, Scikit-learn (xem Pedregosa et al. [2011]). Chúng tôi bắt đầu bằng cách chỉ rõ cấu hình (các siêu tham số, theo cách nói của học máy) của thuật toán rừng ngẫu nhiên. Trong phân tích của mình, chúng tôi chọn các giá trị mặc định cho các siêu tham số của rừng ngẫu nhiên[4]: :contentReference[oaicite:1]index=1

- số lượng cây (`n_estimators`) = 100,
- số đặc trưng tối đa mỗi lần chia (`max_features`) = $\text{int}(\sqrt{6}) = 2$,
- trọng số mẫu (`class_weight`) = nghịch đảo của tổng số mẫu trong lớp của mẫu (`balanced`).

[4]Ký hiệu tương ứng trong Scikit-learn được đặt trong dấu ngoặc.

Số lượng cây là một tham số kiểm soát có bao nhiêu cây quyết định trong rừng ngẫu nhiên. Số đặc trưng tối đa ở đây là căn bậc hai của tổng số đặc trưng, một lựa chọn thường dùng cho rừng ngẫu nhiên. Trọng số mẫu là trọng số gán cho mỗi mẫu trong lớp huấn luyện, và chúng tôi sử dụng cách tiếp cận cân bằng để giảm thiên lệch có thể đến từ sự mất cân đối nhãn. Chúng tôi báo cáo các kết quả từ một rừng ngẫu nhiên *không chuẩn hoá* (tức là các cây quyết định được phép phát triển không giới hạn). Trong Mục 5.4, chúng tôi chạy lại phân tích bằng một rừng ngẫu nhiên *được chuẩn hoá* để kiểm tra rằng các kết quả của chúng tôi ổn định và vững, đồng thời xua tan lo ngại rằng rừng ngẫu nhiên ban đầu bị *overfit*. :contentReference[oaicite:2]index=2

5.1 Kết quả MDI

Trước hết, chúng tôi xem xét tầm quan trọng đặc trưng sử dụng MDI. Nhắc lại, MDI là một phương pháp *trong mẫu* dựa trên sức mạnh giải thích của từng đặc trưng và cho ra các giá trị tầm quan trọng đã được chuẩn hoá (tất cả dương và tổng bằng một). Bảng 2, các Phần A đến E, báo cáo tầm quan trọng đặc trưng theo MDI cho từng trong số sáu biến được dự báo mà chúng tôi xem xét trong phân tích. Mỗi hàng tương ứng với một kích thước cửa sổ nhìn lại cụ thể, như được chỉ trong cột đầu tiên. Mỗi ô được trình bày dưới dạng "giá trị trung bình của điểm tầm quan trọng MDI" ± "độ lệch chuẩn của điểm tầm quan trọng MDI", trong đó trung bình và độ lệch chuẩn được tính trên toàn bộ 87 công cụ. Giá trị tầm quan trọng cao nhất được bôi đậm cho mỗi kích thước cửa sổ. :contentReference[oaicite:0]index=0

Để cung cấp một trực giác về cách các đặc trưng đóng góp vào giải thích trong mẫu của các đặc trưng, chúng tôi đưa ra trong Hình 2 một đồ thị phân tán của các thay đổi dự báo trong chênh lệch đối với chỉ số ES1 như là hàm của các thước đo VPIN và Roll. Rừng ngẫu nhiên gán một dự báo tăng hoặc giảm của chênh lệch với bất kỳ danh sách nào của tất cả các đặc trưng. Việc vẽ đồ thị sự gán này so với các véc-tơ đặc trưng tạo ra một đồ thị trong $\mathbb{R}^6$, mà chúng tôi chiếu xuống $\mathbb{R}^2$ trong Hình 2. Hình chữ L vuông góc trong hình đó chỉ ra một ngưỡng cắt cho các dự báo chênh lệch ở ngay dưới 0.002 đối với thước đo Roll và ngay trên 0.05 đối với VPIN. Rừng ngẫu nhiên dự báo chênh lệch giảm trong góc phần tư tây–bắc và tăng ở những nơi khác.[5] :contentReference[oaicite:1]index=1

