BÁO CÁO VÒNG 2: DATA SCIENCE TALENT COMPETITION 2025

I. TỔNG QUAN

Link github: https://github.com/m1htan/-Round-2-CTE-FTU

Hướng dẫn sử dụng: README.md

Mục lục báo cáo:

II. PHẦN BÁO CÁO CHÍNH

BƯỚC 1: TỔNG QUAN VỀ THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU

1.1 Mục tiêu

Giai đoạn đầu tiên của nghiên cứu tập trung vào việc xây dựng một quy trình thu thập dữ liệu vừa an toàn, vừa có thể lặp lại (reproducible pipeline). Quy trình này đảm bảo khả năng thu thập, xác thực và làm sạch dữ liệu thị trường chứng khoán Việt Nam một cách có hệ thống. Việc kiểm soát độ tin cậy ngay từ cấp độ dữ liệu mang ý nghĩa nền tảng, bởi nó đóng vai trò làm cơ sở thực nghiệm cho tất cả các giai đoạn tiếp theo, bao gồm kỹ thuật đặc trưng (feature engineering), xây dựng chỉ báo (indicator construction) và phát triển mô hình dự báo (predictive modeling).

1.2. Nguồn dữ liệu và phạm vi bao phủ

Bộ dữ liệu được thu thập từ nền tảng **FiinQuantX**, truy cập thông qua FiinSession API với cơ chế xác thực an toàn được quản lý bằng biến môi trường (environment variables). Phạm vi bao phủ được xác định bao gồm tất cả chứng khoán nằm trong rổ HNX Index từ ngày 01/01/2022 đến 30/08/2025. Kết quả tạo ra một panel dữ liệu nhiều năm, cho phép áp dụng cả hai hướng phân tích: chéo (cross-sectional) và theo chiều dọc (longitudinal analysis).

Đối với từng mã chứng khoán, dữ liệu quan sát hàng ngày được thu thập trên cơ sở đã điều chỉnh (adjusted) để phản ánh các sự kiện doanh nghiệp (corporate actions). Bộ biến được bao gồm các chỉ số OHLCV cơ bản (giá mở cửa, giá cao nhất, giá thấp nhất, giá đóng cửa và khối lượng giao dịch) cùng với các chỉ số vi cấu trúc thị trường (order-flow metrics: bu, sd, fn, fs, fb) nhằm phản ánh thanh khoản và hành vi giao dịch ở mức chi tiết hơn.

Để tuân thủ giới hạn của API và duy trì tính ổn định hệ thống, quy trình thu thập được tổ chức theo lô (batch) gồm 80 mã cổ phiếu, với khoảng dừng kiểm soát là 0,8 giây giữa các yêu cầu. Toàn bộ dữ liệu thu thập được sau đó được ghép nối, sắp xếp theo thứ tự thời gian cho từng mã, và lưu trữ trong tệp dữ liệu gốc raw_stocks.csv.

1.3 Xác thực dữ liệu

Sau khi tổng hợp dữ liệu gốc, một quy trình xác thực nhiều bước đã được triển khai nhằm đánh giá tính chính xác và tính nhất quán nội bộ của dữ liệu. Phạm vi thời gian được kiểm tra để đảm bảo trùng khớp với giai đoạn nghiên cứu dự kiến, xác nhận rằng quan sát sớm nhất và muộn nhất đều nằm trong khoảng thời gian yêu cầu.

Các kiểm tra logic được thực hiện trên bộ biến OHLCV: loại bỏ các quan sát có giá âm hoặc bằng 0, khối lượng giao dịch âm, hoặc trường hợp giá mở cửa hay giá đóng cửa nằm ngoài khoảng giá cao nhất - thấp nhất trong ngày. Dữ liệu trùng lặp được phát hiện ở cấp độ mã-thời gian và được thống kê nhằm đánh giá mức độ dư thừa. Phân tích dữ liệu bị thiếu theo ngày giao dịch cũng được thực hiện bằng cách đối chiếu các ngày quan sát với lịch ngày làm việc dự kiến, từ đó nhận diện khoảng trống dữ liệu của từng mã chứng khoán.

Đặc biệt, với các biến order-flow, dữ liệu được kiểm tra chặt chẽ: khối lượng mua và bán không được vượt quá tổng khối lượng giao dịch, phải luôn không âm, và phù hợp với dòng vốn ròng đã báo cáo trong biên độ dung sai 5%. Cuối cùng, thống kê mô tả được tính toán cho các biến OHLCV nhằm phát hiện ngoại lệ và mô tả đặc tính phân phối của mẫu dữ liệu.

1.4 Làm sạch dữ liệu

Giai đoạn làm sạch dữ liệu đã loại bỏ toàn bộ các quan sát không đạt chuẩn xác thực, chỉ giữ lại những bản ghi OHLCV hợp lệ và có tính nhất quán nội bộ. Các bản ghi trùng lặp bị loại bỏ để đảm bảo tính duy nhất theo thời gian ở cấp độ mã chứng khoán.

Mỗi mã được tái chỉ mục theo lịch ngày liên tục trong toàn bộ giai đoạn quan sát. Các giá liên quan được nội suy tuyến tính để đảm bảo tính liên tục, trong khi khối lượng giao dịch được gán giá trị 0 trong các trường hợp không có quan sát, phản ánh giả định về trạng thái không giao dịch. Nhờ vậy, kết quả là một panel dữ liệu có cấu trúc đồng nhất, trong đó tất cả các mã đều được căn chỉnh trên cùng một lưới thời gian, tạo điều kiện cho các bước kỹ thuật đặc trưng tiếp theo. Bộ dữ liệu đã được làm sạch được lưu dưới dạng cleaned_stocks.csv, đóng vai trò là đầu vào chuẩn hóa cho các giai đoạn phân tích tiếp theo.

1.5 Phân tích khám phá dữ liệu

Để cung cấp cái nhìn tổng quan về đặc điểm cấu trúc của bộ dữ liệu sạch, nhóm nghiên cứu đã tiến hành phân tích dữ liệu khám phá. Số lượng chứng khoán và số quan sát, ranh giới thời gian, cùng với phân phối chung của giá và khối lượng giao dịch đều được xác nhận phù hợp với kỳ vọng.

Các trực quan hóa dữ liệu giúp phát hiện những đặc điểm nổi bật: biểu đồ cột thể hiện 10 mã có giá đóng cửa trung bình và khối lượng giao dịch trung bình cao nhất; histogram mô tả phân phối tổng thể của giá đóng cửa; boxplot phân tích mức độ phân tán theo từng sàn giao dịch; đồ thị chuỗi thời gian minh họa xu hướng và biến động của một số chứng khoán tiêu biểu. Bên cạnh đó, ma trận nhiệt (correlation heatmap) về tương quan lợi suất hàng ngày giữa các cổ phiếu thanh khoản cao cho thấy mối quan hệ phụ thuộc và sự đồng biến động trong thị trường.

1.6 Kết luận

Tóm lại, Bước 1 đã thành công trong việc xây dựng một bộ dữ liệu vững chắc và đáng tin cậy, thông qua quy trình thu thập an toàn - xác thực toàn diện - làm sạch hệ thống. Bộ panel kết quả có phạm vi bao phủ rộng, cấu trúc thống nhất và căn chỉnh theo cùng một lưới thời gian, đáp ứng đầy đủ yêu cầu cho các mô hình định lượng nghiêm ngặt.

Một số hạn chế còn tồn tại, bao gồm: phạm vi hiện tại chỉ giới hạn trong chứng khoán niêm yết tại HNX; việc sử dụng lịch ngày dương lịch thay vì lịch giao dịch theo sàn; và cách xử lý đơn giản đối với các biến order-flow. Tuy nhiên, hạ tầng dữ liệu được thiết lập trong bước này đã tạo nền tảng phương pháp luận vững chắc cho các phân tích thực nghiệm ở giai đoạn tiếp theo.

BƯỚC 2: XÂY DỰNG CHIẾN LƯỢC

2.1. Mục tiêu

Giai đoạn thứ hai của tập trung vào việc thiết kế và triển khai một chiến lược thống nhất cho việc xây dựng nhân tố (factor engineering) và xây dựng tín hiệu (signal construction). Chiến lược này kết hợp dữ liệu cơ bản theo quý với thông tin giá - khối lượng hàng ngày, nhằm tạo ra một panel dữ liệu sẵn sàng cho việc lựa chọn cổ phiếu, xác định thời điểm giao dịch và đánh giá chiến lược.

Mục tiêu cốt lõi là chuyển hóa dữ liệu đa dạng từ các nguồn cung cấp khác nhau và vi cấu trúc thị trường thành các biến có thể kiểm toán, bao gồm: định giá, tăng trưởng, xu hướng, động lượng, biến động, và hành vi giá. Đồng thời, quy trình này phải đảm bảo tính lặp lại, độ bền vững trước dữ liệu thiếu, và khả năng so sánh chéo theo thời gian.

