Data Science Talent Competition 2025

Thông tin chung

Tên đội thi	**Votri**
Tài liệu	**Báo cáo dự thi data science talent competition 2025 - vòng 2**

Thành viên của đội

Họ và tên	Vai trò	Số báo danh	Gmail
Trần Mạnh Hùng	Đội trưởng	B0535	manhhungtran2004@gmail.com
Nguyễn Thế Anh	Thành viên	B0396	theanhnguyen16025@gmail.com
Trương Thị Thùy Dung	Thành viên	B0426	truongthuydung109@gmail.com

Mục lục

1. Logic chiến lược lựa chọn cổ phiếu

Logic cốt lõi của chiến lược không phải là dự đoán hướng đi của giá, mà là **dự báo xác suất thành công của một cơ hội giao dịch** dựa trên một kiến trúc Deep Learning tổng hợp, cụ thể là CNN + LSTM + Multi-Head Attention với đầu vào là chỉ số kĩ thuật (TA) .

1.1. Xây dựng Biến mục tiêu: Phương pháp Triple-Barrier và Chuyển đổi sang Bài toán Phân loại Nhị phân

Một thách thức cố hữu trong việc áp dụng machine learning vào tài chính là việc định nghĩa một biến mục tiêu (target variable) vừa có khả năng dự báo, vừa mang ý nghĩa thực tiễn trong giao dịch. Thay vì tiếp cận bài toán một cách truyền thống là dự báo hướng đi hoặc giá trị của giá, chúng tôi lựa chọn một phương pháp luận cao cấp hơn: **Phương pháp Triple-Barrier (Ba Rào cản)**.

Lý do chính cho lựa chọn này là vì Triple-Barrier cho phép chúng tôi **mô phỏng một kịch bản giao dịch hoàn chỉnh**, bao gồm cả ba yếu tố cốt lõi: **mục tiêu lợi nhuận, quản trị rủi ro, và giới hạn thời gian**. Cụ thể, tại mỗi điểm dữ liệu, chúng tôi thiết lập một giao dịch giả định với ba rào cản động, được điều chỉnh theo biến động thị trường thông qua chỉ số ATR (Average True Range):

- **Rào cản Chốt lời (Take Profit):** Giá vào lệnh + (2 * ATR)
- **Rào cản Cắt lỗ (Stop Loss):** Giá vào lệnh - (1 * ATR)
- **Rào cản Thời gian (Time Limit):** Tối đa 7 phiên giao dịch

Quá trình này tạo ra ba kết quả khả dĩ: **1 (Win)** - khi giá chạm rào cản Chốt lời trước; **0 (Loss)** - khi giá chạm rào cản Cắt lỗ trước; và **2 (Timeout)** - khi không có rào cản nào được chạm đến trong khung thời gian cho phép.

Một bước xử lý quan trọng và có chủ đích trong pipeline của chúng tôi là **chuyển đổi bài toán này thành một bài toán phân loại nhị phân**. Dựa trên logic đã triển khai trong mã nguồn, chúng tôi **loại bỏ tất cả các điểm dữ liệu được gán nhãn là 2 (Timeout)**. Quyết định này mang lại hai lợi ích chiến lược:

1. **Tăng cường sự tập trung của mô hình:** Các trường hợp "Timeout" đại diện cho những giai đoạn thị trường đi ngang, biến động thấp, và không có tín hiệu giao dịch rõ ràng. Việc loại bỏ chúng giúp mô hình không bị phân tâm bởi "nhiễu" và có thể tập trung toàn bộ năng lực học máy vào việc phân biệt các mẫu hình thực sự quan trọng: những mẫu hình dẫn đến một cú đột phá tăng giá (Win) và những mẫu hình dẫn đến một cú sụt giảm (Loss).
2. **Tối ưu hóa cho Tín hiệu có Tính Hành động cao:** Bằng cách chỉ giữ lại hai kết cục mang tính quyết định, chúng tôi buộc mô hình phải trở thành một công cụ chuyên biệt để xác định các cơ hội giao dịch có xác suất thành công cao, thay vì chỉ là một công cụ dự báo chung chung.

Kết quả của phương pháp luận này là một biến mục tiêu nhị phân, rõ ràng và có tính ứng dụng cao, tạo nền tảng vững chắc cho việc huấn luyện một mô hình Deep Learning hiệu quả.

1.2. Kiến trúc Mô hình Hybrid và Luồng Xử lý Thông tin

Để giải quyết bài toán phức tạp này, chúng tôi đã triển khai một kiến trúc học sâu hybrid, lấy cảm hứng từ các nghiên cứu tiên tiến trong lĩnh vực dự báo chuỗi thời gian tài chính. Kiến trúc này, được xây dựng dựa trên mã nguồn, bao gồm hai luồng xử lý song song trước khi hợp nhất để đưa ra dự đoán cuối cùng: một luồng xử lý chuỗi thời gian và một luồng xử lý định danh cổ phiếu.

