Tuần lễ sinh viên 5 tốt lần thứ V

trường đại học kinh tế quốc dân

Cuộc thi: The EconomAIst

Báo cáo Bán kết

Đề tài: Phân tích tài chính

Đội thi: F4DT

Sinh viên thực hiện:

Nguyễn Văn Tuấn

Vũ Văn Huy

Nguyễn Đình Tuấn Long

Vũ Tiến Dũng

Ngày 9 tháng 9 năm 2024

Lời mở đầu và Lời cảm ơn

Điều đầu tiên, nhóm F4DT xin cảm ơn Ban Phong trào Sinh viên 5 tốt - Trường Đại học Kinh tế Quốc dân cùng các đơn vị đồng tổ chức cuộc thi THE ECONOMAIST. Đây là một cuộc thi về Ứng dụng Big Data & AI trong kinh tế số. Đội F4DT với tất cả thành viên đến từ Đại học Bách khoa Hà Nội, với giáo án giảng dạy của trường thì các thành viên trong đội đã quá quen với Big Data cũng như AI, tuy nhiên việc ứng dụng Dữ liệu lớn và Trí tuệ nhân tạo trong kinh tế số thì có vẻ chưa thật sự quen thuộc với F4DT. Nhưng nhờ cuộc thi, nhóm F4DT đã có những góc nhìn và trải nghiệm trong việc sử dụng những kiến thức của mình đã có như Phân tích dữ liệu cũng như xây dựng mô hình Trí tuệ nhân tạo ứng dụng vào việc dự đoán giá cổ phiếu. Xin chúc Ban Phong trào Sinh viên 5 tốt - Trường Đại học Kinh tế Quốc dân cùng các đơn vị đồng tổ chức cuộc thi sẽ luôn phát triển mạnh mẽ và tạo ra các cuộc thi cho các sinh viên trên địa bàn có thể thử sức và tranh tài trong các cuộc thi có tính quy mô và chuyên môn cao. Dự đoán giá cổ phiếu là một trong những bài toán phức tạp và hấp dẫn nhất trong tài chính. Giá cổ phiếu chịu ảnh hưởng bởi nhiều yếu tố như tình hình kinh tế vĩ mô, sự thay đổi trong cơ cấu doanh nghiệp, và các yếu tố tâm lý thị trường. Sự biến động lớn và không chắc chắn của thị trường chứng khoán đã thúc đẩy sự phát triển của các công cụ dự đoán nhằm đưa ra các quyết định đầu tư chính xác hơn.

Trân trọng cảm ơn !

Nhóm F4DT

Mục lục

1 Giới thiệu chung

1.1 Ứng dụng của công nghệ phân tích dữ liệu và AI trong dự đoán giá cổ phiếu

- **Trí tuệ nhân tạo (AI)** đã trở thành một công cụ thiết yếu trong lĩnh vực tài chính, đặc biệt là dự đoán giá cổ phiếu. AI có thể xử lý và phân tích lượng dữ liệu khổng lồ từ nhiều nguồn khác nhau như dữ liệu lịch sử giá, khối lượng giao dịch, các chỉ số kinh tế, và thậm chí các thông tin phi cấu trúc như tin tức và tâm lý thị trường.

- Các mô hình AI dựa trên **học máy (Machine Learning)** và **học sâu (Deep Learning)** như **mạng nơ-ron hồi quy LSTM** và **GRU** được sử dụng rộng rãi để xử lý dữ liệu chuỗi thời gian. Những mô hình này có khả năng phát hiện các mẫu ẩn và xu hướng trong dữ liệu mà các phương pháp truyền thống có thể bỏ sót.

1.2 Sự phát triển của công nghệ GenAI trong dự đoán giá cổ phiếu

- **Generative AI (GenAI)** là một bước tiến mới trong lĩnh vực AI, với khả năng tạo ra dữ liệu và mô phỏng các tình huống khác nhau dựa trên dữ liệu học được từ quá khứ. Trong bài toán dự đoán giá cổ phiếu, GenAI không chỉ học từ các mẫu lịch sử mà còn có thể tạo ra các dự báo cho những kịch bản tương lai tiềm năng.