2.2 Kiến trúc và thiết kế dữ liệu

Chiến lược được khởi đầu bằng việc cố định toàn bộ các phép biến đổi trong khung thời gian nghiên cứu từ ngày 01/01/2022 đến 30/08/2025. Dữ liệu giá - khối lượng hàng ngày (OHLCV) đã được xác thực và làm sạch từ Bước 1 đóng vai trò là "xương sống thời gian" (temporal spine) của hệ thống. Trong khi đó, dữ liệu cơ bản (fundamentals) được truy xuất thông qua các lệnh gọi có xác thực tới **FiinQuantX**, sử dụng báo cáo tài chính hợp nhất.

Do các phản hồi từ nhà cung cấp dữ liệu có thể ở nhiều định dạng khác nhau (từ dictionary theo mã cổ phiếu, danh sách bản ghi phẳng, đến mảng đóng gói), pipeline đã được thiết kế với cơ chế chuẩn hóa payloads. Quá trình này sử dụng phương pháp duyệt sâu (deep traversal) để chuẩn hóa toàn bộ khóa đồng nghĩa (synonym keys) về một lược đồ thống nhất, bao gồm các chỉ số cơ bản: PE, PB, ROE, EPS, BVPS.

Các ranh giới quý được xác định rõ ràng bằng ngày kết thúc theo lịch của Q1–Q4, từ đó mỗi bản ghi được gán một **period_end_date** duy nhất nằm trong cửa sổ nghiên cứu. Nhờ đó, toàn bộ dữ liệu cơ bản được liên kết bằng khóa xác định (mã cổ phiếu, năm, quý), giúp cho quá trình ghép nối (merge) diễn ra một cách nhất quán và có thể kiểm toán.

2.3 Xây dựng tín hiệu cơ bản

Để tạo ra các biến tín hiệu từ dữ liệu cơ bản, nhóm nghiên cứu triển khai các bước sau:

- **EPS_TTM:** Lợi nhuận trên cổ phiếu theo quý (EPS) được sắp xếp theo thời gian cho từng mã và tính theo cửa sổ 4 quý liên tiếp (trailing twelve months). Chỉ những trường hợp có đủ tối thiểu 4 quan sát hợp lệ mới được chấp nhận để tránh mẫu số sai lệch.
- **Giá cuối quý (price_eoq):** Tại mỗi nhóm mã - năm - quý, giá đóng cửa của ngày giao dịch cuối cùng được chọn làm giá tham chiếu, thay cho việc dùng hàm cực đại (index-max) vốn có thể gây mất tính xác định.
- **PE_TTM:** Khi giá trị PE do nhà cung cấp báo cáo không có hoặc không hợp lý, pipeline sẽ tính toán PE_TTM bằng cách lấy **price_eoq chia cho EPS_TTM**, với cơ chế bảo vệ tránh mẫu số âm hoặc bằng 0.
- **PB_TTM:** Giá trị sổ sách trên mỗi cổ phiếu (BVPS) được tính trung bình động theo 2 quý trở lên để đảm bảo ổn định, và được điền tiến (forward-fill) trong trường hợp thiếu dữ liệu. Từ đó, chỉ số PB_TTM được xác định bằng **price_eoq / BVPS_TTM_base**.
- **Biến định giá ưu tiên (Preferred Valuation):** Một quy tắc lựa chọn được áp dụng: nếu **PE_filled** tồn tại và dương, thì PE là chỉ số chính; ngược lại, pipeline mặc định sử dụng PB. Cơ chế này được lưu trữ kèm thông tin về loại chỉ số được chọn để đảm bảo tính minh bạch.
- **Xếp hạng định giá:** Trong từng nhóm quý - chỉ số, pipeline tính toán thứ hạng tăng dần, trong đó giá trị thấp hơn tương ứng với cổ phiếu "rẻ hơn".
- **Động lượng lợi nhuận:** EPS được tính tốc độ tăng trưởng theo quý (QoQ), theo năm (YoY), và tăng trưởng TTM (TTM-to-TTM) với độ trễ 4 quý, nhằm phát hiện tín hiệu phục hồi lợi nhuận trung hạn.

2.4 Xây dựng tín hiệu kỹ thuật và hành vi giá

Để bổ sung dữ liệu cơ bản bằng các tín hiệu tần suất cao, bộ chỉ báo kỹ thuật toàn diện đã được xây dựng trên panel hàng ngày cho từng mã. Các nhóm chỉ báo bao gồm:

- **Xu hướng và đường trung bình:** SMA (20, 50, 200 ngày), EMA (12, 26 ngày), WMA (20 ngày).
- **Động lượng và dao động:** RSI(14), MACD(12,26,9) với các thành phần tín hiệu và histogram, Stochastic %K(14) – %D(3), Money Flow Index(14).
- **Biến động và phạm vi:** ATR(14), Bollinger Bands 20 ngày với 2 độ lệch chuẩn.
- **Khối lượng và dòng tiền:** VWAP 14 ngày, OBV (On-Balance Volume).
- **Chỉ báo nâng cao:** PSAR, Supertrend, bộ Ichimoku (conversion, base, spans A/B, lagging line), Aroon, ZigZag swing.
- **Hành vi giá và cấu trúc:** Fair-value gaps, swing high–low, Break of Structure (BOS), Change of Character (CHoCH), order-block detection, liquidity sweeps.

Tất cả các chỉ báo đều được triển khai với cơ chế nhóm theo mã (per-ticker grouping) để tránh sai lệch về độ dài chuỗi, đồng thời trả về chuỗi dữ liệu khớp với chỉ mục gốc.

2.5 Tích hợp, căn chỉnh và hợp nhất dữ liệu

Để đảm bảo tính toàn vẹn dữ liệu, một bước kiểm tra lược đồ (schema check) được thực hiện, bao gồm bắt buộc có các cột timestamp, ticker, OHLCV và xác thực dữ liệu dạng số. Sau đó, dữ liệu được sắp xếp hoàn toàn theo mã - thời gian, đồng thời loại bỏ bản ghi trùng.

Vì dữ liệu cơ bản có tần suất theo quý, còn OHLCV và kỹ thuật theo ngày, nên việc hợp nhất diễn ra theo cơ chế “many-to-one left join”: mỗi ngày giao dịch được gắn với dữ liệu cơ bản của quý gần nhất. Các biến kỹ thuật được điền tiến (forward-fill) trong phạm vi quý để xử lý khoảng trống do độ trễ cửa sổ tính toán, trong khi dữ liệu cơ bản giữ nguyên đến quý kế tiếp.

Để giảm thiểu sai lệch do giai đoạn khởi tạo (cold-start bias), panel loại bỏ các quan sát ban đầu cho đến khi SMA-200 được tính toán đầy đủ. Điều này đảm bảo rằng tất cả các mã khi đưa vào mô hình đều có nền tảng kỹ thuật tương đồng.

2.6 Quản trị độ bền vững, dữ liệu thiếu và dữ liệu rỗng

Quy trình làm sạch chuyên biệt được thiết lập nhằm xử lý dữ liệu thiếu một cách minh bạch và có thể kiểm thử.

- Các cột bắt buộc (timestamp, mã, OHLCV) không được phép thiếu; bản ghi vi phạm sẽ bị loại bỏ.
- Dữ liệu cơ bản được điền tiến trong phạm vi mã để phản ánh tính liên tục kế toán; nếu vẫn thiếu, sẽ được bù chéo theo median trong ngày, và cuối cùng là median toàn cột.
- Các biến phân loại như nhãn chỉ số định giá (PE/PB) mặc định là "unknown" nếu không thể suy ra.
- Các biến tín hiệu dạng sự kiện hoặc cấu trúc (BOS, CHoCH, OB, liquidity…) được mặc định về 0 khi thiếu, đồng thời tạo cột nhị phân đi kèm (_isnull) để lưu giữ thông tin thiếu mang tính gợi ý.
- Các chỉ báo kỹ thuật liên tục (ví dụ RSI, MACD) giữ nguyên NaN trong giai đoạn cửa sổ khởi tạo; các quan sát đó bị loại bỏ khi loại bỏ warm-up. Sau đó, dữ liệu còn thiếu được điền tiến trong phạm vi mã.

Mỗi lần làm sạch đều tạo báo cáo trước/sau về số lượng và tỷ lệ dữ liệu thiếu, giúp việc đánh giá tác động trở nên minh bạch và có thể kiểm toán.

2.7 Xếp hạng, logic lựa chọn và mã hóa ưu tiên

Do các chỉ số định giá có thể được báo cáo trực tiếp hoặc tính toán lại từ dữ liệu TTM, pipeline lưu giữ đồng thời cả giá trị thực và nguồn gốc lựa chọn. Biến định giá ưu tiên (preferred valuation) thể hiện quy tắc: PE được ưu tiên khi khả dụng và hợp lý, ngược lại dùng PB.