Luồng Xử lý Chuỗi thời gian (Sequential Branch):

Luồng này nhận đầu vào là một cửa sổ dữ liệu chuỗi thời gian (seq_in), bao gồm các chỉ báo kỹ thuật đã được lựa chọn qua 40 phiên giao dịch gần nhất. Quá trình xử lý diễn ra qua 3 giai đoạn chính:

1. **Giai đoạn 1: Trích xuất Đặc trưng Cục bộ với CNN (Local Feature Extraction)**
    - **Đầu vào:** Cửa sổ dữ liệu chuỗi thời gian (shape: (40, số lượng đặc trưng)).
    - **Nhiệm vụ:** Lớp Conv1D hoạt động như một bộ lọc đặc trưng, trượt dọc theo trục thời gian để nhận diện các mẫu hình cục bộ và ngắn hạn. Ví dụ, nó có thể học cách nhận biết một mẫu hình 3 ngày của RSI kết hợp với sự gia tăng đột biến của khối lượng.
    - **Đầu ra:** Một chuỗi các "feature maps" đã được tinh lọc, biểu diễn các mẫu hình ngắn hạn quan trọng. Chuỗi này sau đó được truyền tới lớp LSTM.
2. **Giai đoạn 2: Mô hình hóa Phụ thuộc Tuần tự với LSTM (Sequential Dependency Modeling)**
    - **Đầu vào:** Chuỗi các feature maps từ lớp CNN.
    - **Nhiệm vụ:** Lớp LSTM tiếp nhận chuỗi các mẫu hình ngắn hạn này và mô hình hóa các phụ thuộc tuần tự giữa chúng. Nó có khả năng "ghi nhớ" và hiểu được "câu chuyện" đang diễn ra theo thời gian, ví dụ như một mẫu hình "tích lũy" (được nhận diện bởi CNN) *theo sau là* một mẫu hình "bùng nổ khối lượng" (cũng được nhận diện bởi CNN) thường dẫn đến một cơ hội giao dịch.
    - **Đầu ra:** Một chuỗi vector trạng thái ẩn, trong đó mỗi vector tại mỗi bước thời gian đã được làm giàu với thông tin về bối cảnh lịch sử của nó.
3. **Giai đoạn 3: Trọng số hóa Thông tin với Multi-Head Attention (Information Weighting)**
    - **Đầu vào:** Chuỗi vector trạng thái ẩn từ lớp LSTM.
    - **Nhiệm vụ:** Đây là cơ chế tinh vi nhất. Lớp Multi-Head Attention thực hiện cơ chế tự chú ý (self-attention) trên chuỗi đầu ra của LSTM, cho phép mô hình tự động đánh giá và gán trọng số cho tầm quan trọng của từng phiên giao dịch trong cửa sổ 40 ngày. Ví dụ, nó có thể học được rằng đối với một tín hiệu sắp xảy ra, thông tin từ 5 ngày trước quan trọng hơn thông tin của ngày hôm qua. Cơ chế này cho phép mô hình tập trung một cách linh hoạt vào các phần liên

