Evaluating and Enhancing Artificial Intelligence Approaches for Stock Market Prediction: From Single Stock Analysis to Market Index Forecasting

Hồ Tú Minh
Mã số sinh viên: 22022674

Đỗ Quang Dũng
Mã số sinh viên: 22022561

Phạm Long Nhật
Mã số sinh viên: 22022520

May 5, 2025

Abstract

Bài báo cáo dưới đây nhóm em sẽ đưa ra một phân tích chi tiết về việc ứng dụng trí tuệ nhân tạo và học máy trong dự báo thị trường chứng khoán, dựa trên việc tổng hợp hai nghiên cứu khoa học độc lập. Nghiên cứu thứ nhất tập trung vào việc xây dựng mô hình dự báo vô điều kiện cho chỉ số VNINDEX của thị trường chứng khoán Việt Nam, sử dụng các chỉ số dẫn dắt kinh tế và phân tích thành phần chính (PCA) [1]. Nghiên cứu thứ hai [2] đánh giá hiệu quả của AI trong việc dự báo giá cổ phiếu thông qua phân tích kỹ thuật và phân tích cơ bản, sử dụng dữ liệu lịch sử từ Yahoo Finance và dữ liệu cảm xúc từ Twitter. Kết quả từ cả hai nghiên cứu cho thấy tiềm năng đáng kể của AI và ML, nhưng cũng chỉ ra những hạn chế cần khắc phục để đạt được hiệu suất vượt trội trong dự báo chứng khoán.

1 Giới thiệu

Dự báo diễn biến thị trường chứng khoán (TTCK) là một bài toán cực kỳ thách thức do sự phức tạp, tác động của vô số yếu tố và tính khó đoán định vốn có. Các lý thuyết thị trường như Giả thuyết Thị trường Hiệu quả (EMH) cho rằng không thể dự báo được thị trường, trong khi Giả thuyết Thị trường Thích ứng (AMH) lại gợi ý khả năng dự báo tồn tại, đặc biệt thông qua các phương pháp có khả năng học hỏi và thích ứng.

Trước đây, giới phân tích chủ yếu dựa vào hai trường phái truyền thống: Phân tích Kỹ thuật (TA), tập trung vào biểu đồ giá và khối lượng giao dịch lịch sử, và Phân tích Cơ bản (FA), đi sâu vào đánh giá giá trị nội tại của doanh nghiệp thông qua các yếu tố tài chính và kinh doanh. Tuy nhiên, sự bùng nổ của Trí tuệ Nhân tạo (AI) và Học máy (ML), kết hợp với nguồn dữ liệu lớn (Big Data) và năng lực tính toán mạnh mẽ, đã mang lại một cuộc cách mạng cho lĩnh vực dự báo tài chính. Các thuật toán AI/ML có khả năng tự động phát hiện các mối quan hệ phi tuyến tính và những mẫu hình phức tạp ẩn trong dữ liệu mà con người hay các mô hình truyền thống khó có thể nắm bắt.

Bài báo cáo này sẽ tổng hợp và so sánh hai nghiên cứu khoa học độc lập về việc ứng dụng AI/ML trong dự báo TTCK, sử dụng các phương pháp và nguồn dữ liệu khác nhau. Qua đó, chúng tôi mong muốn làm rõ hơn tiềm năng, những hạn chế thực tế và các yếu tố then chốt ảnh hưởng đến hiệu quả của AI/ML trong lĩnh vực đầy thách thức và hấp dẫn này.

2 Phương pháp

2.1 Nghiên cứu 1: Đánh giá hiệu quả của AI trong dự báo thị trường chứng khoán [2]

2.1.1 Mục tiêu và cách tiếp cận

Nghiên cứu [2] đánh giá hiệu quả của AI/ML trong dự đoán thị trường chứng khoán bằng cách tiếp cận hai phương pháp phân tích chính: Phân tích Kỹ thuật và Phân tích Cơ bản (tập trung vào Phân tích Cảm tính).