Để chuẩn hóa giữa các thang đo khác nhau, tất cả giá trị định giá đều được quy đổi thành thứ hạng chéo (cross-sectional rank) theo từng quý. Cách biểu diễn dựa trên thứ hạng này giúp giảm ảnh hưởng từ giá trị ngoại lai, đặc biệt quan trọng trong bối cảnh thị trường mới nổi nơi biến động cực đoan thường xuyên xảy ra. Hệ biến xếp hạng này có thể kết hợp với các tín hiệu khác (động lượng, biến động, cấu trúc giá) để xây dựng chỉ số tổng hợp trong giai đoạn sau.

2.8 Kỹ thuật độ tin cậy và khả năng tái lập

Mọi lệnh gọi dữ liệu từ xa được bọc trong lớp retry theo cơ chế exponential backoff kèm jitter để xử lý lỗi tạm thời từ nhà cung cấp. Tỷ lệ và chỉ số được truy xuất theo lô nhỏ

nhằm tuân thủ giới hạn API, đồng thời hệ thống log ghi nhận hình dạng payload, mẫu khóa, và tỷ lệ null ở các biến then chốt như EPS_TTM và PE_TTM.

Các khóa dữ liệu và bản đồ đồng nghĩa được chuẩn hóa thành token chữ hoa – số để tránh tình trạng sai khác do phân biệt hoa/thường hoặc ký tự đặc biệt. Mỗi bước biến đổi đều ghi lại kết quả trung gian dưới dạng file cụ thể:

- **HNX_fundamental_ratios_quarterly.csv** - dữ liệu cơ bản theo quý
- **HNX_ohlcv_with_fundamentals.csv** - OHLCV hàng ngày kèm dữ liệu cơ bản
- **HNX_technical_indicators.csv** - tập chỉ báo kỹ thuật
- Bảng panel hàng ngày đã hợp nhất với toàn bộ tín hiệu
- Các phiên bản đã làm sạch cùng báo cáo dữ liệu thiếu

Nhờ cơ chế này, toàn bộ kết quả có thể tái tạo chính xác từ dữ liệu thô ban đầu.

2.9 Đầu ra và các cân nhắc thực tiễn

Đầu ra cuối cùng của Bước 2 là một panel dữ liệu dày đặc theo tần suất ngày, trong đó mỗi quan sát được gắn kèm:

- Dữ liệu cơ bản theo quý đã căn chỉnh
- Chỉ số định giá ưu tiên và nguồn gốc của nó
- Các tín hiệu tăng trưởng lợi nhuận
- Bộ chỉ báo kỹ thuật toàn diện (xu hướng, động lượng, biến động, khối lượng)
- Các cờ hành vi giá mang tính cấu trúc (price-action markers)

Panel được cắt bỏ giai đoạn warm-up, điền tiến hợp lý để duy trì tính liên tục, đồng thời được gắn cờ dữ liệu thiếu nhằm bảo toàn thông tin về quá trình phát sinh dữ liệu.

Hai vấn đề thực tiễn cần nhấn mạnh:

1. Việc sử dụng điền tiến và bù median là hợp lý để duy trì tính liên tục, nhưng cần được kiểm tra sức chịu đựng (stress-test) trong các thí nghiệm ablation, đặc biệt đối với các chiến lược nhạy cảm với thời điểm công bố và vi cấu trúc.

2. Framework được thiết kế dạng mô-đun: các ưu tiên (như PB thay cho PE), định nghĩa thay thế cho BVPS rolling, hay lịch giao dịch theo sàn có thể thay đổi mà không làm gián đoạn logic quản trị dữ liệu.

Nhờ đó, hệ thống không chỉ đảm bảo tính bền vững và có thể tái lập, mà còn tạo sự linh hoạt để thích ứng với các tùy chọn nghiên cứu khác nhau.

BƯỚC 3: THIẾT KẾ LOGIC LỌC CỔ PHIẾU

3.1 Mục tiêu

Bước thứ ba trong pipeline tập trung vào việc xây dựng một logic lọc cổ phiếu có hệ thống cho tập hợp chứng khoán niêm yết trên Sở Giao dịch Chứng khoán Hà Nội (HNX). Mục tiêu chính là xác định một tập con cổ phiếu đáp ứng đồng thời các điều kiện nghiêm ngặt về cả kỹ thuật lẫn cơ bản, sau đó sắp xếp chúng dựa trên một cơ chế xếp hạng tổng hợp. Nhờ đó, quá trình này không chỉ dừng lại ở phân tích mô tả, như sàng lọc theo xu hướng và định giá, mà còn đóng vai trò cầu nối đến giai đoạn mô hình dự báo, nơi các nhãn huấn luyện được tạo ra để phục vụ học có giám sát ở Bước 4.

3.2 Tiền xử lý dữ liệu và chính sách xử lý dữ liệu thiếu

Tập dữ liệu đầu vào tại bước này là sự tích hợp của ba nhóm thành phần chính: dữ liệu giá (OHLCV), dữ liệu cơ bản và các chỉ báo kỹ thuật đã được xây dựng trong Bước 2. Dữ liệu được sắp xếp theo trình tự thời gian, đồng bộ cho từng mã cổ phiếu và từng mốc thời gian, nhằm đảm bảo tính nhất quán trước khi tiến hành lọc.

Đối với vấn đề dữ liệu thiếu, pipeline áp dụng chính sách **“ffill_then_drop”**. Theo đó, các giá trị bị thiếu trong mỗi chuỗi thời gian của một mã cổ phiếu sẽ được điền tiến (forward-fill) dựa trên quan sát gần nhất hợp lệ. Sau khi điền tiến, những bản ghi vẫn còn thiếu ở các biến then chốt sẽ bị loại bỏ hoàn toàn. Cách tiếp cận này bảo đảm rằng dữ liệu sử dụng trong quá trình lọc vừa duy trì được tính liên tục trong từng chuỗi cổ phiếu, vừa loại bỏ được các điểm quan sát không đầy đủ - vốn có thể gây sai lệch cho logic sàng lọc.

Các biến bắt buộc phải đầy đủ bao gồm: giá đóng cửa, các đường trung bình động đơn giản (SMA50, SMA200), đường trung bình động hàm mũ (EMA12, EMA26), histogram của MACD, chỉ số RSI, ADX, Supertrend, Parabolic SAR, tỷ lệ định giá ưu tiên, tốc độ tăng trưởng lợi nhuận trên mỗi cổ phiếu (EPS TTM YoY), tỷ suất lợi nhuận trên vốn chủ sở hữu (ROE), khối lượng giao dịch, và mã định danh sàn giao dịch.

3.3 Logic sàng lọc theo luật định (Bước 3A)

Bộ lọc sàng lọc đầu tiên áp dụng một tập hợp quy tắc xác định, trong đó kết hợp đồng thời xác nhận xu hướng kỹ thuật và sự vững chắc về nền tảng cơ bản. Một cổ phiếu chỉ được đánh dấu là đủ điều kiện khi đồng thời thỏa mãn tất cả các điều kiện sau:

Xác nhận xu hướng (Trend confirmation):

- $P_{close} > SMA_{50} > SMA_{200}$, thể hiện cấu trúc "bullish" nhất quán giữa các đường trung bình.
- $EMA_{12} > EMA_{26}$ và $MACD_{hist} > 0$, phản ánh động lượng ngắn hạn tích cực.
- RSI >= 50 và ADX >= 20, cho thấy xu hướng có độ mạnh và động lượng không yếu.
- P_{close} > Supertrend và Parabolic SAR < P_{close}, xác nhận tín hiệu xu hướng tiếp diễn.

Ràng buộc cơ bản (Fundamental constraints):

- Cổ phiếu phải nằm trong nhóm 40% có định giá rẻ nhất trên toàn sàn HNX (valuation percentile rank $\leq 40\%$).
- Tăng trưởng lợi nhuận trên mỗi cổ phiếu phải dương ($EPS_{TTM,yoy} > 0$), đồng thời tỷ suất lợi nhuận trên vốn chủ sở hữu (ROE) cũng phải duy trì giá trị dương.
- Khối lượng giao dịch phải khác 0 để bảo đảm tính thanh khoản tối thiểu.

Kết quả cuối cùng của bộ lọc này là một tín hiệu nhị phân có tên signal_rule_trend_value, phản ánh sự đồng thuận giữa tín hiệu xu hướng và giá trị cơ bản.

3.4 Cơ chế chấm điểm tổng hợp (Bước 3B)

Để sắp hạng các cổ phiếu đủ điều kiện sau bước sàng lọc, hệ thống xây dựng một cơ chế chấm điểm tổng hợp (composite score), trong đó tích hợp nhiều chiều đo lường hiệu suất. Các chỉ báo được chuẩn hóa thông qua xếp hạng phân vị chéo (cross-sectional percentile ranking) trong phạm vi toàn bộ tập cổ phiếu HNX tại từng ngày giao dịch, nhằm bảo đảm tính so sánh đồng nhất giữa các mã.