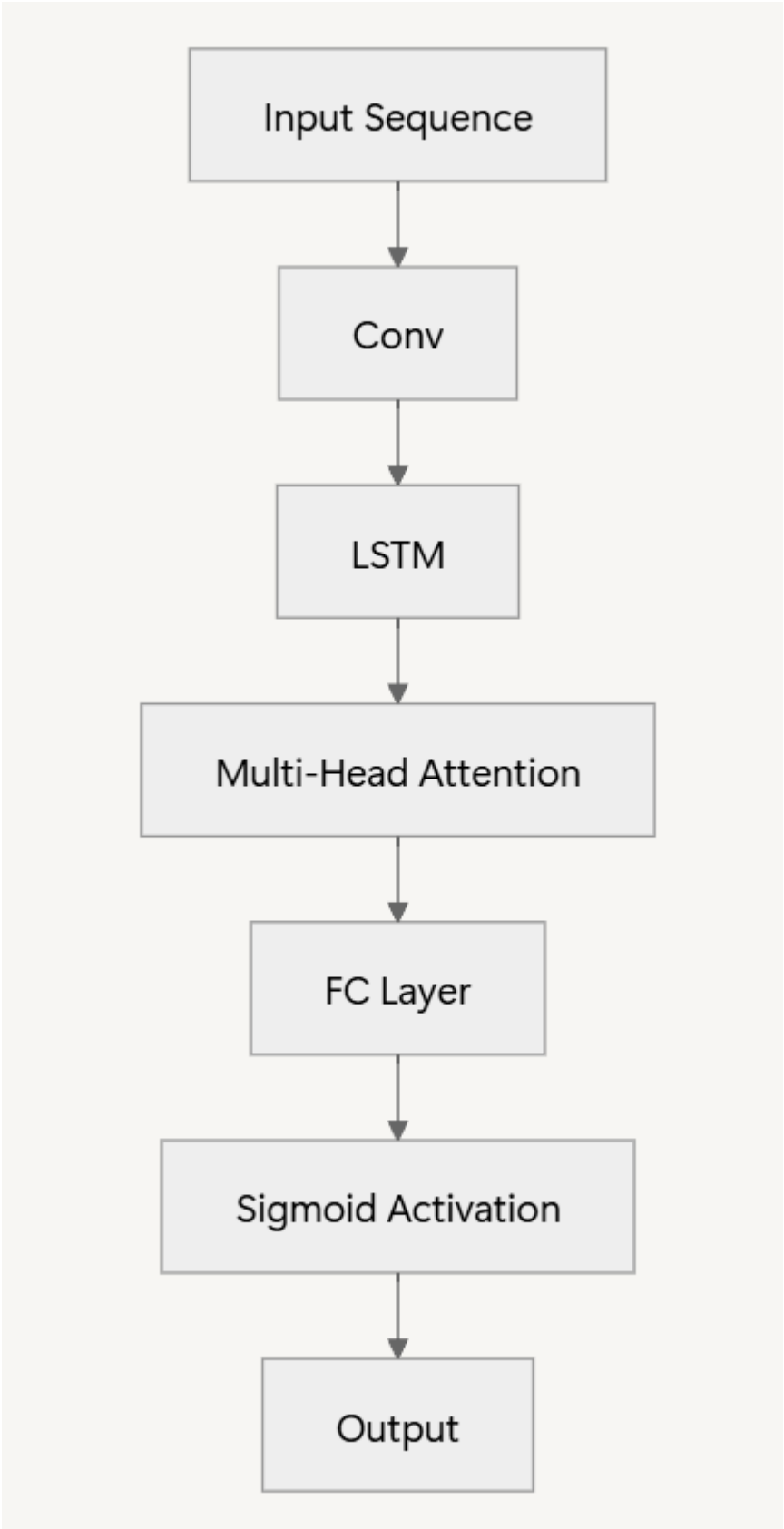
quan nhất của chuỗi đầu vào, một kỹ thuật đã được chứng minh là giúp cải thiện đáng kể độ chính xác dự báo.
    - **Đầu ra:** Một chuỗi vector mới đã được "trọng số hóa", trong đó các thông tin từ các bước thời gian quan trọng được khuếch đại.

Luồng Xử lý Định danh Cổ phiếu (Ticker Identity Branch):

- **Đầu vào:** Mã định danh của cổ phiếu (tick_in).
- **Nhiệm vụ:** Lớp Embedding chuyển đổi mã định danh này thành một vector đặc trưng dày đặc. Trong quá trình huấn luyện, vector này sẽ học cách mã hóa các "đặc tính" riêng, tiềm ẩn của từng cổ phiếu (ví dụ: mức độ biến động trung bình, thuộc nhóm ngành nào, hành vi giá đặc thù).
- **Đầu ra:** Một vector embedding duy nhất đại diện cho bản sắc của cổ phiếu.

Giai đoạn Hợp nhất và Ra quyết định:

1. **Tổng hợp (Aggregation):** Chuỗi vector đã được trọng số hóa từ luồng tuần tự được nén lại thành một vector duy nhất thông qua lớp GlobalAveragePooling1D, tóm tắt toàn bộ thông tin của cửa sổ thời gian.
2. **Kết hợp (Combination):** Vector tóm tắt chuỗi thời gian này được **kết hợp** (concatenate) với vector embedding của cổ phiếu. Bước này cực kỳ quan trọng, vì nó cho phép mô hình đưa ra quyết định dựa trên cả **bối cảnh thị trường gần đây** và **bản chất riêng của cổ phiếu**.
3. **Dự báo (Prediction):** Vector tổng hợp cuối cùng được đưa qua một lớp Dense với hàm kích hoạt sigmoid để tạo ra đầu ra cuối cùng: một giá trị xác suất từ 0 đến 1, biểu thị mức độ tự tin của mô hình rằng cơ hội giao dịch này sẽ "Thành công" (Win).


Input Sequence
Conv
LSTM
Multi-Head Attention
FC Layer
Sigmoid Activation
Output

2. Dữ liệu và chỉ báo được sử dụng

2.1. Nguồn Dữ liệu và Tiền xử lý

2.1.1. Thu thập Dữ liệu

Nền tảng của bất kỳ chiến lược định lượng nào là một bộ dữ liệu chất lượng cao và đáng tin cậy. Trong nghiên cứu này, chúng tôi sử dụng thư viện **FiinQuantX** để truy xuất dữ liệu một cách có hệ thống, đảm bảo tính nhất quán và toàn vẹn.