2.1.2 Phân tích kỹ thuật

- **Nguồn dữ liệu:** Dữ liệu lịch sử giá cổ phiếu của Apple (AAPL) từ năm 2010 đến 2021, bao gồm giá mở cửa, đóng cửa, cao nhất, thấp nhất, và khối lượng giao dịch.
- **Đặc trưng đầu vào:** Các chỉ số tài chính phổ biến được tính toán, bao gồm:
  - Đường trung bình động đơn giản (Simple Moving Average).
  - Đường trung bình động lũy thừa (Exponential Moving Average).
  - Chỉ số sức mạnh tương đối (Relative Strength Index).
  - Chỉ báo MACD (Moving Average Convergence Divergence).
- **Thuật toán:**
  - Hồi quy tuyến tính: Một mô hình đơn giản để dự đoán giá cổ phiếu dựa trên các đặc trưng trên.
  - LSTM (Long Short-Term Memory): Một loại mạng nơ-ron hồi quy phù hợp với dữ liệu chuỗi thời gian.
- **Đánh giá:** Các chỉ số như R^2, MAPE (Mean Absolute Percentage Error), và RMSE (Root Mean Squared Error) được sử dụng để so sánh hiệu suất của hai mô hình.

2.1.3 Phân tích cơ bản

- **Nguồn dữ liệu:** Dữ liệu Twitter được thu thập để phân tích cảm xúc công chúng về cổ phiếu AAPL.
- **Tiền xử lý dữ liệu:** Các tweet được làm sạch (loại bỏ ký tự đặc biệt, từ dừng), sau đó trích xuất

đặc trưng văn bản bằng phương pháp TF-IDF (Term Frequency-Inverse Document Frequency).

- **Thuật toán:**
  - Logistic Regression: Mô hình phân loại cơ bản để dự đoán xu hướng giá (tăng/giảm).
  - SVM (Support Vector Machine): Một mô hình phân loại mạnh mẽ hơn.
  - ANN (Artificial Neural Network): Mạng nơ-ron nhân tạo để phân tích các mối quan hệ phức tạp.
- **Đánh giá:** Hiệu suất được đo bằng độ chính xác (accuracy), precision, recall, F1-score, và AUC (Area Under the Curve).

2.2 Nghiên cứu 2: Xây dựng mô hình dự báo không điều kiện cho VNINDEX [1]

2.2.1 Mục tiêu và cách tiếp cận

Mục tiêu của bài báo thứ hai nhóm em nghiên cứu là xây dựng một mô hình dự đoán chỉ số thị trường VNINDEX dựa vào các chỉ số tài chính [1]. Bài báo tập trung vào pha trích chọn đặc trưng tốt nhất (từ gần 300 chỉ số tài chính, thu thập theo tháng từ 1/2010 đến 4/2016) để xây dựng các đặc trưng đầu vào cho mô hình dự đoán. Nhóm nghiên cứu sử dụng kết hợp phương pháp Leading indicators và PCA để chọn đặc trưng đầu vào tốt nhất, mô hình lựa chọn để dự đoán chỉ số chứng khoán theo tháng là Multiple Regression. Kết quả nhóm nghiên cứu đạt được là mô hình dự đoán với sai số (MAPE) rất thấp.

2.2.2 Các bước thực hiện

Phương pháp thực hiện nghiên cứu của bài báo gồm các bước:

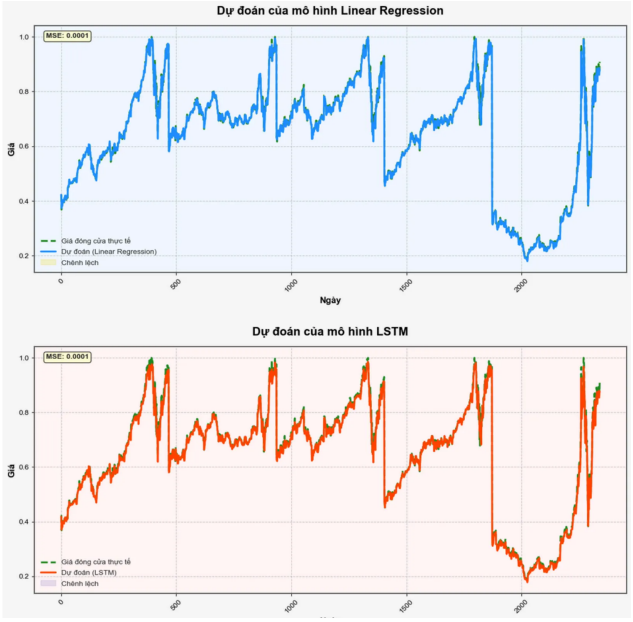
1. Tìm ra các chỉ số trước (leading indicators) bằng cách kiểm tra tính dừng bằng kiểm định Augmented Dickey-Fuller, xác định các biến dừng ở mức độ khác nhau.
2. Sau đó thực hiện kiểm định nhân quả Granger (với độ trễ tối đa 5) xác định 42 biến là chỉ số dẫn dắt của VNINDEX, gồm 4 biến kinh tế-tài chính và 38 biến giao dịch cổ phiếu.
3. Tiếp theo, phân tích thành phần chính (PCA) được áp dụng, kết quả chọn ra 9 thành phần chính (giải thích 72,2% biến thiên dữ liệu) làm các biến đầu vào cho mô hình dự đoán.
4. Mô hình Multiple Regression được xây trên dữ liệu từ tháng 1/2010 đến tháng 12/2015, với phần còn lại (4 tháng đầu 2016) dùng để kiểm tra hiệu suất.