Đối với ước lượng chênh lệch mua–bán, Phần A cho thấy thước đo Amihud có tầm quan trọng cao nhất, tiếp theo là chỉ số VPIN. Tầm quan trọng đặc trưng tăng theo kích thước cửa sổ đối với Amihud, Kyle và VPIN, nhưng

[5]Nhìn trong hình có vẻ nhiễu vì chúng tôi đang điều kiện hoá chỉ trên hai trong sáu đặc trưng. Nếu không, sẽ có những vùng ranh giới sắc nét hơn.

không phải đối với các thước đo Roll hoặc đối với VIX. Hiệu suất khác biệt (và thấp hơn) của VIX so với VPIN bác bỏ quan điểm cho rằng VPIN chỉ đơn thuần bắt được các hiệu ứng biến động. Thước đo Amihud cũng là quan trọng nhất đối với dự báo độ lệch tuyệt đối (Phần E). :contentReference[oaicite:2]index=2

Phần B đưa ra các kết quả tầm quan trọng đặc trưng cho dự báo biến động. Ở đây chúng tôi tìm thấy các kết quả pha trộn tuỳ theo kích thước cửa sổ. Đối với cả cửa sổ ngắn nhất (25) và dài nhất (lớn hơn hoặc bằng 1000) *bars*, VPIN chiếm ưu thế. Amihud là quan trọng nhất nếu được đo trên cửa sổ 250–500 *bars*, trong khi VIX chiếm ưu thế cho cửa sổ 50 *bars* (mặc dù VIX và VPIN cũng rất tương tự nhau đối với cửa sổ 25 *bars*). Tầm quan trọng đặc trưng đối với dự báo kiểm định Jarque–Bera trong Phần C cũng cho thấy các kết quả pha trộn. Nhìn chung, Amihud là quan trọng nhất, nhưng đối với một số cửa sổ thì VIX và VPIN lại chiếm ưu thế. Thước đo Amihud cũng đạt kết quả tốt khi sử dụng các cửa sổ dài hơn cho dự báo tương quan tuần tự (Phần D), trong khi VIX chiếm ưu thế đối với các cửa sổ ngắn hơn. Thú vị là các thước đo Roll, vốn có thể kỳ vọng sẽ làm tốt với dự báo thay đổi tương quan, lại không đạt kết quả tốt. Các kết quả cho dự báo độ nhọn một lần nữa thiên về Amihud đối với các cửa sổ dài, nhưng lại thiên về VIX và VPIN đối với các cửa sổ ngắn hơn. :contentReference[oaicite:3]index=3

Nhìn chung, dữ liệu gợi ý rằng nếu đo bằng hiệu suất trong mẫu thì thước đo Amihud làm tốt nhất, với VPIN và VIX cũng có tầm quan trọng đặc trưng mạnh. Hệ số λ của Kyle và các thước đo Roll không bao giờ là thước đo quan trọng nhất để dự báo bất kỳ biến nào trong sáu biến. Tuy nhiên, tất cả các thước đo đều có kết quả MDI tương tự nhau đối với hầu hết các biến. Có lẽ quan trọng nhất là các thước đo này đều cung cấp sức mạnh giải thích trong mẫu đáng kể, mặc dù chúng là những thước đo đơn giản được thiết kế cho một thế giới đơn giản hơn. [6]

5.2 Kết quả MDA

Tiếp theo, chúng tôi chuyển sang đánh giá tầm quan trọng đặc trưng theo MDA. Bảng 3 tóm tắt kết quả tầm quan trọng đặc trưng theo MDA cho từng biến được dự báo. Trái ngược với MDI, MDA là một phương pháp *ngoài mẫu* phản ánh sức mạnh dự báo của từng đặc trưng. Theo đó, các đầu ra của MDA không được đảm bảo là dương (một số đặc trưng thậm chí có thể gây bất lợi cho mục đích dự báo), và cũng không được chuẩn hoá. Như có thể thấy trong bảng, có một số giá trị âm nhưng gần bằng 0, và cách diễn

[6]Tuy nhiên cần lưu ý rằng tại mỗi phép chia trong các cây chúng tôi chỉ xét hai đặc trưng. Một đặc trưng chưa bao giờ hữu ích sẽ có MDI bằng không, nhưng một đặc trưng đôi khi tốt hơn đặc trưng thay thế đơn lẻ mà nó được so sánh sẽ có MDI khác không.