- Một trong những ứng dụng nổi bật của GenAI là khả năng phân tích các yếu tố ngoại vi như tâm lý thị trường thông qua các bài viết trên mạng xã hội, tin tức và các báo cáo tài chính. Các mô hình GenAI như **GPT** có thể tự động phân tích nội dung phi cấu trúc để đưa ra các dự đoán dựa trên các yếu tố không định lượng như tin đồn hoặc các sự kiện kinh tế, chính trị.

- GenAI cũng có khả năng phân tích và mô phỏng các chiến lược đầu tư khác nhau, cho phép các nhà đầu tư kiểm tra nhiều kịch bản và lựa chọn chiến lược tối ưu

nhất trong các điều kiện thị trường khác nhau.

2 Kĩ thuật dự đoán giá cổ phiếu

2.1 Định nghĩa 1 vài chỉ số trước khi phân tích

Thay đổi phần trăm hàng ngày:

$$r_t = \frac{p_t}{p_{t-1}} - 1$$

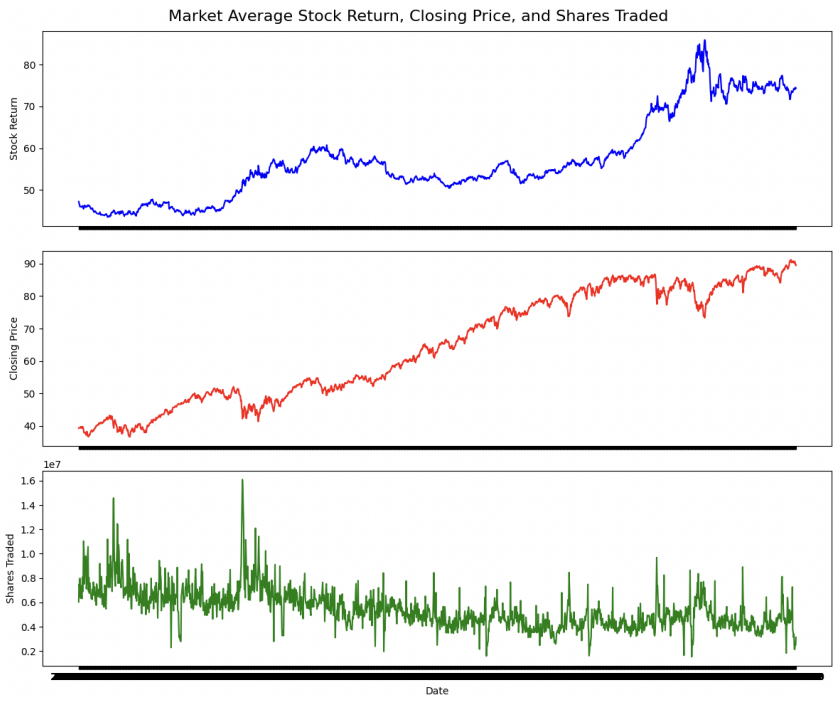
- r_t là lợi nhuận tại thời điểm t
- p_t là giá tại thời điểm t

Lợi nhuận tích lũy

$$i_t = (1 + r_t)i_{t-1} = \left(1 + \frac{p_t}{p_{t-1}} - 1\right) i_{t-1} = \frac{p_t}{p_{t-1}} i_{t-1}$$