3 Kết quả

3.1 Nghiên cứu 1 [2]

3.1.1 Phân tích kĩ thuật

Hai mô hình có hiệu suất khá tương tự, đều đưa ra kết quả rất tích cực khi dự đoán giá trị đóng cửa cổ phiếu với độ chính xác cao. Trong trường hợp này, Linear Regression là lựa chọn tốt hơn do đơn giản mà vẫn đạt hiệu suất cao. Kết quả này đạt được nhờ việc sử dụng các chỉ số tài chính phổ biến như SMA (Simple Moving Average), EMA (Exponential Moving Average), RSI (Relative Strength Index), MACD (Moving Average Convergence Divergence), và OBV (On-Balance Volume) làm đầu vào cho mô hình. Điều này cho thấy phân tích kỹ thuật, với sự hỗ trợ của học máy, là một công cụ hữu ích trong việc dự đoán biến động giá cổ phiếu.

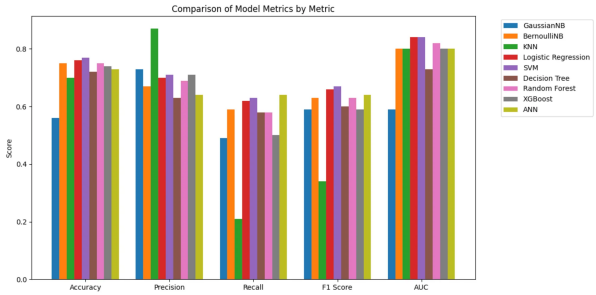


Metric	Linear Regression	LSTM
R-squared	0.9983	0.9982
Explained Variance	0.9984	0.9982
MAE	0.0053	0.0062
RMSE	0.0080	0.0084
MAPE (%)	0.8779	1.0260

Tuy nhiên, kết quả tích cực này chủ yếu dựa trên việc dự đoán giá trị trong ngắn hạn hoặc sử dụng các chỉ số được tính toán gần với thời điểm dự đoán. Để đánh giá khả năng ứng dụng thực tế cho mục tiêu đầu tư dài hạn hơn, nhóm đã thực hiện thêm một thử nghiệm. Khi tham gia vào thị trường cổ phiếu, để có thể thu được lợi nhuận lớn, người chơi cần đưa ra những lựa chọn nhìn xa hơn trong tương lai (ví dụ: 1-3 tháng), và vào thời điểm đó thì họ sẽ không tiếp cận được với những chỉ số tài chính trong quãng thời gian gần. Vì vậy, nhóm đã thử nghiệm thêm việc dự đoán giá đóng sau 90 ngày sử dụng các chỉ số tài chính của cách đó một khoảng thời gian. Mục đích là đánh giá mức độ hiệu quả của model trong bối cảnh thực tế. Kết quả cho thấy rằng độ chính xác của các mô hình còn chưa cao, sai số trung bình (sử dụng độ đo Mean Absolute Percentage Error) của mô hình Linear Regression là 13.94% và của mô hình LSTM là 13.35%. Điều này cho thấy thách thức đáng kể khi áp dụng các mô hình này cho dự báo dài hạn trong thực tế.