giải là chúng đóng góp ít vào dự báo ngoài mẫu dù có thể có sức mạnh giải thích trong mẫu. Mỗi hàng tương ứng với một kích thước cửa sổ nhìn lại như chỉ ra ở cột thứ nhất. Mỗi ô được trình bày dưới dạng "giá trị trung bình điểm tầm quan trọng MDA" ± "độ lệch chuẩn điểm tầm quan trọng MDA", trong đó trung bình và độ lệch chuẩn được tính trên toàn bộ 87 công cụ. Giá trị tầm quan trọng cao nhất được in đậm cho mỗi kích thước cửa sổ. Cột cuối cùng tóm tắt độ chính xác dự báo ngoài mẫu được tính trung bình trên tất cả các công cụ.

Đối với dự báo chênh lệch mua–bán, VPIN có tầm quan trọng đặc trưng cao nhất cho mọi kích thước cửa sổ, và nó cũng có tầm quan trọng cao nhất cho 5 hoặc 6 kích thước cửa sổ trong dự báo độ nhọn và kiểm định Jarque–Bera. Thước đo Roll chiếm ưu thế trong dự báo tương quan tuần tự. Đối với dự báo biến động thực tế, thước đo Roll tốt hơn cho các cửa sổ ngắn, với VPIN bám sát ngay sau. Tuy nhiên, với các cửa sổ nhìn lại dài hơn, VPIN lại mang lại tầm quan trọng lớn hơn cho dự báo biến động thực tế, trong khi thước đo Roll nhìn chung đóng góp ít cho dự báo ngoài mẫu. Điều đáng chú ý là, VIX hầu như không có sức mạnh dự báo ngoài mẫu bất kể kích thước cửa sổ. Cuối cùng, đối với dự báo độ lệch tuyệt đối, kết quả tầm quan trọng đặc trưng là hỗn hợp, với λ của Kyle, VPIN, Roll Impact và Roll Measure mỗi thước đo có tầm quan trọng cao hơn ở các kích thước cửa sổ cụ thể. :contentReference[oaicite:1]index=1 :contentReference[oaicite:2]index=2 :contentReference[oaicite:3]index=3

Chúng tôi diễn giải các kết quả này như là bằng chứng ủng hộ sức mạnh dự báo của các thước đo vi cấu trúc phản ánh ma sát trên thị trường. VPIN nhìn chung là quan trọng nhất trong việc dự báo các biến mà lẽ ra phải chịu ảnh hưởng của sự hiện diện giao dịch dựa trên thông tin: chênh lệch và các thước đo "đuôi dày" trong phân phối lợi suất. Thước đo Roll được xây dựng từ tương quan trong biến động giá, vì vậy không có gì ngạc nhiên khi thước đo này có một mức độ sức mạnh giải thích đối với tương quan tuần tự của lợi suất. Cuối cùng, dù chúng tôi đưa VIX vào tập đặc trưng, nó không nhằm phản ánh các ma sát vi cấu trúc, vì vậy việc nó có ít sức mạnh giải thích đối với các biến mà chúng tôi cố gắng dự báo cũng không có gì bất ngờ. :contentReference[oaicite:4]index=4

Các mức độ chính xác tổng thể trong Bảng 3 cho thấy thuật toán học máy của chúng tôi đang nắm bắt được điều gì đó có giá trị. Đối với phân loại chuỗi thời gian tài chính nhị phân, một bộ phân loại thường cho độ chính xác quanh mức 0.5. Sự bất lực mang tính chuẩn mực này trong việc làm tốt hơn đoán ngẫu nhiên phù hợp với giả thuyết thị trường hiệu quả: đối với các thị trường có tính thanh khoản, thị trường hiệu quả phần lớn thời gian và hành xử như một bước ngẫu nhiên. Do đó, bất cứ mức nào vượt quá 0.5 đều có thể được xem là nắm bắt một khả năng kém hiệu quả của thị trường và

vì thế là một kết quả tích cực. Ngoại trừ ước lượng chênh lệch mua–bán, các mức độ chính xác ngoài mẫu của chúng tôi đạt đỉnh trong khoảng từ 0.54 đến 0.61 (tuỳ thuộc kích thước cửa sổ nhìn lại), điều mà theo chuẩn của học máy tài chính là rất tốt. :contentReference[oaicite:5]index=5