Các thành phần chính của điểm tổng hợp bao gồm:

- **Điểm động lượng (Momentum score):** dựa trên histogram của MACD, sau khi được xử lý giới hạn (winsorized) trong khoảng ±3 để giảm thiểu ảnh hưởng của ngoại lệ.
- **Điểm xu hướng (Trend score):** phản ánh sức mạnh tương đối của giá so với đường SMA200, qua đó đánh giá vị thế trung hạn của cổ phiếu.
- **Điểm ADX:** sử dụng giá trị ADX nhưng giới hạn trần ở mức 40 nhằm tránh nhiễu do các ngoại lệ quá lớn.
- **Điểm Aroon:** đo lường sự khác biệt tương đối giữa chỉ báo Aroon-up và Aroon-down, thể hiện mức độ chiếm ưu thế của xu hướng tăng so với xu hướng giảm.
- **Điểm định giá (Valuation score):** được tính dựa trên thứ hạng nghịch đảo của chỉ số định giá ưu tiên, trong đó cổ phiếu có định giá rẻ hơn sẽ được gán điểm cao hơn.
- **Điểm tăng trưởng lợi nhuận và khả năng sinh lời:** phản ánh thứ hạng phân vị của tốc độ tăng trưởng EPS và chỉ số ROE, qua đó đo lường sức khỏe tài chính và triển vọng kinh doanh của doanh nghiệp.

Điểm tổng hợp cuối cùng được xác định như một trung bình có trọng số (weighted average) của tất cả các thành phần trên, từ đó cung cấp một thước đo duy nhất để xếp hạng và ưu tiên các cổ phiếu trong tập hợp đủ điều kiện.

$$Score_{composite} = 0.15 \cdot Score_{mom} + 0.15 \cdot Score_{trend} + 0.30 \cdot Score_{val} + 0.20 \cdot Score_{eps} + 0.10 \cdot Score_{roe} + 0.05 \cdot Score_{adx} + 0.05 \cdot Score_{aroon}$$

Điểm tổng hợp cuối cùng được xác định như một trung bình có trọng số (weighted average) của tất cả các thành phần trên. Trong đó, hệ thống ưu tiên:

- Định giá (Valuation): 30%
- Tăng trưởng lợi nhuận (Earnings growth): 20%
- Các thành phần còn lại (động lượng, sức mạnh xu hướng, ADX và Aroon) được phân bổ trọng số nhỏ hơn nhưng vẫn được tích hợp nhằm phản ánh toàn diện sức mạnh kỹ thuật và động lượng của cổ phiếu.

Cách tiếp cận này đảm bảo rằng yếu tố nền tảng tài chính (valuation và earnings) giữ vai trò trung tâm, trong khi các tín hiệu kỹ thuật bổ sung giúp tối ưu hóa khả năng nhận diện cơ hội đầu tư.

3.5 Phát hiện tín hiệu giao cắt (Bước 3C)

Bên cạnh các bộ lọc tĩnh dựa trên quy tắc, hệ thống còn tính toán các tín hiệu giao cắt động tại cấp độ từng mã cổ phiếu. Các tín hiệu chính bao gồm:

- **Golden Cross (GC):** giao cắt hướng lên của đường SMA50 vượt lên trên SMA200.
- **Giao cắt tín hiệu MACD:** đường MACD cắt lên trên đường tín hiệu MACD (signal line).
- **Giao cắt của bộ dao động Stochastic:** bộ lọc tùy chọn nếu dữ liệu sẵn có, yêu cầu đường %K cắt lên trên đường %D trong khi vẫn nằm dưới ngưỡng quá bán (30).

Các tín hiệu dựa trên sự kiện này (gồm signal_cross_gc, signal_cross_macd, signal_cross_stoch) cho phép giám sát bổ sung các điểm đảo chiều của động lượng, bổ trợ cho logic sàng lọc tổng thể.

3.6 Lựa chọn cổ phiếu: Xếp hạng và chọn Top-N

Sau khi vượt qua bộ lọc quy tắc, các cổ phiếu thuộc rổ HNX sẽ được xếp hạng hằng ngày dựa trên điểm tổng hợp. Từ kết quả này, hệ thống giữ lại danh sách Top-N cổ phiếu có điểm cao nhất, với giá trị mặc định là $N = 20$. Cơ chế lựa chọn được thiết kế linh hoạt: trong trường hợp số lượng cổ phiếu đáp ứng đủ điều kiện ít hơn 20, danh mục đầu ra sẽ chỉ bao gồm số cổ phiếu hợp lệ hiện có. Bên cạnh đó, đối với mỗi mã chứng khoán, hệ thống lưu lại thứ hạng (rank_composite) cũng như cờ lựa chọn (pick_topN_composite) nhằm phục vụ cho các bước phân tích và dự báo tiếp theo.

3.7 Sinh nhãn cho mô hình dự báo

Để chuẩn bị cho bước 4 - huấn luyện học có giám sát, nghiên cứu tiến hành tính toán lợi suất kỳ vọng (forward-looking returns) trong các khoảng thời gian $k = 10$ và *20 phiên giao dịch*. Lợi suất kỳ vọng được xác định theo công thức:

$$fwd\_ret_k = \frac{P_{t+k}}{P_t} - 1$$

Trong đó, P_t là giá đóng cửa tại thời điểm ttt, còn P_{t+k} là giá đóng cửa sau k phiên. Trên cơ sở này, đối với từng ngày, trung vị lợi suất kỳ vọng của toàn bộ rổ HNX được sử

dụng làm mốc so sánh. Cổ phiếu nào có lợi suất cao hơn trung vị sẽ được gán nhãn 1 (outperformer), ngược lại cổ phiếu có lợi suất thấp hơn trung vị sẽ được gán nhãn 0 (underperformer). Việc gán nhãn này tạo điều kiện cho đánh giá hiệu quả mang tính tương đối trong phạm vi sàn HNX, thay vì tuyệt đối.

3.8. Thảo luận và Hàm ý

3.8.1. Lựa chọn sàn nghiên cứu: HNX

Sàn giao dịch chứng khoán Hà Nội (HNX) được lựa chọn làm phạm vi nghiên cứu vì ba lý do chính. Thứ nhất, so với HOSE, HNX thể hiện mức độ dị biệt cao hơn về quy mô doanh nghiệp, thanh khoản và phân bố ngành, nhờ đó tạo điều kiện thử nghiệm phong phú hơn cho các thuật toán sàng lọc cổ phiếu. Thứ hai, nhiều cổ phiếu trên HNX ít được nghiên cứu, định giá chưa hiệu quả, từ đó mở ra cơ hội cho các mô hình sàng lọc hệ thống khai thác những điểm phi hiệu quả này. Thứ ba, việc tập trung vào một sàn giao dịch duy nhất giúp đảm bảo tính đồng nhất về môi trường giao dịch và khung pháp lý, giảm thiểu thiên lệch từ sự khác biệt liên sàn. Cách tiếp cận này đồng thời cho phép thực hiện xếp hạng theo phần trăm một cách nhất quán trên toàn bộ vũ trụ phân tích.

3.8.2. Ngưỡng định giá 40%

Việc giới hạn lựa chọn trong nhóm 40% cổ phiếu có định giá rẻ nhất của HNX phản ánh sự cân bằng giữa thiên hướng giá trị và độ rộng cơ hội đầu tư. Nếu cắt ngưỡng quá chặt (ví dụ 20%), danh mục có thể tập trung quá nhiều vào các cổ phiếu khó giao dịch hoặc gặp vấn đề tài chính, gây suy giảm tính thanh khoản và tính ổn định. Ngược lại, nếu nới lỏng quá mức (ví dụ 60%), danh mục sẽ mất đi thiên hướng giá trị, đồng thời suy yếu cơ sở kinh tế học đằng sau tiêu chí sàng lọc. Ngưỡng 40% này phù hợp với các nghiên cứu thực nghiệm trước đây (Fama & French, 1992; Asness et al., 2013), vốn chỉ ra rằng việc hình thành danh mục dựa trên phân vị định giá mang lại mức bù rủi ro ổn định và có ý nghĩa thống kê.

3.8.3. Trọng số trong điểm tổng hợp

Cách phân bổ trọng số trong điểm tổng hợp phản ánh sự kết hợp giữa cơ sở lý thuyết và tính thực nghiệm. Trong đó, định giá (30%) và tăng trưởng lợi nhuận (20%) chiếm tỷ trọng cao nhất, bởi chúng đại diện cho hai phong cách đầu tư cốt lõi: giá trị (value) và tăng

trưởng (growth) – những yếu tố đã được chứng minh có ảnh hưởng lớn đến lợi suất cổ phiếu trong dài hạn. Các yếu tố xu hướng (trend, 15%) và động lượng (momentum, 15%) đóng vai trò bổ trợ, nhằm xác nhận rằng cổ phiếu hấp dẫn về cơ bản cũng được hỗ trợ bởi tín hiệu giá. Bên cạnh đó, ROE (10%) được đưa vào để đo lường hiệu quả sinh lời, trong khi các chỉ báo về sức mạnh xu hướng như ADX (5%) và Aroon (5%) được phân bổ trọng số thấp nhằm nắm bắt thông tin bổ sung mà vẫn tránh rủi ro nhiễu động. Tổng thể, đây là một mô hình lai cơ bản – kỹ thuật, vừa duy trì ưu thế từ các yếu tố nền tảng, vừa đảm bảo tín hiệu xu hướng ngắn hạn không bị bỏ qua. Thiết kế này phù hợp với các tài liệu kinh điển về mô hình đa nhân tố (Carhart, 1997; Jegadeesh & Titman, 1993), đồng thời điều chỉnh linh hoạt theo đặc thù thanh khoản và biến động của thị trường Việt Nam.