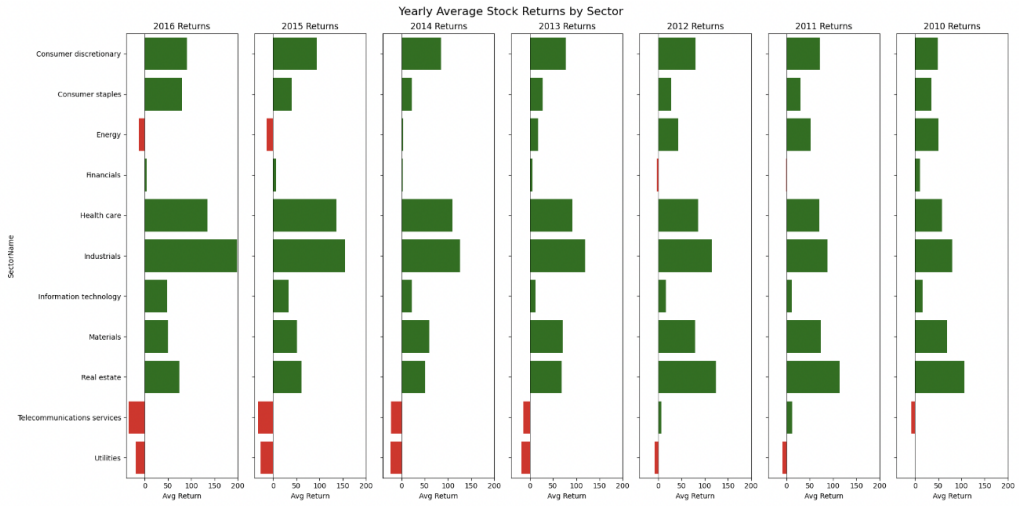
- r_t là % thay đổi hằng ngày ở công thức trên
- i_t là ngày thứ t

2.2 Đánh giá thị trường chung



Nhận xét: Có sự tăng trưởng vững chắc về lợi nhuận và giá cổ phiếu từ năm 2012 đến 2015, sau đó giảm nhẹ và ổn định hơn. Khối lượng giao dịch có xu hướng biến động mạnh vào các thời điểm điều chỉnh lớn trên thị trường.

2.3 Lợi nhuận cổ phiếu trung bình hàng năm theo ngành



Nhận xét: Các màu sắc từ đỏ (lợi nhuận âm) đến xanh lá (lợi nhuận dương) cho thấy rõ sự phân cực giữa các ngành có lợi nhuận tích cực và tiêu cực. Những ngành như Công nghiệp, Y tế và Bất động sản luôn duy trì màu sắc xanh, cho thấy khả năng sinh lời cao, trong khi các ngành như Năng lượng và Dịch vụ viễn thông luôn ở mức âm hoặc rất thấp.

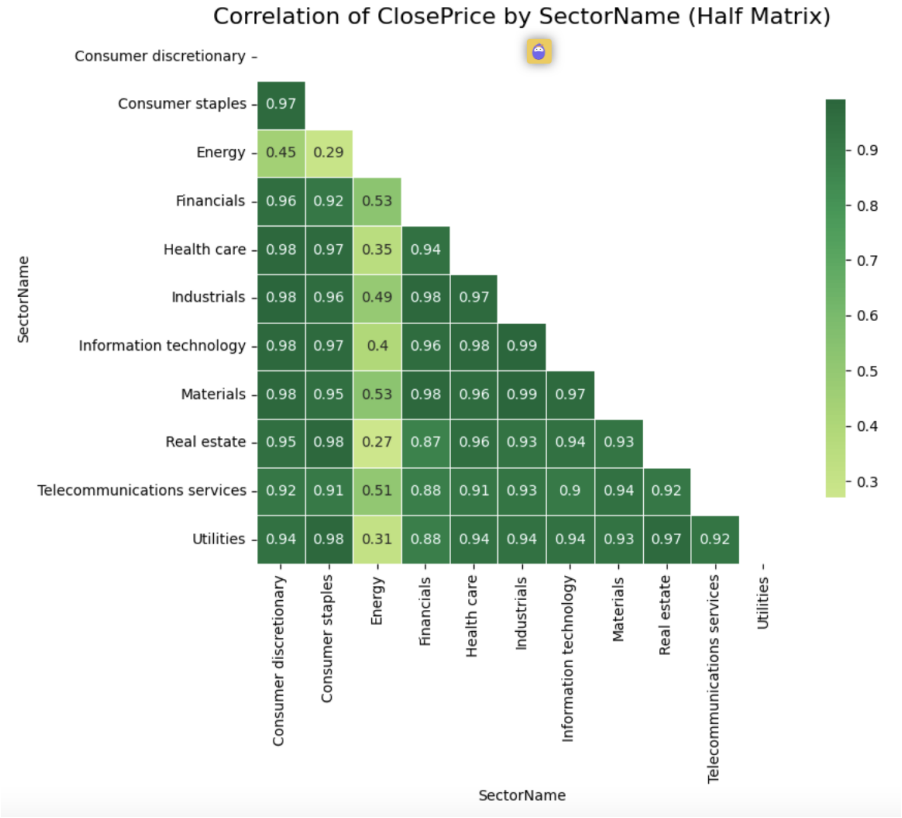
Xu hướng tăng trưởng theo từng năm:

- **Industrials (Công nghiệp)** duy trì lợi nhuận cao nhất so với các ngành khác qua toàn bộ giai đoạn, đặc biệt là năm 2016, khi lợi nhuận đã vượt mốc 200.
- **Health care (Y tế)** cũng có xu hướng tăng trưởng mạnh mẽ từ năm 2010 đến 2016, với mức lợi nhuận vượt trội, đặc biệt là từ 2012 trở đi, khi lợi nhuận của ngành này bắt đầu tăng đều. ⇒ Nên đầu tư vào 2 ngành này.

Nhóm ngành ổn định: **Real estate (Bất động sản)** và **Materials (Vật liệu)** có xu hướng ổn định với lợi nhuận tăng trưởng đều và không bị ảnh hưởng nhiều qua các năm. Đây có thể là những ngành có tiềm năng an toàn cho các nhà đầu tư tìm kiếm lợi

nhuận ổn định.

2.4 Phân tích tương quan giá đóng cửa theo ngành

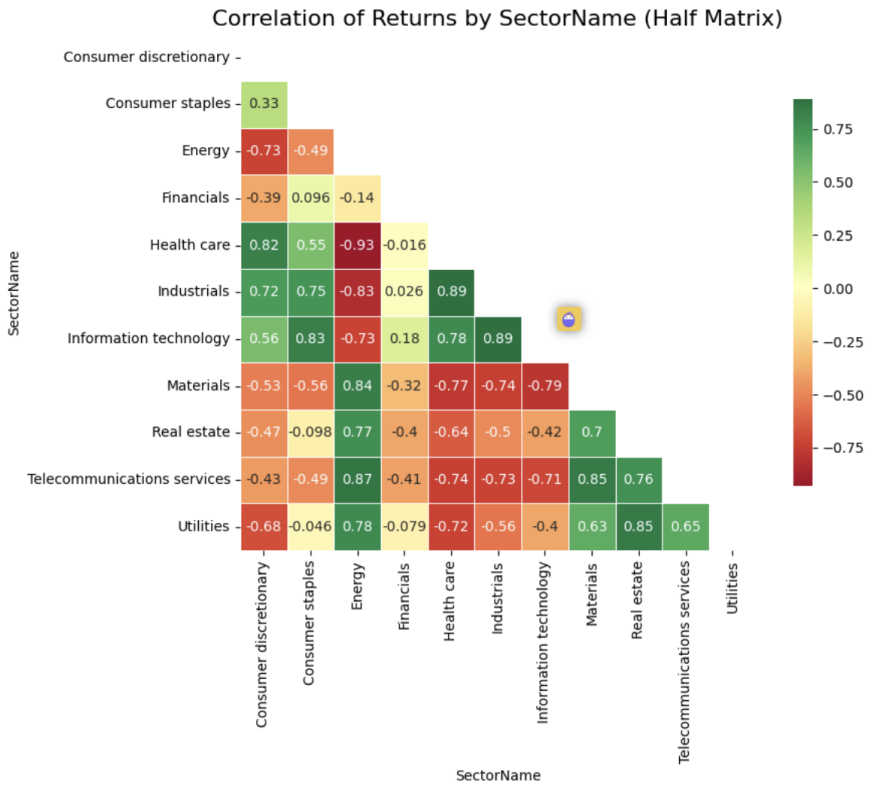


1. Tương quan cao giữa các ngành:

- Hầu hết các ngành có mối tương quan mạnh (trên 0.9), đặc biệt giữa Consumer discretionary, Industrials, Information technology, Materials, và Utilities.
- Điều này cho thấy các yếu tố chung của thị trường, như điều kiện kinh tế vĩ mô, ảnh hưởng đồng bộ đến giá cổ phiếu của các ngành này.