Độ chính xác đối với chênh lệch mua–bán thì không tốt bằng. Chúng tôi suy đoán rằng điều này là do thiếu một chênh lệch mua–bán quan sát được trong thị trường hợp đồng tương lai; chúng tôi ước tính chênh lệch này bằng bộ ước lượng Corwin–Schultz. Có thể bản thân sai số của kỹ thuật này khi áp dụng cho hợp đồng tương lai khiến việc ước lượng bằng phương pháp rừng ngẫu nhiên kém hiệu quả. Ngoài ra, cũng có thể cách tiếp cận *dollar-volume bar* được sử dụng ở đây không phù hợp với ước lượng cụ thể này. Chúng tôi khảo sát khả năng này trong Mục 5.4, nơi chúng tôi chạy lại phân tích bằng *time bars*.

6 Kết quả và phân tích

Trong phần này, chúng tôi trình bày kết quả dự báo được tạo ra bởi mô hình rừng ngẫu nhiên, sử dụng các biến vi cấu trúc đã mô tả ở trên. Các kết quả được đánh giá cả trong mẫu (MDI) và ngoài mẫu (MDA).

Kết quả trong mẫu (MDI)

Bảng 1 báo cáo giá trị MDI trung bình của các biến vi cấu trúc đối với từng biến mục tiêu. Kết quả cho thấy tất cả các biến vi cấu trúc đều có khả năng giải thích trong mẫu, nhưng tầm quan trọng tương đối của chúng khác nhau.

Ví dụ:

- Đối với dự báo chênh lệch mua – bán, thước đo Amihud và VPIN có tầm quan trọng cao nhất.
- Đối với dự báo biến động, VPIN và VIX có sức mạnh giải thích nổi bật.
- Đối với dự báo thống kê Jarque–Bera, các thước đo biến động và Amihud chiếm ưu thế.

Kết quả MDI cho thấy rằng một số biến vi cấu trúc truyền thống, mặc dù đơn giản, vẫn có khả năng nắm bắt các yếu tố quan trọng trong hành vi thị trường.

Kết quả ngoài mẫu (MDA)

Bảng 2 báo cáo giá trị MDA trung bình. Trong khi kết quả MDI gợi ý rằng nhiều biến có sức mạnh giải thích trong mẫu, kết quả ngoài mẫu cho thấy VPIN có sức mạnh dự báo vượt trội.

Cụ thể:

- VPIN là biến quan trọng nhất cho 5 trong số 6 mục tiêu dự báo.
- Roll measure là biến quan trọng nhất cho mục tiêu còn lại (tương quan tuần tự).
- Amihud có tầm quan trọng cao trong mẫu, nhưng ngoài mẫu tầm quan trọng giảm đáng kể.

Kết quả này nhấn mạnh sự khác biệt giữa giải thích trong mẫu và dự báo ngoài mẫu. Nhiều biến truyền thống có thể giải thích được động lực giá, nhưng chỉ một số ít (như VPIN) thực sự hữu ích trong dự báo ngoài mẫu.

So sánh theo loại tài sản

Khi chia mẫu theo loại tài sản (cổ phiếu, tiền tệ, hàng hoá, lãi suất, thu nhập cố định), chúng tôi quan sát thấy một số khác biệt:

- VPIN đặc biệt mạnh trong thị trường cổ phiếu và tiền tệ.
- Amihud quan trọng hơn trong thị trường thu nhập cố định.
- Các thước đo biến động (cao–thấp) có tầm quan trọng lớn hơn trong hàng hoá.

Độ chính xác dự báo

Chúng tôi tính độ chính xác dự báo bằng 10-fold Purged Cross-Validation. Độ chính xác trung bình dao động từ 55% đến 65% tuỳ biến mục tiêu, cao hơn đáng kể so với mức 50% ngẫu nhiên.

Kết quả này xác nhận rằng các biến vi cấu trúc, đặc biệt là VPIN, có khả năng cung cấp thông tin hữu ích cho dự báo động lực giá thị trường.