3.8.4. Hàm ý rộng hơn

Khung sàng lọc cổ phiếu được xây dựng có hai hàm ý đáng chú ý trong lĩnh vực khoa học dữ liệu tài chính. Thứ nhất, nó cho thấy khả năng ứng dụng của các phương pháp chọn lọc cổ phiếu dựa trên đa nhân tố tại thị trường mới nổi như Việt Nam, nơi dữ liệu còn thiếu hụt, thanh khoản thấp và biến động cao. Thứ hai, mô hình này cho thấy cách kết hợp hiệu quả giữa sàng lọc quy tắc định tính và mô hình học máy. Cụ thể, các tín hiệu, điểm tổng hợp và nhãn dự báo được thiết kế song song, giúp thu hẹp không gian giả thuyết trước khi áp dụng thuật toán dự báo. Điều này đóng góp vào dòng nghiên cứu về các hệ thống hỗ trợ quyết định lai (hybrid decision-support systems), vốn đang ngày càng được quan tâm trong khoa học tài chính hiện đại.

BƯỚC 4: Phát triển chiến lược và xác định trọng số

4.1 Mục tiêu

Step 4 phát triển một khung học máy theo thời gian (walk-forward) để: (i) ước tính xác suất vượt trội tương đối (excess return) của mỗi cổ phiếu HNX trong các chân trời $k \in \{10,20\}$ phiên; (ii) chuyển hóa xác suất này thành chiến lược Top-N (mặc định N=20) theo ngày; và (iii) ghi nhận thông tin giải thích (feature importance/coefficients) phục vụ quản trị mô hình. Khung được thiết kế "thực dụng-trong-sản-xuất": vừa tôn trọng kỷ luật chống rò rỉ dữ liệu (no look-ahead), vừa bảo đảm hiệu năng chạy nhanh (FAST MODE), và duy trì khả năng lặp lại (reproducibility).

4.2 Dữ liệu, Universe, and Labels

Dữ liệu đầu vào là bộ hợp nhất OHLCV-fundamentals-technicals (Step 2), được lọc theo HNX và điền khuyết forward-fill theo mã trước khi model hóa. Nhãn học máy được dựng theo logic hiệu năng tương đối nội vũ trụ (cross-section):

$$fwd\_ret_k(i,t) = \frac{P_{i,t+k}}{P_{i,t}} - 1, \quad label_k(i,t) = 1\{fwd\_ret_k(i,t) > \text{median}_{j \in \text{HNX}}\, fwd\_ret_k(j,t)\}.$$

Cách gán nhãn theo median hằng ngày chuẩn hóa điều kiện thị trường và có khuynh hướng cân bằng lớp (≈50/50) theo ngày, giúp thuật toán ổn định hơn so với ngưỡng tuyệt đối.

4.3. Feature Engineering

Tập đặc trưng (FEATS) bao quát bốn khối thông tin:

- **Trend/Momentum**: $\text{MACD}, \text{MACD hist}, \text{RSI}, \text{ADX}, \text{Aroon}_{up/down}, \text{Stoch}_{K/D}$, các đường trung bình $\text{SMA}_{20,50,200}, \text{EMA}_{12,26}, \text{WMA}_{20}$, dải Bollinger $(\text{mid}, \text{up}, \text{low})$.
- **Volatility/Volume**: $\text{ATR}_{14}, \text{OBV}, \text{VWAP}_{14}, \text{MFI}_{14}, \text{volume}, \text{VMA}_{\text{SMA/EMA}}$.
- **Value/Quality**: $\text{valuation\_pref}, \text{EPS TTM yoy}, \text{ROE}, \text{PB\_filled}, \text{PE\_filled}$.
- **Ratio phát sinh**: $\text{close}/\text{SMA}_{200}$ (nếu khả dụng) để bắt "distance-to-trend".

Toàn bộ FEATS được ép kiểu float32, lọc NaN/Inf, giúp giảm bộ nhớ và nhiễu số.

4.4. Model Zoo and Rationale

Khung mô hình ưu tiên đa dạng hóa giả thuyết (hypothesis diversification):

a. Logistic Regression (LR) - tuyến tính, diễn giải tốt; dùng class_weight="balanced" để xử lý lệch lớp theo ngày; yêu cầu chuẩn hóa bằng StandardScaler.
b. Random Forest (RF) - phi tuyến, chống nhiễu, không cần scaling, cung cấp Gini gain importance.
c. (Tuỳ chọn) LGBM/XGB - nếu thư viện sẵn có: boosting cây quyết định, hiệu quả với tương tác phi tuyến, hỗ trợ importance nội sinh.

Mỗi mô hình trả về xác suất $\hat{p}_k^{(m)}(i,t)$ cho nhãn $label_k$.

4.5. Walk-Forward Training Protocol

Các mốc thời gian được sắp tăng dần. Tại mỗi ngày kiểm định t:

- Tập huấn luyện dùng toàn bộ ngày < t với tối thiểu 126 ngày (≈ 6 tháng) khởi động.
- Tái huấn luyện định kỳ mỗi 10 ngày (RETRAIN_EVERY_N_D = 10) để thích ứng phi-tĩnh (non-stationarity) mà không quá tốn chi phí tính toán.
- Bỏ qua phiên nếu lớp nhãn trong huấn luyện chưa đủ hai trạng thái (an toàn thống kê).
- Ghi checkpoint mỗi 50 ngày (tùy chọn) nhằm theo dõi tiến độ và kiểm soát rủi ro tính toán.

Việc chuẩn hóa (scaling) chỉ áp dụng cho LR và được fit trên train-set duy nhất tại thời điểm tái huấn luyện để tránh rò rỉ thống kê.

4.6. Probability Aggregation and Strategy Formation

Với mỗi chân trời k, hệ thống tạo các cột xác suất theo mô hình {proba_lr_k, proba_rf_k , (proba_$lgbm_k$), (proba_xgb_k)}. Xác suất tổ hợp ensemble bằng trung bình đều:

$$\hat{p}_k^{ens}(i,t) = \frac{1}{M}\sum_{m=1}^{M}\hat{p}_k^{(m)}(i,t),$$

Trong đó M là số mô hình khả dụng tại thời điểm chạy. Trung bình đều ổn định, ít rủi ro over-fitting trọng số, phù hợp khi tập mô hình đã đủ dị biệt (linear vs. tree-based).

Danh mục Top-N theo ngày: với mỗi t và k, sắp xếp cổ phiếu theo $\hat{p}_k^{ens}(i,t)$ giảm dần, chọn tối đa N=20 mã (chấp nhận ít hơn nếu thiếu ứng viên).

4.7. Explainability and Model Governance

Khung ghi nhận "dấu vết giải thích" theo đợt tái huấn luyện:

- LR: vector hệ số (đã scale) β cho từng đặc trưng - diễn giải dấu/độ lớn (định hướng và cường độ ảnh hưởng trên log-odds).
- RF/LGBM/XGB: feature importance dựa trên gain/split.

Các bản ghi kèm mốc train_day, model, horizon, importance_type, cho phép: (i) kiểm soát trôi dạt mô hình (model drift), (ii) truy vết nguyên nhân thay đổi hiệu quả, và (iii) áp dụng kiểm định về tính ổn định đóng góp của đặc trưng.

Ghi chú: Tham số VAL_DAYS_FOR_IMPORTANCE=0 vô hiệu hóa permutation importance nhằm ưu tiên tốc độ. Cơ chế này có thể bật lại khi cần audit sâu.

4.8. Risk Controls and Data-Leakage Mitigations

Thiết kế đã cài đặt các chốt chặn rủi ro chính:

- Walk-forward nghiêm ngặt (huấn luyện < kiểm định).
- Chuẩn hóa chỉ từ train-set; forward-fill theo mã trước khi dựng nhãn để tránh "nhìn trước" mức giá tương lai.
- Class weighting cho LR giúp ổn định học trong các phiên có lệch lớp vi mô.
- Min-train-days và skip if single class bảo đảm khả năng nhận dạng ranh giới quyết định có ý nghĩa.

4.9 Strategy Implications

- **Định thời phi tuyến:** Việc phối hợp LR (tuyến tính) và cây (phi tuyến) cùng ensemble giúp chiến lược nhạy với cả cấu trúc xu hướng đơn giản lẫn tương tác bậc cao (ví dụ: "value-in-uptrend").
- **Ổn định theo chế độ thị trường**: Nhãn tương đối theo median cross-section giảm độ nhạy với trôi dạt beta thị trường, làm cho xác suất dự báo bền hơn giữa các chế độ biến động.