3.1.2 Phân tích cơ bản

Trong phân tích cơ bản, nghiên cứu [2] sử dụng dữ liệu từ Twitter để phân tích tâm lý công chúng và đánh giá tác động của nó lên thị trường chứng khoán. Tuy nhiên, kết quả không khả quan như mong đợi. Thuật toán SVM (Support Vector Machine) đạt độ chính xác cao nhất là 75.5%, nhưng các chỉ số khác như Precision,



Recall, và F1-score dao động từ 62% đến 75.7%, cho thấy hiệu suất dự đoán vẫn còn hạn chế.

3.2 Nghiên cứu 2 [1]

Kết quả của nghiên cứu [1] cho thấy kết quả khả quan, với sai số tuyệt đối trung bình rất thấp (dưới 1.6% MAPE) khi dự báo cho 4 tháng tiếp theo, và sai số giảm đáng kể nếu rút ngắn xuống 1-2 tháng, chứng tỏ mô hình hiệu quả hơn cho dự báo ngắn hạn do thị trường Việt Nam có nhiều biến động khó lường trong dài hạn. Mô hình bị hạn chế bởi dữ liệu tại Việt Nam không có sẵn theo quý hoặc năm, khiến việc mở rộng dự báo dài hạn khó khăn hơn và phản ánh hạ tầng dữ liệu thị trường chứng khoán Việt Nam còn chưa phát triển. Điểm mạnh của phương pháp là khả năng xử lý tốt các dao động bất thường như cú sốc kinh tế hoặc thay đổi chính sách nhờ PCA. Vì vậy, nên ưu tiên áp dụng mô hình này cho dự báo ngắn hạn trong các giai đoạn thị trường ổn định để đảm bảo độ chính xác cao.

4 Các Đề xuất Cải tiến và Tối ưu hóa

Dựa trên các kết quả và hạn chế được trình bày trong hai nghiên cứu [1, 2], một số hướng cải tiến và tối ưu hóa tiềm năng có thể được đề xuất để nâng cao hiệu quả của các mô hình AI/ML trong dự báo thị trường chứng khoán:

4.1 Tối ưu hóa Dữ liệu

- **Phân tích Kỹ thuật:**
  - **Đa dạng hóa:** Sử dụng dữ liệu từ nhiều cổ phiếu khác nhau thuộc các ngành và vốn hóa thị trường khác nhau, không chỉ AAPL. Bao gồm cả dữ liệu từ các chỉ số thị trường rộng hơn (ví dụ: S&P 500, VNIndex) làm đặc trưng đầu vào, vì thị trường chung ảnh hưởng đến cổ phiếu riêng lẻ.
  - **Thêm Đặc trưng:** Tích hợp thêm các yếu tố kinh tế vĩ mô (lãi suất, lạm phát, tỷ giá), chỉ số biến động (VIX), dữ liệu phái sinh (hợp đồng tương lai, quyền chọn - ví dụ: tỷ lệ put/call), và có thể cả dữ liệu thay thế (alternative data) nếu có.
  - **Tăng tần suất:** Xem xét sử dụng dữ liệu tần suất cao hơn (phút, giờ) nếu mục tiêu là dự đoán trong ngày (intraday).
- **Phân tích Cơ bản (Cảm tính và xa hơn):**
  - **Mở rộng Nguồn:** Không chỉ dựa vào Twitter. Tích hợp dữ liệu từ các nguồn tin tức tài chính uy tín (Reuters, Bloomberg, CafeF, Vietstock), báo cáo tài chính công ty, biên bản họp cổ đông/phân tích, hồ sơ nộp cho UBCKNN, nhận định của các nhà phân tích.
  - **Phân tích NLP Nâng cao:** Sử dụng các kỹ thuật NLP tiên tiến hơn để phân tích cảm tính (ví dụ: nhận diện sắc thái, mỉa mai, aspect-based sentiment - cảm tính về khía cạnh cụ thể). Sử dụng các mô hình ngôn ngữ lớn được huấn luyện trước và tinh chỉnh cho văn bản tài chính (ví dụ: FinBERT).
  - **Chất lượng Dữ liệu:** Áp dụng các bộ lọc mạnh mẽ hơn để loại bỏ nhiễu, tin rác, bot trên mạng xã hội.