- **Phạm vi Dữ liệu (Universe):** Chiến lược được xây dựng và kiểm thử trên rổ cổ phiếu **VN30**, bao gồm 30 mã cổ phiếu có giá trị vốn hóa lớn và tính thanh khoản cao nhất thị trường. Việc lựa chọn rổ cổ phiếu này giúp đảm bảo tính đại diện và giảm thiểu rủi ro từ các cổ phiếu có thanh khoản thấp. Danh sách cụ thể bao gồm: ACB, BCM, BID, BVH, CTG, FPT, GAS, GVR, HDB, HPG, và các mã khác trong rổ VN30.
- **Khung thời gian:** Dữ liệu được thu thập trong khoảng thời gian **3 năm**, từ ngày **01/01/2021 đến ngày 01/01/2024**, với tần suất theo ngày (daily).
- **Các trường Dữ liệu Gốc:** Dữ liệu thô cho mỗi cổ phiếu bao gồm các trường cơ bản: timestamp, ticker, open, high, low, close, và volume.

2.1.2. Quy trình Tiền xử lý

Dữ liệu tài chính trong thực tế thường không hoàn hảo và có thể chứa các giá trị bị thiếu (missing values) do các phiên ngừng giao dịch hoặc lỗi dữ liệu. Để đảm bảo chất lượng đầu vào cho mô hình, chúng tôi đã áp dụng một quy trình tiền xử lý gồm hai bước:

1. **Lọc Đặc trưng Ban đầu:** Sau khi các chỉ báo kỹ thuật được tính toán (sẽ được trình bày ở mục sau), bất kỳ cột đặc trưng nào có **tỷ lệ dữ liệu thiếu vượt quá 10%** sẽ bị loại bỏ hoàn toàn. Quy tắc này nhằm tránh việc đưa vào mô hình các đặc trưng không đáng tin cậy hoặc phải nội suy quá nhiều.
2. **Nội suy Dữ liệu (Imputation):** Đối với các giá trị thiếu còn lại (thường là các điểm dữ liệu riêng lẻ), chúng tôi sử dụng phương pháp **nội suy tuyến tính theo từng mã cổ phiếu (per-ticker linear interpolation)**. Việc áp dụng nội suy trên từng mã riêng lẻ là cực kỳ quan trọng, vì nó đảm bảo rằng dữ liệu của một cổ phiếu này không bị rò rỉ (data leakage) để làm đầy cho một cổ phiếu khác, qua đó bảo toàn tính độc lập và toàn vẹn của chuỗi thời gian cho từng tài sản.

2.2. Pipeline Kỹ thuật và Lựa chọn Đặc trưng: Từ Dữ liệu thô đến Tín hiệu Thông minh

Chúng tôi tin rằng hiệu suất của một mô hình học sâu không chỉ đến từ kiến trúc của nó, mà còn phụ thuộc phần lớn vào chất lượng của dữ liệu đầu vào. Vì vậy, chúng tôi đã xây dựng một pipeline 3 giai đoạn chặt chẽ để biến đổi dữ liệu thị trường thô thành một không gian đặc trưng (feature space) tinh gọn, giàu thông tin và tối ưu cho việc dự báo.

Giai đoạn 1: Xây dựng Không gian Đặc trưng Toàn diện với FiinQuantX

Đây là giai đoạn nền tảng, nơi chúng tôi tạo ra một bức tranh đa chiều về hành vi của từng cổ phiếu.

- **Khai thác FiinQuantX:** Để đảm bảo tính chính xác và đồng bộ, toàn bộ các chỉ báo kỹ thuật cơ sở được tính toán thông qua các hàm được tối ưu hóa của thư viện **FiinQuantX**. Như trong hàm compute_indicators của mã nguồn, chúng tôi ưu tiên gọi các hàm như fi.ema(), fi.rsi(), fi.macd(), v.v. Cách tiếp cận này không chỉ giúp tăng tốc độ xử lý mà còn đảm bảo các công thức tính toán tuân thủ theo tiêu chuẩn ngành.
- **Bộ chỉ báo Đa dạng:** Chúng tôi đã tính toán một bộ chỉ báo cực kỳ phong phú (hơn 100 đặc trưng ban đầu) bao phủ mọi khía cạnh của thị trường:
    - **Xu hướng (Trend):** Các loại đường trung bình (SMA, EMA, WMA), MACD với nhiều bộ tham số, ADX, Ichimoku Cloud.
    - **Động lượng (Momentum):** RSI, Stochastic Oscillator với các chu kỳ khác nhau.
    - **Biến động (Volatility):** ATR, Bollinger Bands với nhiều độ lệch chuẩn.
    - **Khối lượng (Volume):** OBV, MFI, VWAP.
    - **Chỉ báo Nâng cao:** Chúng tôi cũng khai thác các chỉ báo phức tạp hơn do FiinQuantX cung cấp như Supertrend, Zigzag, FVG.