- **Khả năng vận hành**: Tái huấn luyện mỗi 10 ngày cân bằng giữa thích ứng (concept drift) và chi phí. Cách chọn Top-N theo ngày thuận tiện chuyển hóa sang danh mục equal-weight hoặc risk-parity ở Step 5 (nếu mở rộng).

4.10 Limitations and Future Extensions

- **Không hiệu chỉnh xác suất**: Chưa áp dụng probability calibration (Platt/Isotonic). Có thể bổ sung để cải thiện xếp hạng khi tham gia nhiều mô hình boosting.
- **Thiếu kiểm soát rủi ro giao dịch**: Khung hiện chưa tích hợp constraints (spread, turnover, capacity, T+2, lot size) và transaction costs; nên đưa vào backtest sau.
- **Ensemble bằng trung bình đều**: Dù bền vững, đôi khi kém tối ưu so với stacking hoặc Bayesian model averaging; có thể thử nghiệm khi đủ dữ liệu xác thực.
- **Permutation importance đang tắt**: khi bật, cần cấu hình VAL_DAYS_FOR_IMPORTANCE > 0 và chọn cửa sổ xác thực chéo theo thời gian (purged/embargoed CV).

BƯỚC 5: BACKTESTING AND EVALUATING PERFORMANCE

5.1 Mục tiêu

Bước 5 xây dựng một khung mô phỏng danh mục theo thời gian (time-series portfolio simulation) để đánh giá hiệu quả chiến lược lựa chọn cổ phiếu HNX được hình thành ở Step 3–4. Khung này quy định chặt chẽ cách hình thành trọng số, định nghĩa lợi suất danh mục, chi phí giao dịch, chuẩn so sánh (benchmark), và tập hợp các thước đo hiệu quả/rủi ro tiêu chuẩn (CAGR, biến động thường niên, Sharpe, Max Drawdown, tần suất thắng theo ngày), đồng thời lưu trữ nhật ký giao dịch ở cấp độ từng mã. Mục tiêu là đảm bảo tính không rò rỉ dữ liệu (no look-ahead), tái lập (reproducible) và minh bạch trong đánh giá chiến lược.

5.2. Data and Backtest Setup

Nguồn dữ liệu:

- Giá đóng cửa đã làm sạch và căn chỉnh từ Step 2 (IN_PRICES), lọc HNX theo trường Exchange.
- Danh sách picks từ Step 4 (ML ensemble, ưu tiên) hoặc Step 3 (quy tắc/score composite), tương ứng PICKS_STEP4/PICKS_STEP3.
- Chuẩn so sánh (benchmark) đọc từ BENCH_PATH ở chế độ FILE; nếu không khả dụng, dùng proxy equal-weight HNX (trung bình lợi suất chéo mã theo ngày).

Cửa sổ mô phỏng: có thể giới hạn bằng START_DATE, END_DATE, nếu không sẽ dùng toàn bộ lịch sử khả dụng.

Tham số chính: chân trời K=10 ngày (đồng bộ với mô hình Step 4), Top-N =20, chi phí giao dịch =10 bps (trên tổng biến động trọng số ngày), chế độ chọn signal source tự động ưu tiên Step 4.

5.3. Return Alignment and No Look-Ahead

Từ ma trận giá đóng cửa $P_{t,i}$ (ngày t, mã i), lợi suất **t→t+1** được định nghĩa:

$$r_{t,i} = \frac{P_{t+1,i}}{P_{t,i}} - 1,$$

và gán về ngày t (sử dụng pct_change().shift(-1)), để đảm bảo rằng trọng số $w_{t,i}$ quyết định ở cuối ngày t được nhân với lợi suất di chuyển từ t sang t+1. Cách căn chỉnh này phòng ngừa rò rỉ dữ liệu trong chuẩn hóa thời điểm giao dịch.

5.4. Portfolio Construction

5.4.1. Weight Matrix from Picks

Từ Step 4: với mỗi ngày, lọc theo chân trời K, xếp hạng cổ phiếu theo xác suất ensemble $\hat{p}_k^{ens}$ giảm dần và chọn tối đa Top-N; phân bổ **đều**: $w_{t,i}=1/N_t$ cho các mã được chọn (có thể $N_t<N$ nếu thiếu ứng viên).

Từ Step 3: nếu Step 4 không có, sử dụng cờ chọn Top-N sẵn có hoặc chuyển hóa danh sách picks thành trọng số đều.

5.4.2. Handling Missing Returns and Trading Halts

Để xử lý mã không có lợi suất (ngừng giao dịch, thiếu dữ liệu), khung áp dụng mặt nạ hợp lệ và chuẩn hóa lại theo hàng:

$$\tilde{w}_{t,i} = \frac{w_{t,i} \cdot \mathbf{1}_{\{r_{t,i} \text{ valid}\}}}{\sum_j w_{t,j} \cdot \mathbf{1}_{\{r_{t,j} \text{ valid}\}}}.$$

Nhờ vậy tổng trọng số có hiệu lực trong ngày luôn bằng 1 khi còn ít nhất một mã hợp lệ.

5.4.3. Gross and Net Returns

Lợi suất gộp (gross) của danh mục:

$$r_t^{gross} = \sum_i \tilde{w}_{t,i} \cdot r_{t,i}.$$

Chi phí giao dịch dựa trên **biến động tổng trọng số**:

$$\text{turnover}_t = \sum_i |\tilde{w}_{t,i} - \tilde{w}_{t-1,i}|, \quad \text{cost}_t = \text{TC\_RATE} \times \text{turnover}_t,$$

với TC_RATE=10 bps=0.001.

5.4.4. Benchmark

Nếu có tệp benchmark (ví dụ VN-Index) với giá đóng cửa, lợi suất **t→t+1** được suy ra bằng sai phân log tuyến tính trên tỷ lệ P_{t+1}/P_t và gán về *t*. Nếu không, dùng **proxy equal-weight HNX**: trung bình đơn lợi suất chéo mã ngày.

5.5. Performance and Risk Metrics

Từ chuỗi lợi suất ngày rtr_trt (portfolio net/benchmark), các thước đo được tính theo chuẩn thực nghiệm:

- **Đường cong vốn**: $\text{Equity}_t = \prod_{\tau \le t}(1 + r_\tau)$
- **CAGR** (annualized return) với tần suất 252 phiên: $\text{CAGR} = \left(\prod_t (1 + r_t)\right)^{252/n} - 1.$
- **Annualized Volatility:** $\sigma_{ann} = \text{stdev}(r_t) \cdot \sqrt{252}$
- **Sharpe Ratio** (lãi suất phi rủi ro = 0): $\text{Sharpe} = \frac{\overline{r_t} \cdot 252}{\sigma_{ann}}.$
- **Maximum Drawdown (MaxDD):** $\min_t \left(\frac{\text{Equity}_t}{\max_{\tau \le t} \text{Equity}_\tau} - 1\right)$
- **Số ngày/Hit-Rate**: số quan sát hợp lệ và tỷ lệ $r_t > 0$.

Các thước đo này được tính cho Portfolio (net) và Benchmark, ghi vào bảng tóm tắt HNX_backtest_summary.csv để so sánh.

5.6. Trade Formation and Trade-Level Analytics

Khung trích xuất nhật ký giao dịch bằng cách nhận diện chuỗi nắm giữ liên tục trên từng mã (đoạn có $w_{t,i} > 0$ liên tiếp). Mỗi giao dịch ghi: ticker, entry, exit, số ngày nắm giữ, và PnL:

$$\boxed{\text{PnL} = \prod_{\tau \in [t_{in}, t_{out}]} (1 + r_{\tau,i}) - 1.}$$

Từ nhật ký, tạo các chỉ tiêu: **Trades**, **WinRate**, **MedianPnL**, **AvgDays**. Các phân tích này giúp hiểu **đặc tính vi mô** của chiến lược (chu kỳ nắm giữ điển hình, phân bố PnL, độ bền giao dịch).

5.7. Visualization and Reporting Artifacts

Hệ thống xuất bộ tạo tác kết quả chuẩn hoá:

1. Chuỗi ngày (HNX_backtest_daily.csv): lợi suất gộp/ròng, turnover, chi phí, lợi suất benchmark.
2. **Equity curves** (CSV & PNG): danh mục vs. benchmark.
3. **Summary tables** (CSV): hiệu quả/rủi ro tổng hợp.
4. **Trade summary & Trade log** (CSV): thống kê và chi tiết giao dịch.
5. **Hình hoá**: đường vốn, đồ thị drawdown, phân bố lợi suất ngày (histogram).

Những tệp này hỗ trợ kiểm toán, tái lập, và trực quan hoá kết quả.