2. Ngành Energy có tương quan thấp:

- Ngành Energy có mối tương quan thấp với các ngành khác, đặc biệt với Health care (0.35), Industrials (0.49), và Materials (0.53).
- Giá cổ phiếu ngành năng lượng bị ảnh hưởng bởi các yếu tố đặc thù như giá dầu, nguồn cung năng lượng, và chính trị, hơn là các yếu tố kinh tế chung.



2.5 Phân tích tương quan biến động cổ phiếu theo ngành

1. Tương quan cao (dương):

- Consumer staples và Information technology: tương quan cao 0.83.
- Real estate và Telecommunications services: tương quan 0.85, lợi nhuận biến động cùng chiều.
- Industrials và Information technology: tương quan 0.78, phản ứng tương tự với thị trường.

2. Tương quan âm (ngược chiều):

- Energy và Health care: tương quan âm mạnh -0.93, do khác biệt yếu tố tác động (giá dầu, chính sách y tế).
- Energy và Information technology: tương quan -0.73; Consumer discretionary và Utilities: tương quan -0.68.

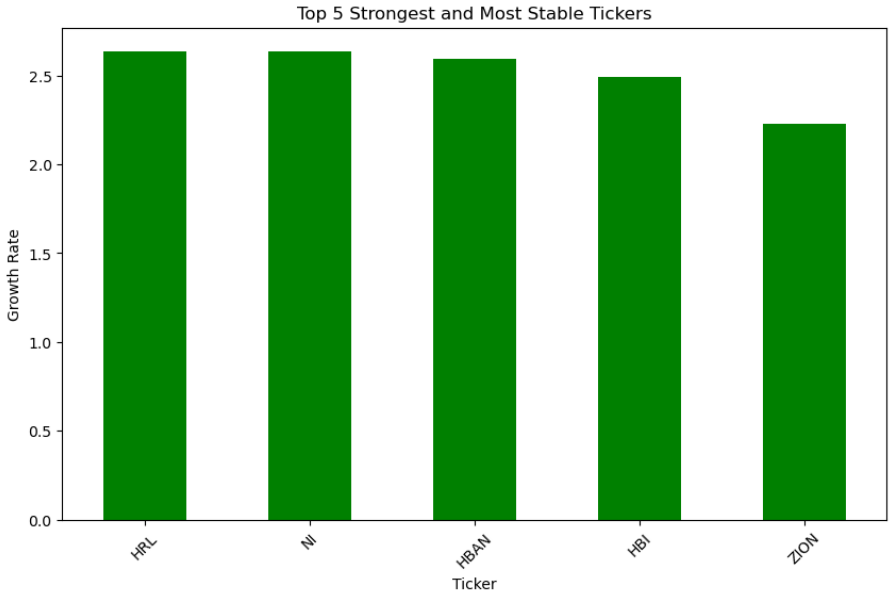
3. Ngành có mối tương quan thấp với nhiều ngành khác:

- Energy có nhiều mối tương quan âm, đặc biệt với Health care (-0.93), Information technology (-0.73), và Materials (-0.77).
- Utilities có tương quan âm với Consumer discretionary (-0.68) và Industrials (-0.72), do sự ổn định cao và ít nhạy cảm với biến động kinh tế.

2.6 Phân tích và chọn mã chứng khoán tiềm năng

Chọn mã chứng khoán tiềm năng Tập dữ liệu ban đầu có quá nhiều công ty, nhóm ngành mỗi công ty đều có những đặc điểm, chỉ số riêng nên ta sẽ chỉ chọn ra công ty theo những tiêu chí cụ thể để phân tích (nếu phân tích hết sẽ rất loạn và rối). Ở đây chúng tôi sẽ lựa chọn phân tích các công ty dựa trên yếu tố: Sự phát triển mạnh mẽ và tính ổn định.