Kết quả của giai đoạn này là một không gian đặc trưng ban đầu rất lớn, chứa đựng gần như mọi tín hiệu kỹ thuật có thể có.

Giai đoạn 2: Làm giàu Dữ liệu qua Kỹ thuật Đặc trưng Nâng cao

Các chỉ báo thô chỉ cung cấp một góc nhìn. Để mô hình có thể "hiểu" được bối cảnh, chúng tôi đã tạo ra các đặc trưng bậc cao hơn:

- **Đặc trưng theo Bối cảnh (Contextual Features):** Chúng tôi tính toán **Z-Scores trượt** cho các chỉ báo dao động như RSI, MACD. Thay vì chỉ biết RSI hiện tại là 75, đặc trưng này cho mô hình biết rằng giá trị 75 này cao hơn 2 độ lệch chuẩn so với mức trung bình của 30 ngày qua, tức là một tín hiệu "quá mua" có ý nghĩa thống kê.
- **Đặc trưng Cấu trúc & Tương quan (Structural & Relational Features):** Các tỷ lệ như price_vs_ema20 hay ema5_vs_ema20 được tạo ra. Chúng không đo giá trị tuyệt đối mà đo lường **mối quan hệ và cấu trúc** giữa giá và xu hướng của nó, giúp mô hình học các tín hiệu về sự phân kỳ hoặc hội tụ.
- **Đặc trưng Tương tác & Xác nhận (Interaction & Confirmation Features):** Đây là bước thể hiện sự sáng tạo cao nhất.
    - **Tương tác:** Chúng tôi tạo ra đặc trưng rsi_x_adx14. Đặc trưng này nắm bắt một insight quan trọng: một động lượng RSI cao sẽ có ý nghĩa dự báo mạnh mẽ hơn nhiều khi nó xảy ra trong một bối cảnh có xu hướng rõ ràng (ADX cao).
    - **Xác nhận:** Đặc trưng **bullish_confirmation_score** được tạo ra bằng cách đếm số lượng các điều kiện tăng giá (ví dụ: giá > EMA50, MACD > 0, RSI > 50) xảy ra đồng thời. Nó hoạt động như một bộ lọc nhiễu, hiện thực hóa nguyên tắc

"weight of evidence" (trọng lượng của bằng chứng), nơi sự hội tụ của nhiều tín hiệu yếu tạo thành một tín hiệu tổng hợp mạnh mẽ.

Giai đoạn 3: Quy trình Tinh lọc Đặc trưng Đa tầng

Sau giai đoạn 2, chúng ta có hàng trăm đặc trưng. Việc đưa tất cả vào mô hình sẽ dẫn đến "lời nguyền của không gian nhiều chiều" (curse of dimensionality), gây ra overfitting và lãng phí tài nguyên tính toán. Do đó, một quy trình tinh lọc gồm 2 bước được áp dụng:

1. **Bước 1: Sàng lọc Tín hiệu với Recursive Feature Elimination (RFE)**
    - **Phương pháp:** Chúng tôi sử dụng RFE, một thuật toán sàng lọc mạnh mẽ, với RandomForestClassifier làm mô hình lõi. Random Forest được chọn vì khả năng nắm bắt các mối quan hệ phi tuyến và đánh giá tầm quan trọng của đặc trưng một cách hiệu quả.
    - **Quy trình:** RFE liên tục huấn luyện mô hình và loại bỏ dần các đặc trưng yếu nhất cho đến khi chỉ còn lại **40 đặc trưng** có sức mạnh dự báo cao nhất (N_RFE_FEATURES = 40). Toàn bộ quá trình này được thực hiện **chỉ trên tập huấn luyện** để đảm bảo không có sự rò rỉ thông tin từ tương lai.
2. **Bước 2: Tối ưu hóa bằng cách Loại bỏ Tương quan (Correlation Filtering)**
    - **Vấn đề:** Trong 40 đặc trưng đã chọn, có thể vẫn tồn tại các cặp cung cấp thông tin trùng lặp (ví dụ: EMA10 và EMA12). Hiện tượng này, gọi là **đa cộng tuyến (multicollinearity)**, có thể làm giảm sự ổn định của mô hình.
    - **Giải pháp:** Chúng tôi tính toán ma trận tương quan giữa 40 đặc trưng và tự động loại bỏ một trong hai đặc trưng nếu chúng có độ tương quan vượt ngưỡng **0.8**.
    - **Kết quả:** Từ bộ đặc trưng đã được sàng lọc, chúng tôi chọn ra **20 đặc trưng cuối cùng (FINAL_FEATURE_COUNT = 20)** có thứ hạng cao nhất từ RFE.