Kiểm tra độ vững (Robustness Checks)

Để đảm bảo kết quả không bị ảnh hưởng bởi lựa chọn siêu tham số hoặc cửa sổ quan sát, chúng tôi tiến hành một số kiểm tra:

- Sử dụng các kích thước cửa sổ khác nhau ($W = 1, 2, 3$ tuần).
- Thay đổi số lượng cây trong rừng ngẫu nhiên (100, 500, 1000).
- So sánh các loại bars khác nhau (time bars, volume bars, dollar-volume bars).

Kết quả nhất quán: VPIN vẫn duy trì vai trò nổi bật ngoài mẫu, trong khi Amihud và Roll quan trọng hơn trong mẫu.

7 Kết luận và định hướng nghiên cứu tương lai

Trong nghiên cứu này, chúng tôi đã chỉ ra rằng các thước đo vi cấu trúc thị trường cung cấp thông tin hữu ích cho việc dự báo động lực giá và biến động trong các thị trường tài chính hiện đại. Mặc dù các thước đo này được phát triển trong bối cảnh thị trường đơn giản hơn nhiều, chúng vẫn mang lại giá trị trong kỷ nguyên mà giao dịch dựa trên máy tính và dữ liệu lớn thống trị.

Bằng cách sử dụng các kỹ thuật học máy, cụ thể là rừng ngẫu nhiên, chúng tôi có thể đánh giá sức mạnh dự báo của nhiều thước đo vi cấu trúc khác nhau mà không cần giả định trước dạng hàm của mối quan hệ giữa chúng và các biến kết quả. Kết quả của chúng tôi cho thấy:

- Các thước đo như Amihud, Roll, và các ước lượng biến động có sức mạnh giải thích trong mẫu.
- VPIN nổi bật như một thước đo có khả năng dự báo ngoài mẫu mạnh mẽ, vượt trội so với các biến truyền thống.
- Sức mạnh dự báo có sự khác biệt giữa các loại tài sản, nhưng tính ưu việt của VPIN vẫn được duy trì trên hầu hết các nhóm.

Những phát hiện này mang lại một số hàm ý thực tiễn:

1. **Quản lý rủi ro.** Các thước đo vi cấu trúc có thể được sử dụng để nâng cao mô hình rủi ro, giúp dự báo các giai đoạn biến động cao hoặc thiếu thanh khoản.
2. **Chiến lược giao dịch.** Thông tin từ VPIN và các thước đo khác có thể được tích hợp vào thuật toán thực thi để điều chỉnh tỷ lệ tham gia, giảm chi phí giao dịch và rủi ro thị trường.

3. **Tạo lập thị trường.** Các nhà tạo lập thị trường có thể sử dụng những dự báo này để điều chỉnh báo giá, từ đó cải thiện khả năng quản lý tồn kho và giảm rủi ro bị giao dịch với đối tác có thông tin.

Về mặt học thuật, nghiên cứu của chúng tôi cho thấy rằng sự kết hợp giữa lý thuyết vi cấu trúc và các công cụ học máy phi tham số có thể mở ra hướng nghiên cứu mới. Thay vì tập trung vào việc xác định đặc tả mô hình tham số, các nhà nghiên cứu có thể khai thác sức mạnh của học máy để phát hiện ra biến quan trọng và hiểu vai trò của chúng trong hành vi thị trường.

Trong tương lai, chúng tôi đề xuất một số hướng nghiên cứu:

- Khảo sát các thuật toán học máy khác (như boosting, mạng nơ-ron sâu) và so sánh với rừng ngẫu nhiên.
- Mở rộng phân tích sang các thị trường mới nổi hoặc thị trường phi tập trung.
- Kết hợp thêm dữ liệu vi mô (như độ sâu sổ lệnh) để cải thiện sức mạnh dự báo.

Tóm lại, kết quả của chúng tôi chứng minh rằng nghiên cứu vi cấu trúc không chỉ còn giữ vai trò giải thích, mà còn có khả năng dự báo mạnh mẽ khi được tích hợp với công cụ học máy. Trong kỷ nguyên máy móc, các thước đo vi cấu trúc tiếp tục là công cụ không thể thiếu cho cả nhà nghiên cứu lẫn người tham gia thị trường.