Dữ liệu phân tích từ file financial_metrics.csv



Hình 1: Top 5 công ty

Bỏ qua công ty NI do thiếu dữ liệu các chỉ số tài chính từ file các chỉ số tài chính. Tiến hành tổng hợp các chỉ số quan trọng mang tính đặc trưng của 4 công ty còn lại.

HRL: Doanh thu: 9.2 tỷ, CapEx: -173.8 triệu, Vốn: 3.79 tỷ, Nợ: 1.79 tỷ, ROE: 17.67%,

EPS: 1.58

HBAN: Doanh thu: 3.0 tỷ, CapEx: -95.95 triệu, Vốn: 6.2 tỷ, Nợ: 57.04 tỷ, ROE: 10.75%, EPS: 0.75

ZION: Doanh thu: 2.33 tỷ, CapEx: -122.66 triệu, Vốn: 6.85 tỷ, Nợ: 50.26 tỷ, ROE: 4.75%, EPS: 1.36

HBI: Doanh thu: 5.23 tỷ, CapEx: -69.10 triệu, Vốn: 1.3 tỷ, Nợ: 3.67 tỷ, ROE: 30%, EPS: 2.80

Hình 2: Trực quan các metric

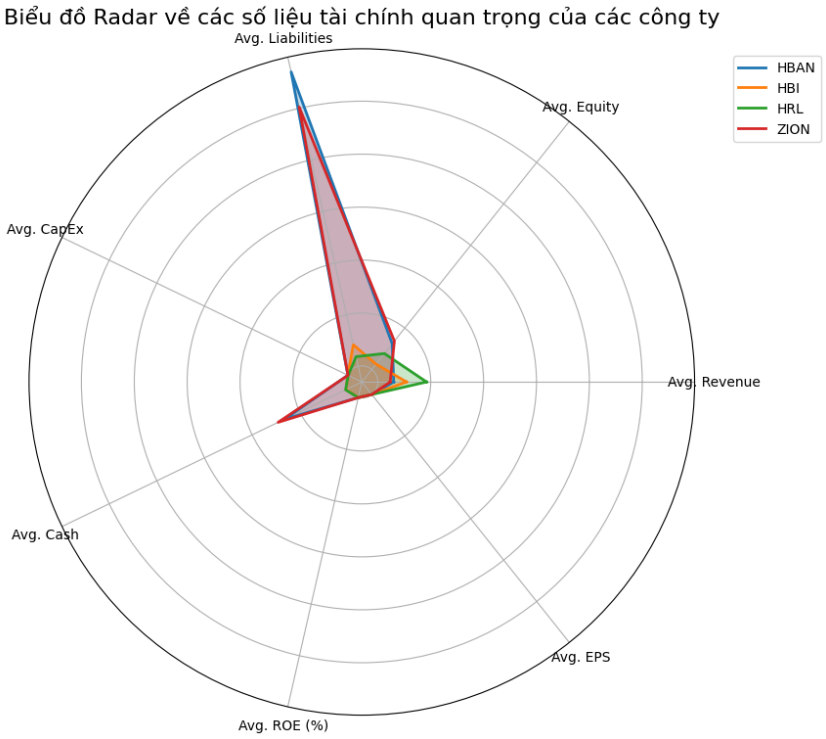
Nhận xét:

1. HRL: Lợi thế: Doanh thu cao, biên lợi nhuận ổn định, quản lý nợ tốt. Đầu tư: An toàn, tiềm năng dài hạn nhờ lợi nhuận và quản lý tài chính vững chắc.

2. HBI: Lợi thế: Doanh thu ổn định, khả năng sinh lời khá. Rủi ro: Dòng tiền đầu tư âm lớn, cần theo dõi khả năng sinh lời từ các khoản đầu tư.

3. HBAN: Lợi thế: Dòng tiền ròng tích cực. Rủi ro: Nợ cao, biên lợi nhuận thấp, phụ thuộc nhiều vào vốn vay.

4. ZION: Lợi thế: Dòng tiền ổn định. Rủi ro: Nợ cao nhất, biên lợi nhuận biến động, tiềm ẩn rủi ro tài chính.



Hình 3: Biểu đồ radar về trung bình các chỉ số

HRL và HBI nổi bật về tài chính vững chắc, trong khi HBAN và ZION cần cải thiện quản lý nợ và hiệu quả sử dụng vốn.