Bộ 20 đặc trưng cuối cùng này là kết quả của một quá trình chắt lọc nghiêm ngặt, đảm bảo rằng mỗi đặc trưng đưa vào mô hình Deep Learning đều mang lại một góc nhìn độc đáo, có sức mạnh dự báo cao và không bị nhiễu, tạo ra nền tảng vững chắc nhất cho mô hình học tập.

3. Kết quả backtest (biểu đồ, bảng, hiệu suất)

3.1. Huấn luyện Mô hình

Dữ liệu được phân chia theo trình tự thời gian thành 3 tập: **Training (6940 mẫu), Validation (1487 mẫu), và Test (1488 mẫu)**. Các cơ chế EarlyStopping và Class Weight (trọng số 3.85 cho lớp 'Win') được áp dụng trong quá trình huấn luyện để chống overfitting và xử lý mất cân bằng dữ liệu.

3.2. Đánh giá Hiệu suất Phân loại

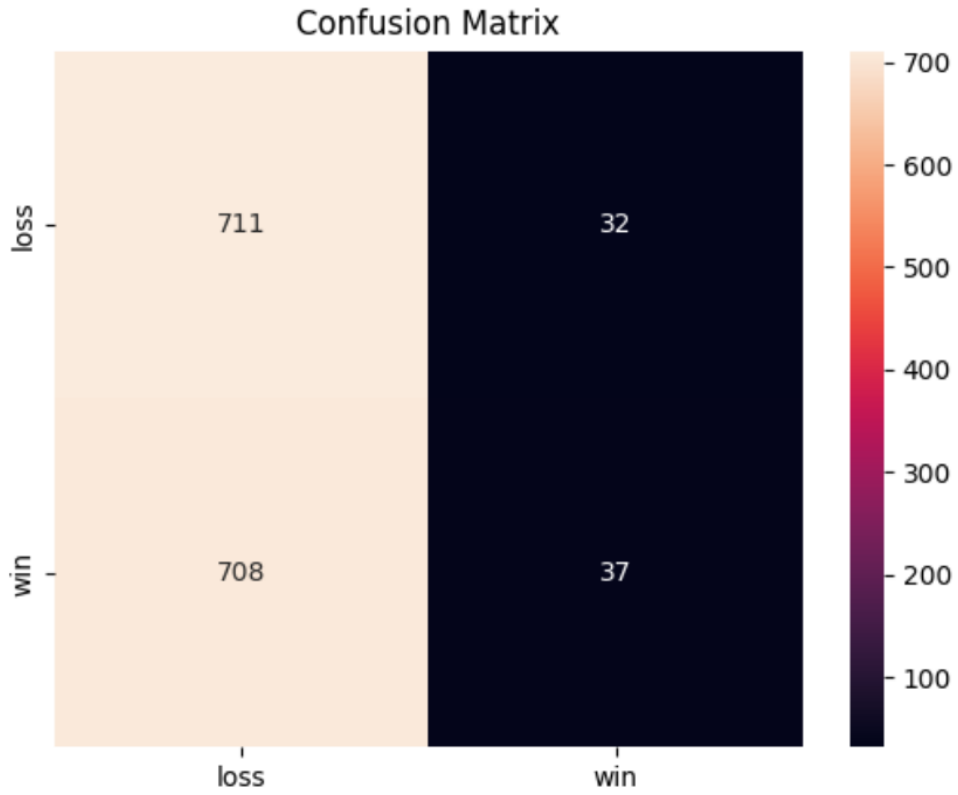
Một ngưỡng quyết định là 0.75 được áp dụng lên xác suất đầu ra của mô hình để tạo tín hiệu. Kết quả trên tập Test được trình bày dưới đây.

Bảng 1: Báo cáo Phân loại (Classification Report)

	precision	recall	f1-score	support
loss	0.50	0.96	0.66	743
win	0.54	0.05	0.09	745
			accuracy	0.50 (1488)
macro avg	0.52	0.50	0.37	1488
weighted avg	0.52	0.50	0.37	1488

Bảng 2: Ma trận Nhầm lẫn (Confusion Matrix)

	Predicted Loss	Predicted Win
Actual Loss	711	32
Actual Win	708	37



Việc áp dụng ngưỡng xác suất cao (0.75) dẫn đến **Recall (0.05)** thấp, cho thấy chiến lược được thiết kế để có tính chọn lọc cao, ưu tiên chất lượng tín hiệu hơn số lượng. Chỉ số **Precision (0.54)** cho thấy 54% các tín hiệu "Mua" được tạo ra là chính xác. Ma trận nhầm lẫn định lượng hành vi này, với số lượng False Positives (32) thấp và False Negatives (708) cao.