4.2 Tối ưu hóa Mô hình

- **Phân tích Kỹ thuật:**
  - **Mô hình Chuỗi Thời gian Nâng cao:** Khám phá các mô hình khác ngoài LR và LSTM như ARIMA/SARIMA, Prophet (của Facebook), các mô hình Gradient Boosting (LightGBM, CatBoost) được điều chỉnh cho chuỗi thời gian, và đặc biệt là các kiến trúc Transformer được thiết kế cho dữ liệu tuần tự.
  - **Tinh chỉnh Siêu tham số (Hyperparameter Tuning):** Thực hiện tinh chỉnh siêu tham số một cách có hệ thống cho các mô hình phức tạp như LSTM (sử dụng GridSearch, RandomizedSearch, Bayesian Optimization).
- **Phân tích Cơ bản:**
  - **Mô hình Transformer:** Sử dụng các mô hình Transformer như BERT, PhoBERT (cho tiếng Việt), FinBERT để vector hóa và phân loại văn bản. Chúng nắm bắt ngữ cảnh tốt hơn nhiều so với các phương pháp truyền thống.
- **Mô hình Lai (Hybrid Models):**
  - **Kết hợp Đặc trưng:** Đây là hướng đi quan trọng. Đưa cả đặc trưng kỹ thuật và đặc trưng cơ bản vào cùng một mô hình để nhận được kết quả tốt hơn.

4.3 Tối ưu hóa Đánh giá

- **Thêm Chỉ số Tài chính (Backtesting):** Cần xây dựng một hệ thống kiểm thử lại (backtesting) để mô phỏng chiến lược giao dịch dựa trên dự đoán của mô hình. Tính toán các chỉ số tài chính thực tế như:
  - Tổng lợi nhuận (Cumulative Return)
  - Tỷ lệ Sharpe (lợi nhuận điều chỉnh theo rủi ro)
  - Tỷ lệ Sortino (lợi nhuận điều chỉnh theo rủi ro thua lỗ)
  - Mức sụt giảm tối đa (Maximum Drawdown)
  - Tỷ lệ thắng (Win Rate), Hệ số lợi nhuận (Profit Factor)

5 Kết luận

Phân tích hai nghiên cứu cho thấy Trí tuệ Nhân tạo (AI) và Học máy (ML) có tiềm năng lớn nhưng cũng đối mặt thách thức đáng kể trong dự báo thị trường

chứng khoán. Nghiên cứu dự báo VNINDEX (Nghiên cứu 2) [1] cho thấy hiệu quả tích cực với phương pháp kết hợp Chỉ báo dẫn dắt và PCA, đặc biệt trong dự báo ngắn hạn. Ngược lại, nghiên cứu trên cổ phiếu AAPL (Nghiên cứu 1) [2] thể hiện sự thận trọng: dù mô hình Phân tích Kỹ thuật có thể bám sát giá lịch sử, độ tin cậy cho giao dịch thực tế còn bỏ ngỏ, và Phân tích Cảm tính từ Twitter cho hiệu quả hạn chế. Điều này cho thấy cần cẩn trọng với các tuyên bố về khả năng "đánh bại thị trường" của AI.

Hiệu quả của AI/ML phụ thuộc mạnh mẽ vào nhiều yếu tố như mục tiêu dự báo, chất lượng dữ liệu, phương pháp luận, và đặc biệt là quy trình đánh giá thực tế thông qua kiểm thử lại chiến lược (backtesting). Để tiến bộ, cần tập trung vào việc phát triển mô hình lai, ứng dụng các kỹ thuật học sâu/NLP tiên tiến, và xây dựng quy trình đánh giá chặt chẽ hơn, chú trọng tính thực tiễn và khả năng diễn giải.

Tóm lại, AI/ML không phải là giải pháp toàn năng nhưng là những công cụ hỗ trợ mạnh mẽ. Thành công trong lĩnh vực này đòi hỏi sự kết hợp nhuần nhuyễn giữa chuyên môn tài chính, kỹ năng khoa học dữ liệu và tư duy phản biện thực tế để khai thác hiệu quả tiềm năng của công nghệ trong một thị trường đầy phức tạp và biến động.

References

[1] Building Unconditional Forecast Model of Stock Market Indexes Using Combined Leading Indicators and Principal Components: Application to Vietnamese Stock Market. *Indian Journal of Science and Technology*. Available at: `https://indjst.org/download-article.php?Article_Unique_Id=INDJST157&Full_Text_Pdf_Download=True`

[2] Effectiveness of Artificial Intelligence in Stock Market Prediction Based on Machine Learning *arXiv preprint arXiv:2107.01031*, 2021. Available at: `https://arxiv.org/abs/2107.01031`