2.7 Các thuật toán sử dụng

2.7.1 Thuật toán Voting Regressor

Bước 1: Chuẩn bị dữ liệu

- Tính toán các chỉ báo: ROC, RSI, MACD, CCI, OBV.
- Sử dụng GridSearch để tối ưu tham số.

Bước 2: Huấn luyện Voting Regressor

- Kết hợp 3 mô hình: RandomForest, LinearRegression, KNeighborsRegressor.
- **Voting không trọng số:**

$$\hat{y} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i$$

- **Voting có trọng số:**

$$\hat{y} = \frac{\sum_{i=1}^{n} w_i \hat{y}_i}{\sum_{i=1}^{n} w_i}$$

Bước 4: Dự đoán và Đánh giá mô hình

- Tính Overall Score từ các chỉ số đã chuẩn hóa:

$$Overall\ Score = (weight_{R^2} \times R^2_{normalized}) + \dots$$

Ticker	Best params	R^2	MSE	RMSE	MAE	MAPE	SMAPE	Overall Score
HRL	{'roc_period': 12, 'rsi_period': 30, 'macd_fast': 12, 'macd_slow': 26, 'macd_signal': 9}	0.946462	4.423544	2.103222	1.505267	6.644377	6.713307	0.963516
HBAN	{'roc_period': 20, 'rsi_period': 10, 'macd_fast': 12, 'macd_slow': 26, 'macd_signal': 9}	0.689485	1.374797	1.172517	0.769363	10.051904	9.271435	0.911951
HBI	{'roc_period': 14, 'rsi_period': 10, 'macd_fast': 26, 'macd_slow': 50, 'macd_signal': 9}	0.967750	3.091938	1.758391	1.323582	8.906190	8.931538	0.965081

2.7.2 Thuật toán ARIMA

Bước 1: Chuẩn bị dữ liệu

Bước 2: Huấn luyện Auto ARIMA

- Sử dụng `auto_arima` để tìm tham số tối ưu, tìm các tham số p, d, q tối ưu.

Bước 3: Huấn luyện ARIMA theo Walk-forward

Ticker	Best ARIMA Order	MSE	RMSE	MAE	MAPE	SMAPE	R^2	Overall Score
HRL	(2, 1, 0)	0.339230	0.582435	0.418234	1.097712	1.095638	0.958429	0.986065
HBAN	(2, 1, 2)	0.037996	0.194925	0.140081	1.412925	1.412071	0.975684	0.991166
HBI	(3, 1, 2)	0.301924	0.549476	0.370653	1.384472	1.379178	0.931581	0.980415

- Huấn luyện mô hình ARIMA dựa trên tham số tối ưu.

- **Công thức ARIMA**:

$$ARIMA(p,d,q) : y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

trong đó y_t là giá trị dự đoán, ϕ là hệ số tự hồi quy, và θ là hệ số MA.

2.7.3 Thuật toán GRU

Bước 1: Chuẩn bị dữ liệu

- Chuẩn hóa dữ liệu với Min-Max scaling, chia dữ liệu thành các tập huấn luyện, kiểm tra và xác thực. Dữ liệu đầu vào là chuỗi giá mở, cao, thấp và đóng.

Bước 2: Xây dựng mô hình GRU

- Mô hình GRU nhiều lớp với kích thước ẩn $n\_neurons$.
- Dự đoán các giá trị: Open, High, Low, Close.
- **Công thức GRU:**

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t$$

$$\tilde{h}_t = \tanh(W_h[r_t \odot h_{t-1}, x_t])$$

$$z_t = \sigma(W_z[h_{t-1}, x_t]), \quad r_t = \sigma(W_r[h_{t-1}, x_t])$$

trong đó z_t là cổng cập nhật, r_t là cổng xóa, h_t là trạng thái ẩn.

Bước 3: Huấn luyện mô hình

- Huấn luyện với `Adam Optimizer` và **Mean Squared Error (MSE)** làm hàm mất mát.
- Sử dụng tập xác thực để theo dõi hiệu quả.