3.3. Tín hiệu Giao dịch và Kết quả Backtest

Để đánh giá giá trị thực tiễn của các tín hiệu do mô hình tạo ra, chúng tôi đã tiến hành kiểm thử lại (backtest) trên tập dữ liệu ngoài mẫu (out-of-sample test set). Quá trình backtest được thiết kế theo phương pháp event-driven (dựa trên sự kiện), trong đó một giao dịch chỉ được thực hiện khi mô hình phát ra tín hiệu "Mua" (xác suất > 0.75).

Hai chiến lược backtest đơn giản đã được triển khai để đo lường hiệu quả của mô hình ở hai khung thời gian khác nhau:

1. Chiến lược Bám theo Xu hướng Hàng ngày (Daily Trend-Following)

- **Logic:** Chiến lược này mô phỏng việc nắm bắt xu hướng giá từ cuối ngày hôm trước đến cuối ngày hôm sau. Khi có tín hiệu "Mua" cho ngày T, một vị thế mua được giả

định mở tại mức **giá đóng cửa của ngày T-1** (prev_test) và đóng tại mức **giá đóng cửa của ngày T** (closes_test). Lợi nhuận của mỗi giao dịch được tính bằng closes_test - prev_test.

- **Mục đích:** Đánh giá khả năng của mô hình trong việc dự đoán xu hướng giá **liên phiên (interday)**.
- **Kết quả:**
    - **Tổng lợi nhuận cộng dồn: 13,032.56** (đơn vị giả định)

2. Chiến lược Giao dịch trong Ngày (Intraday Trading)

- **Logic:** Chiến lược này tập trung vào việc khai thác biến động giá trong cùng một phiên giao dịch. Khi có tín hiệu "Mua" cho ngày T, một vị thế mua được giả định mở tại mức **giá mở cửa của ngày T** (opens_test) và đóng tại mức **giá đóng cửa của cùng ngày T** (closes_test). Lợi nhuận của mỗi giao dịch được tính bằng closes_test - opens_test.
- **Mục đích:** Đánh giá khả năng của mô hình trong việc dự đoán các chuyển động giá có lợi nhuận **trong phiên (intraday)**.
- **Kết quả:**
    - **Tổng lợi nhuận cộng dồn: 6,764.65** (đơn vị giả định)

Diễn giải và Hạn chế:

Các kết quả backtest trên đều cho thấy tổng lợi nhuận cộng dồn dương, cung cấp bằng chứng cho thấy các tín hiệu do mô hình tạo ra có **kỳ vọng thống kê dương (positive statistical expectancy)**. Mô hình đã thể hiện khả năng xác định các mẫu hình mà sau đó, trung bình, giá có xu hướng tăng cả trong phiên và giữa các phiên.

Cần lưu ý rằng các backtest này được thực hiện trong điều kiện đơn giản hóa và **chưa tính đến các yếu tố ma sát của thị trường** như phí giao dịch, thuế, trượt giá (slippage) hay tác động thị trường của lệnh lớn.

4. Phân tích khả năng mở rộng và hạn chế của chiến lược

4.1. Khả năng Mở rộng và Điểm mạnh

- **Kiến trúc Mô hình Hiện đại:** Sự kết hợp CNN, LSTM và Attention giúp mô hình có khả năng học các mẫu hình cực kỳ phức tạp mà các mô hình truyền thống không thể nắm bắt.
- **Thích ứng với Từng Cổ phiếu:** Lớp **Ticker Embedding** cho phép mô hình học các "đặc tính" riêng của từng mã cổ phiếu, giúp chiến lược dễ dàng mở rộng ra toàn bộ thị trường mà không cần điều chỉnh thủ công.
- **Hướng Mở rộng với Dữ liệu Thay thế:** Để đối phó với các sự kiện bất ngờ ("Thiên Nga Đen"), chiến lược có thể được mở rộng bằng cách tích hợp một mô hình **FinBERT** để

phân tích cảm xúc từ tin tức tài chính. Điều này sẽ bổ sung một lớp thông tin định tính, giúp hệ thống phản ứng nhanh hơn với các sự kiện ngoại sinh.
- **Tối ưu hóa cơ chế chú ý (Attention Mechanism):** Tối ưu hóa hơn nữa cơ chế chú ý có thể tăng cường khả năng của mô hình trong việc nắm bắt các tín hiệu thị trường

4.2. Hạn chế và Rủi ro

- **Phụ thuộc vào Dữ liệu Lịch sử:** Chiến lược hoạt động dựa trên giả định rằng các mẫu hình trong quá khứ sẽ lặp lại. Các thay đổi cấu trúc đột ngột trên thị trường có thể làm giảm hiệu quả của mô hình.
- **Tính chất "Hộp đen":** Việc diễn giải đầy đủ lý do đằng sau một quyết định cụ thể của mô hình Deep Learning vẫn còn là một thách thức, đòi hỏi các kỹ thuật XAI (Explainable AI) phức tạp hơn.
- **Chi phí Giao dịch:** Backtest hiện tại chưa mô phỏng chi phí giao dịch và trượt giá, điều này có thể ảnh hưởng đến lợi nhuận thực tế.