5.8. Methodological Considerations

- **Tái cân bằng theo ngày**: Top-N được xác định **mỗi ngày giao dịch**, do đó turnover phản ánh thay đổi danh mục do tín hiệu cập nhật.
- **Phân bổ đều (equal-weight)**: phù hợp khi xác suất dự báo là **thứ hạng** hơn là mức độ rủi ro tuyệt đối; tránh quá-khớp trọng số theo xác suất chưa hiệu chỉnh (uncalibrated).
- **Chi phí giao dịch tuyến tính**: mô hình hoá chi phí theo tổng $|\Delta w|$ là thực tế và thận trọng ở thị trường có thanh khoản thay đổi như HNX; người dùng có thể mở rộng sang **hàm chi phí phi tuyến** (tác động thị trường) nếu có dữ liệu độ sâu sổ lệnh.
- **Xử lý mã ngừng giao dịch**: chuẩn hoá lại theo hàng giúp danh mục **bảo toàn tổng trọng số** và không phóng đại lợi suất do dữ liệu thiếu.

5.9. Limitations and Extensions

- Không mô phỏng ràng buộc giao dịch: lô tối thiểu, T+2/T+3, trần/sàn giá, giới hạn biên độ, và các hard constraints khác chưa được đưa vào.

- Không kiểm soát rủi ro nhân tố: chưa hạn chế tiếp xúc theo ngành, size, value, momentum ở cấp danh mục; có thể thêm ex-post factor neutrality hoặc risk budgeting (volatility targeting).
- Benchmark đại diện: nếu dùng proxy EW HNX, kết luận tương đối có thể khác với VN-Index/ HNX-Index; khuyến nghị sử dụng chỉ số chuẩn thức khi có dữ liệu chất lượng.
- Hiệu chỉnh xác suất: danh mục Top-N hiện dựa trên thứ hạng $\hat{p}_i$; trong mở rộng, có thể áp dụng probability calibration và weighting theo xác suất hoặc risk-adjusted weighting.

STEP 6: ASSESSING RESILIENCE TO MARKET FLUCTUATIONS

6.1 Mục tiêu

Bước 6 xây dựng một khung stress testing theo thời gian nhằm đánh giá mức độ bền vững (resilience) của chiến lược lựa chọn cổ phiếu HNX trước các cú sốc thị trường. Phương pháp luận kết hợp (i) phát hiện tự động các giai đoạn "stress" dựa trên suy giảm lợi suất tích lũy và/hoặc bùng nổ biến động của chuẩn so sánh; (ii) mô phỏng danh mục thực thi tín hiệu (từ Step 3 – RULE hoặc Step 4 – ML), có xét trễ thi hành và chi phí ma sát; (iii) đo lường hiệu năng/rủi ro trong từng cửa sổ stress và trên toàn giai đoạn, kèm nhật ký giao dịch và tạo tác kết quả để kiểm toán.

6.2. Dữ liệu và thiết lập

Universe & Prices: Dữ liệu giá/khối lượng đã làm sạch (Step 2) lọc theo HNX. Hai ma trận theo ngày × mã được dựng: ret (lợi suất từ t→t+1, gán về t) và vol (khối lượng).

Signals/Picks: Hai chế độ:

- RULE (mặc định trong mã): đọc HNX_picks_daily.csv từ Step 3;
- ML: lọc theo chân trời dự báo K= ML_HORIZON và rank từ HNX_ml_picks_daily.csv.

Benchmark: Ưu tiên đọc từ tệp (BENCH_MODE="FILE"); nếu không, cung cấp các proxy nội bộ (HNX equal-weight, median, hoặc volume-weighted).

Execution knobs: Top-N = 20; SHIFT_PICK_D = 1 (thi hành tín hiệu vào ngày t dựa trên picks của ngày t−1); COST_BPS và SLIPPAGE_BPS (mặc định 0 trong mã, dễ mở rộng).

6.3. Portfolio Construction and Returns

6.3.1. Holdings and Delay

Với mỗi ngày ttt, danh mục được phân bổ **equal-weight** trên các mã nằm trong Top-N của ngày t−1:

$$w_{t,i} = \frac{1}{N_t}\,1\{i \in \mathrm{TopN}_{t-1}\}, \quad N_t \le 20.$$

Cơ chế trễ **SHIFT_PICK_D = 1** phản ánh ràng buộc thực thi trong thực tế (không "đi kịp" lúc picks phát sinh).

6.3.2. Net Returns with Frictions

Lợi suất gộp danh mục:

$$r_t^{gross} = \sum_i w_{t,i}\, r_{t,i}.$$

Chi phí giao dịch tuyến tính theo **turnover**:

$$\mathrm{TO}_t = \sum_i |w_{t,i} - w_{t-1,i}|, \qquad \mathrm{cost}_t = \left(\frac{\mathrm{COST\_BPS} + \mathrm{SLIPPAGE\_BPS}}{10{,}000}\right) \cdot \mathrm{TO}_t.$$

Lợi suất ròng:

$$r_t^{net} = r_t^{gross} - \mathrm{cost}_t.$$

Trong mã mẫu, chi phí đặt 0 bps (có thể tăng để kiểm định độ nhạy).

6.4. Automatic Detection of Stress Windows

Stress windows được nhận diện trên chuỗi lợi suất **benchmark** theo hai tiêu chí và một chế độ kết hợp:

1. Return drawdown window (RET_DROP)
2. Volatility spike (VOL_SPIKE)
3. BOTH

Các ngày thỏa điều kiện được gom cụm thành các đoạn liên tục (cho phép gap ≤ 3 ngày). Mỗi đoạn được định lượng mức độ nghiêm trọng bằng hiệu suất tích lũy benchmark của

chính đoạn đó sev. Cuối cùng, lựa chọn tối đa MAX_WINDOWS đoạn nghiêm trọng nhất (sev nhỏ nhất).

III. THẢO LUẬN, HẠN CHẾ VÀ HƯỚNG PHÁT TRIỂN TƯƠNG LAI

1) Thảo luận

- **Thiết kế lai cơ bản–kỹ thuật**: Điểm tổng hợp (value/growth chiếm trọng số chính) kết hợp với tín hiệu xu hướng–động lượng giúp lọc cổ phiếu "rẻ nhưng đang đi lên", đồng thời giảm nhiễu so với thuần kỹ thuật hoặc thuần cơ bản.
- **Nhãn tương đối theo median HNX**: Chuẩn hóa điều kiện thị trường theo ngày, tăng ổn định mô hình và phù hợp cho xếp hạng Top-N.
- **Walk-forward + ensemble**: Tránh rò rỉ dữ liệu, thích nghi trôi dạt (refit định kỳ), và tăng độ bền nhờ đa dạng hóa giả thuyết (LR vs. cây quyết định).
- **Backtest thực dụng**: Căn chỉnh t→t+1, chi phí theo turnover, xử lý mã ngừng giao dịch, và stress test theo cả sụt giảm lợi suất lẫn bùng nổ biến động → đánh giá được tính chịu đựng (resilience) trong các chế độ thị trường.

2) Hạn chế

- **Phạm vi**: Chỉ HNX; proxy benchmark có thể khác biệt với chỉ số chính thức.
- **Tín hiệu & trọng số**: Ensemble trung bình đều, chưa hiệu chỉnh xác suất; equal-weight chưa tính đến rủi ro cá biệt/nhân tố.
- **Chi phí & thực thi**: Mô hình chi phí tuyến tính; chưa mô phỏng lô, T+2, trần/sàn, tác động thị trường, giới hạn công suất.
- **Đánh giá thống kê**: Chưa có kiểm định chính thức (bootstrap, Diebold–Mariano), chưa báo cáo IC/Rank-IC và độ ổn định importance theo thời gian.
- **Dữ liệu**: Một số chỉ báo vi cấu trúc/quality có thể thiếu; xử lý missing chủ yếu ffill/median.