2.7.4 Thuật toán XGBoost

Tổng quan về thuật toán XGBoost Các bước thực hiện

Ticker	R^2	RMSE	MAE	MSE	MAPE	SMAPE	Overall Score
HRL (Test)	0.9239	0.0122	0.0082	0.0001591	1.0722	1.0692	0.9826
HRL (Validation)	0.9903	0.0130	0.0095	0.0001772	1.2260	1.2275	0.9955
HBAN (Test)	0.9847	0.0164	0.0121	0.0002747	1.8425	1.8257	0.9932
HBAN (Validation)	0.9656	0.0189	0.0139	0.0003624	2.1867	2.1537	0.9887
HBI (Test)	0.9208	0.0162	0.0123	0.0002619	1.7728	1.7594	0.9805
HBI (Validation)	0.8274	0.0220	0.0157	0.0004933	1.9589	1.9430	0.9614

- Bước 1: Chuẩn bị dữ liệu
  - Tạo ra các đặc trưng bổ sung như đường trung bình động (Moving Averages), dải Bollinger Bands, Chỉ số Sức mạnh Tương đối (RSI), v.v.
- Bước 2: Chia tách dữ liệu
- Bước 3: Tạo ma trận DMatrix
- Bước 4: Xác định siêu tham số
- Bước 5: Huấn luyện mô hình
- Bước 6: Dự đoán và đánh giá mô hình: Sử dụng những chỉ số sau để đánh giá mô hình

Ticker	R^2	MSE	RMSE	MAE	MAPE	SMAPE	Overall Score
HRL	0.9794	1.7298	1.3152	1.0946	5.7760	5.6237	0.9761
HBAN	0.9759	0.1122	0.3349	0.2766	3.7352	3.6728	0.9857
HBI	0.9807	1.7906	1.3381	1.2081	11.1375	10.3957	0.9657

3 Kết luận và định hướng tương lai

3.1 Kết luận

Qua vòng thi bán kết, nhóm F4DT đã học được rất nhiều điều bổ ích. Đầu tiên là những kiến thức chuyên sâu về tài chính và phân tích thị trường chứng khoán, từ các chỉ số tài

chính đến các mô hình dự đoán giá cổ phiếu. Việc áp dụng các thuật toán hiện đại như XGBoost, Voting hồi quy, mạng hồi quy dài hạn và thuật toán chuỗi thời gian ARIMA đã giúp nhóm cải thiện khả năng dự đoán và hiểu rõ hơn về các mô hình học máy trong tài chính.

Cuộc thi THE ECONOMAIST thực sự là một cơ hội tuyệt vời giúp nhóm phát triển toàn diện cả về kiến thức lẫn kỹ năng, chuẩn bị tốt hơn cho những dự án trong tương lai.

3.2 Định hướng phát triển trong tương lai

1. Thử nghiệm với dữ liệu lớn hơn và đa dạng hơn: Thu thập dữ liệu từ nhiều nguồn khác nhau và thử nghiệm với các bộ dữ liệu lớn hơn để cải thiện tính tổng quát của mô hình.

2. Thực hiện đánh giá và phân tích về các nội tại công ty và các hoạt động kinh doanh của họ một cách sâu sắc hơn để thực hiện việc dự đoán và đầu tư.

3. Xây dựng thêm các đặc trưng mới: Kết hợp thêm các đặc trưng mới hoặc thử nghiệm với các kết hợp khác nhau của các chỉ báo hiện có có thể cung cấp thêm thông tin mới và cải thiện độ chính xác dự đoán.

4. Mở rộng thời gian huấn luyện: Nếu tài nguyên cho phép, các thí nghiệm trong tương lai nên xem xét việc tăng số lượng epoch huấn luyện. Điều này có thể giúp mô hình hội tụ hoàn toàn và có khả năng phát hiện ra những mẫu phức tạp hơn trong dữ liệu.

5. Tối ưu hóa phần cứng: Sử dụng tài nguyên tính toán mạnh hơn hoặc nền tảng đám mây để vượt qua giới hạn phần cứng và tiến hành huấn luyện sâu rộng hơn.