5. Insight rút ra từ mô hình

Quá trình xây dựng và đánh giá mô hình không chỉ mang lại một chiến lược giao dịch tiềm năng mà còn cung cấp những insight sâu sắc về bản chất của thị trường và các yếu tố dẫn đến một cơ hội giao dịch thành công.

5.1. Tầm quan trọng của Việc Định nghĩa Bài toán

Việc chuyển đổi bài toán từ "dự đoán giá" sang **"dự đoán xác suất thành công của một giao dịch"** thông qua phương pháp Triple-Barrier là insight nền tảng quan trọng nhất. Nó buộc mô hình phải học trong một khuôn khổ thực tế, có ràng buộc về lợi nhuận kỳ vọng (2R), rủi ro chấp nhận (1R) và thời gian. Điều này giúp tạo ra các tín hiệu có tính ứng dụng cao, thay vì các dự báo giá trị đơn thuần có thể không đi kèm với một kế hoạch hành động rõ ràng.

5.2. Sức mạnh của Sự hội tụ và Bối cảnh Tín hiệu

Phân tích bộ 20 đặc trưng cuối cùng được lựa chọn bởi quy trình RFE và lọc tương quan cho thấy một insight rõ ràng: mô hình không dựa vào một chỉ báo "thần thánh" nào cả. Thay vào đó, nó học cách nhận diện sự thành công dựa trên sự kết hợp và tương tác của nhiều nhóm tín hiệu:

- **Động lượng và Tốc độ thay đổi (Momentum & RoC):** Các đặc trưng như rsi_14_zscore, macd_diff_*, ema_5_roc_10 chiếm ưu thế, cho thấy mô hình ưu tiên các tín hiệu về sự "thay đổi" và "gia tốc" của giá hơn là mức giá tuyệt đối.
- **Biến động và Sức mạnh Xu hướng (Volatility & Trend Strength):** Các đặc trưng như atr_7, volatility_60d, và đặc biệt là adx_10 và rsi_x_adx14 cho thấy mô hình đã học được rằng: một động lượng mạnh chỉ đáng tin cậy khi được xác nhận bởi một xu

hướng rõ ràng và một mức biến động đủ lớn để giá có thể chạm đến mục tiêu chốt lời.
- **Xác nhận từ Khối lượng:** Sự hiện diện của volume và obv trong bộ đặc trưng cuối cùng khẳng định rằng mô hình xem khối lượng là một yếu tố xác nhận quan trọng cho sự bền vững của một chuyển động giá.

5.3. Chiến lược Ưu tiên Độ chính xác (High-Precision Strategy)

Việc thiết lập một ngưỡng quyết định cao (0.75) đã định hình chiến lược theo hướng **"chất lượng hơn số lượng"**. Kết quả Precision cao hơn Recall cho thấy mô hình được tối ưu hóa để giảm thiểu rủi ro từ các tín hiệu sai (False Positives). Đây là một insight quan trọng về quản trị rủi ro: trong một môi trường nhiễu như thị trường tài chính, việc kiên nhẫn chờ đợi một tín hiệu có độ tin cậy cao sẽ bền vững hơn là giao dịch liên tục dựa trên các tín hiệu có xác suất thấp.

5.4. Không có "Một Công thức" cho Toàn thị trường

Hai quyết định kỹ thuật quan trọng trong mã nguồn đã tiết lộ insight này:

- **Per-Ticker Scaling:** Việc chuẩn hóa dữ liệu theo từng mã cổ phiếu riêng lẻ (thay vì trên toàn bộ tập dữ liệu) cho thấy mô hình nhận thức được rằng mỗi cổ phiếu có một "tính cách" thống kê riêng (mức biến động, biên độ dao động của chỉ báo...). Điều này giúp mô hình so sánh một tín hiệu với chính lịch sử của cổ phiếu đó, làm cho tín hiệu trở nên có ý nghĩa hơn.
- **Ticker Embedding:** Lớp Embedding trong kiến trúc mô hình cho phép nó học một vector đại diện cho "bản sắc" của từng cổ phiếu. Điều này có nghĩa là mô hình không chỉ học từ các chỉ báo kỹ thuật, mà còn học được các yếu tố tiềm ẩn như cổ phiếu đó thuộc ngành nào, mức độ được thị trường quan tâm ra sao. Insight ở đây là: cùng một bộ tín hiệu kỹ thuật có thể dẫn đến kết quả khác nhau đối với FPT (công nghệ) và VNM (tiêu dùng), và mô hình đã học được cách phân biệt điều đó.