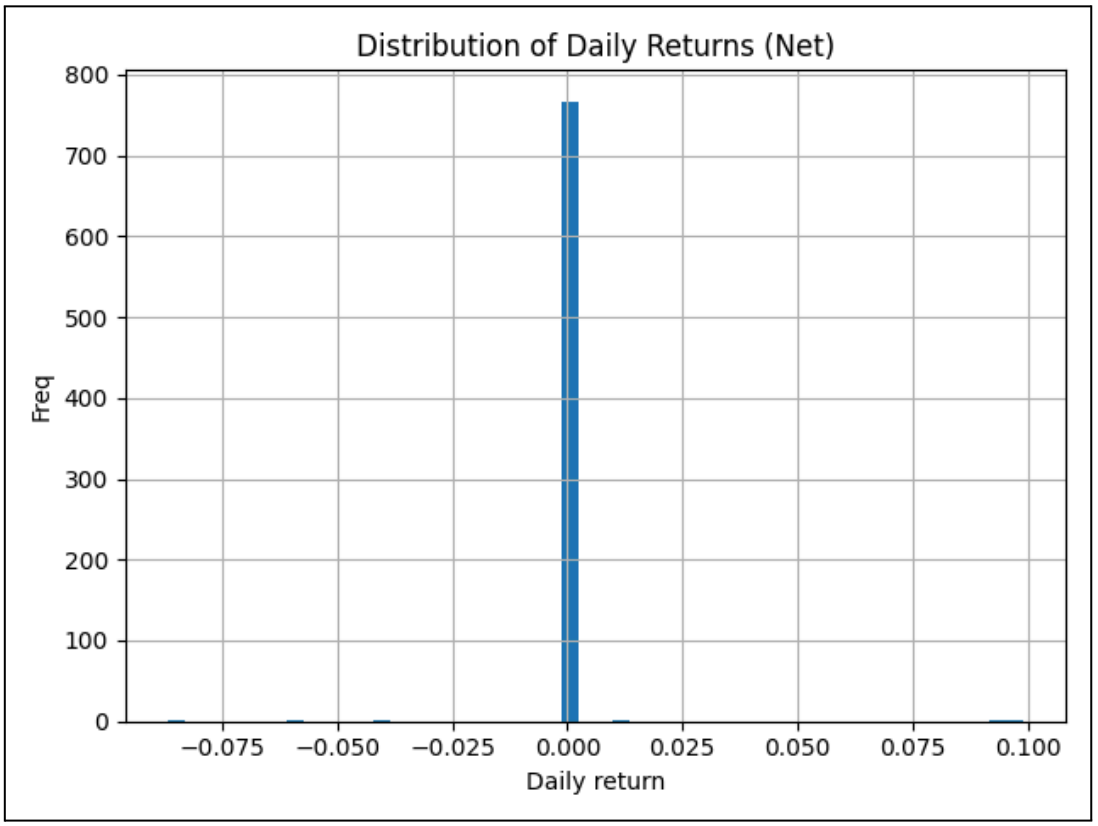
3) Hướng phát triển tương lai

- **Hiệu chỉnh & phân bổ**: Calibration xác suất (Platt/Isotonic); trọng số theo xác suất đã hiệu chỉnh và/hoặc theo rủi ro (vol-targeting, risk-parity); neutral hóa nhân tố/ngành.
- **Chi phí & thanh khoản**: Hàm chi phí phi tuyến (market impact), mô phỏng trễ/độ sâu sổ lệnh, ràng buộc công suất và giới hạn giao dịch thực tế.

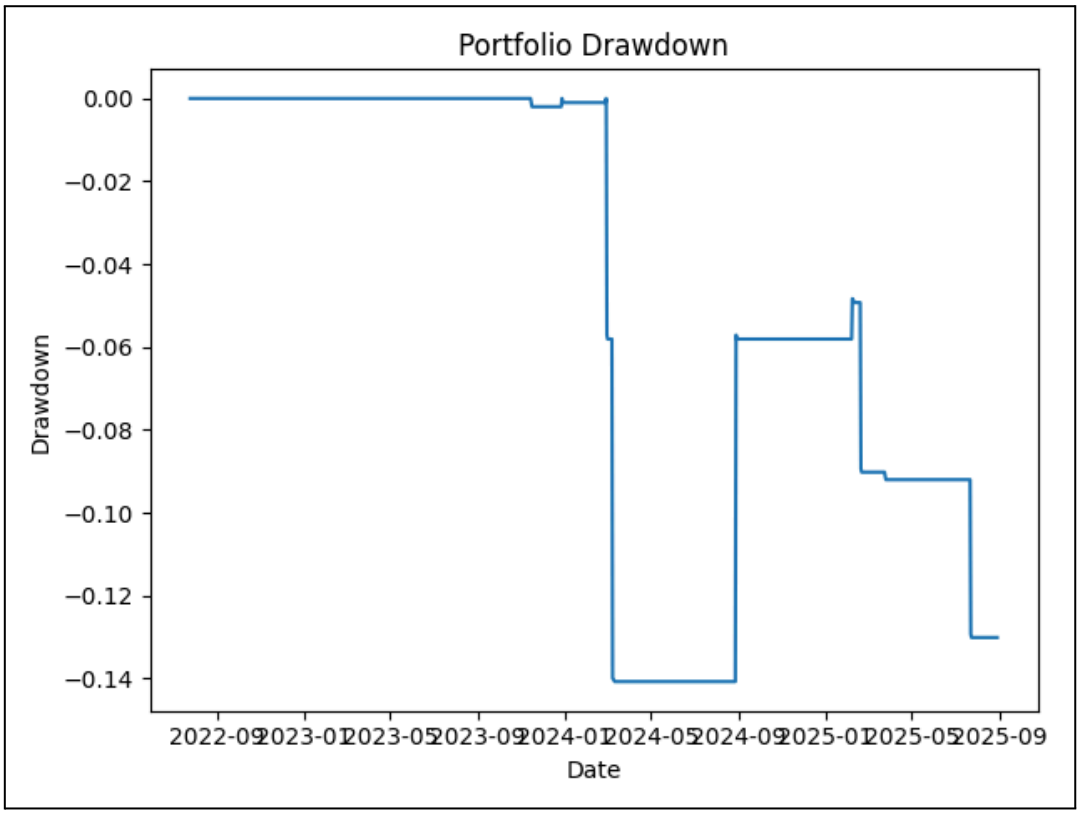
- **Đánh giá nâng cao**: IC/Rank-IC, AUC/PR@N, rolling Sharpe/MaxDD, kiểm định chênh lệch hiệu quả (DM test), phân tích độ nhạy (Top-N, COST, SHIFT, ngưỡng stress).
- **Nhận diện chế độ**: Bộ phân loại regime (bull/bear/high-vol) để điều chỉnh Top-N/đòn bẩy/tiền mặt theo trạng thái thị trường.
- **Mở rộng dữ liệu**: Mở sang HOSE/UPCoM, thêm dữ liệu intraday và yếu tố thanh khoản/phi tuyến; tinh luyện chỉ số valuation/quality.
- **Quản trị mô hình**: Giám sát drift, nhật ký explainability theo thời gian, và chu trình tái huấn luyện thích nghi/early-stopping dựa trên hiệu quả ngoài mẫu.

IV. PHỤ LỤC

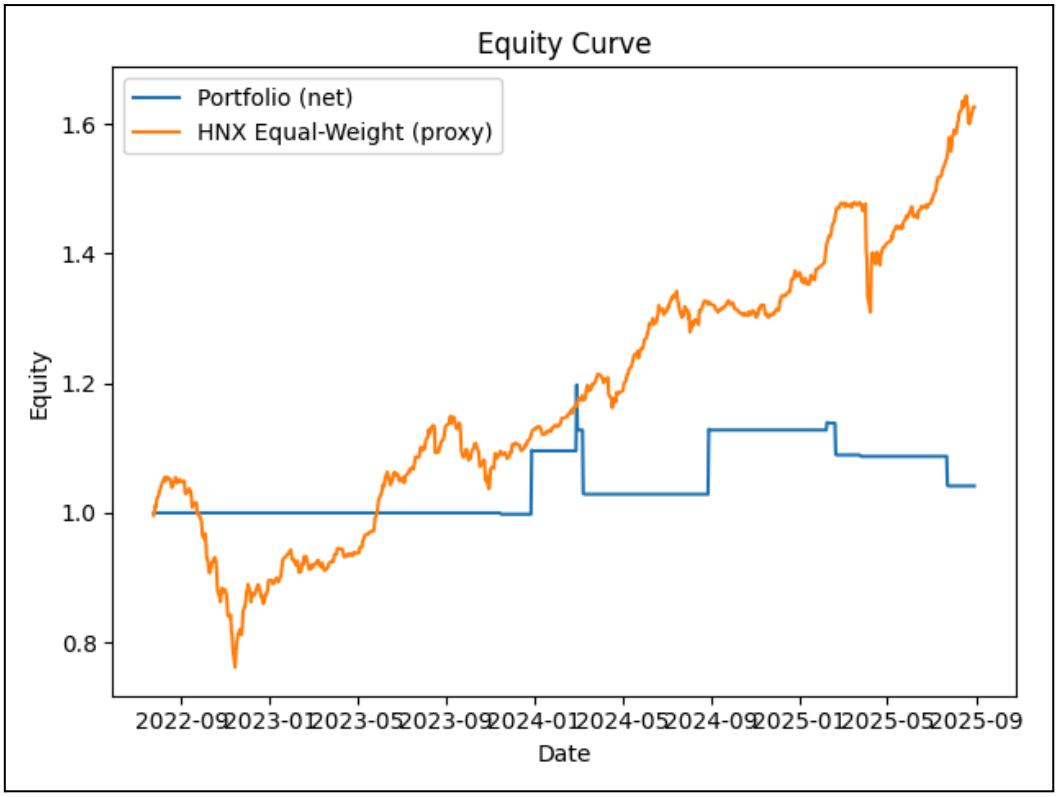
Các bảng biểu, hình ảnh của bước backtest và stress_test được lưu tại folder output_backtest và folder output_stress_test trong github nộp bài



Hình: phân phối lợi nhuận hàng ngày (lợi nhuận ròng)



Hình: Sụt giảm nguồn vốn



Hình: Đường cong lợi nhuận