

Phân tích, dự đoán giá chứng khoán bằng mô hình thống kê, học máy và học sâu

Lê Anh Tuấn Dũng

MSSV: 21521974

Trường đại học Công Nghệ Thông Tin

Khoa Hệ Thống Thông Tin

21521974@gm.uit.edu.vn

Đỗ Sĩ Đạt

MSSV: 21521932

Trường đại học Công Nghệ Thông Tin

Khoa Hệ Thống Thông Tin

21521932@gm.uit.edu.vn

Lê Thị Ánh Hồng

MSSV: 21520245

Trường đại học Công Nghệ Thông Tin

Khoa Hệ Thống Thông Tin

21520245@gm.uit.edu.vn

Nguyễn Thị Mai Liên

MSSV: 21522283

Trường đại học Công Nghệ Thông Tin

Khoa Hệ Thống Thông Tin

21522283@gm.uit.edu.vn

Trần Thị Mỹ Xoan

MSSV: 215222815

Trường đại học Công Nghệ Thông Tin

Khoa Hệ Thống Thông Tin

215222815@gm.uit.edu.vn

Tóm tắt nội dung:

Khi thực hiện đầu tư vào các cổ phiếu chứng khoán, việc sử dụng các mô hình dự đoán giá cổ phiếu trước khi thực hiện một giao dịch đầu tư có thể giúp cho ta có các quyết định đúng đắn hơn khi ra quyết định thực hiện 1 giao dịch. Bài báo này trình bày phương pháp dự đoán giá cổ phiếu của ba công ty lớn Samsung, LG và Sony. Chúng tôi sẽ thực hiện dự đoán bằng cách sử dụng các mô hình thống kê, học máy, học sâu: VARMA, XGBoost, NBeats, Gradient Boosting, LightGBM. Bộ dữ liệu được sử dụng lấy từ ngày 1/3/2019 đến ngày 1/6/2024. Sau khi thực hiện dự đoán thì ta sẽ thực hiện sử dụng các độ đo MAPE, MSE, RMSE để đánh giá các mô hình. Cuối cùng, sử dụng các mô hình để thực hiện dự đoán giá chứng khoán cho 30, 60 và 90 ngày tiếp theo.

Từ khoá: VARMA, XGBoost, NBeats, Gradient Boosting, LightGBM. Mô hình thống kê, mô hình học máy, mô hình học sâu, dự đoán giá chứng khoán.

I. GIỚI THIỆU

Dự đoán giá cổ phiếu là một vấn đề quan trọng trong lĩnh vực đầu tư tài chính. Việc dự đoán chính xác giá cổ phiếu có thể giúp các nhà đầu tư đưa ra quyết định đầu tư sáng suốt và giảm thiểu rủi ro. Có nhiều phương pháp khác nhau để dự đoán giá cổ phiếu, bao gồm phân tích kỹ thuật, phân tích cơ bản và học máy. Samsung, LG, SONY là những công ty hàng đầu trong ngành công nghệ, giá cổ phiếu của 3 công ty đóng vai trò là chỉ số quan trọng về động lực thị trường, khiến việc phân tích của họ trở nên cần thiết đối với các nhà đầu tư.

Báo cáo này nhằm mục đích dự đoán giá cổ phiếu của các công ty có ảnh hưởng bằng cách sử dụng 5 thuật toán VARMA, XGBoost, NBeats, Gradient Boosting, LightGBM để đưa ra dự đoán cho giá cổ phiếu. Nhờ đó các nhà đầu tư có thể lựa chọn chính xác hạng mục đầu tư vào giá cổ phiếu.

Bằng cách tận dụng những phương pháp này, chúng tôi mong muốn cung cấp hướng dẫn có giá trị cho các nhà đầu tư trong việc định hướng bối cảnh năng động của lĩnh vực công nghệ.

II. NGHIÊN CỨU LIÊN QUAN

[1] Thực hiện giả định có hai cách tiếp cận giá trị đầu vào cho mô hình, dữ liệu liên tục và dữ liệu nhị phân. Ở hướng tiếp cận thứ nhất với dữ liệu liên tục, kết quả cho ra là RNN và LSTM là hai yếu tố dự đoán hàng đầu (khoảng 86% điểm F1). Ở hướng tiếp cận thứ hai với dữ liệu nhị phân thì hai yếu tố dự đoán tốt nhất là RNN và LSTM (với khoảng 90% điểm F1) và quá trình dự đoán cho tất cả các mô hình đều nhanh hơn. Các công trình thử nghiệm cho thấy sự cải thiện đáng kể về hiệu suất của các mô hình khi sử dụng dữ liệu nhị phân thay vì dữ liệu liên tục.

[2]: Bài báo này đã tham gia vào cuộc tranh luận về tính hữu ích của phân tích tâm lý đối với việc dự đoán diễn biến thị trường chứng khoán. Bài báo đã sử dụng dữ liệu Twitter làm kho thông tin của mình để dự đoán những thăng trầm của sáu công ty NASDAQ nổi tiếng. Bài báo đã đề xuất một phương pháp bắt đầu bằng cách trích xuất nhiều đặc điểm dựa trên văn bản để làm phong phú thêm việc thể hiện tình cảm. Sau đó kết quả thực nghiệm đã cho ra kết luận rằng sự tang giảm giá cổ phiếu của một công ty bị ảnh hưởng bởi quan điểm hoặc công chúng thể hiện trên Twitter.

[3]: Bài báo này nhằm mục đích dự đoán hướng đi của giá cổ phiếu Mỹ bằng cách tích hợp entropy truyền hiệu quả thay đổi theo thời gian (ETE) và các thuật toán học máy khác nhau. Đầu tiên, bài báo khám phá rằng ETE dựa trên thời hạn biến động 3 và 6 tháng có thể được coi là biến giải thích thị trường bằng cách phân tích mối liên hệ giữa các cuộc khủng hoảng tài chính và mối quan hệ nhân quả Granger giữa các cổ phiếu. Sau đó, bài báo phát hiện ra rằng hiệu suất dự đoán theo hướng giá cổ phiếu có thể được cải thiện khi biến điều khiển ETE được tích hợp như một tính năng mới trong hồi quy logistic, perceptron đa lớp, rừng ngẫu nhiên, XGBoost và mạng bộ nhớ ngắn hạn

dài. Cuối cùng, bài báo xác nhận rằng mạng perceptron đa lớp và bộ nhớ ngắn hạn dài phù hợp hơn cho việc dự đoán giá cổ phiếu. Nghiên cứu này là nỗ lực đầu tiên nhằm dự đoán hướng giá cổ phiếu bằng cách sử dụng ETE, có thể áp dụng thuận tiện vào thực tế.

[4] Mô hình VARMA thường được sử dụng để dự báo dữ liệu chuỗi thời gian đa biến. Mô hình VARMA là sự mở rộng của mô hình ARMA trong chuỗi thời gian đơn biến (Lütkepohl, 2005; Wei, 1990) và được sử dụng với điều kiện rằng dữ liệu phải ổn định theo thời gian (Lütkepohl, 2005). Mô hình VARMA (p,q) là sự kết hợp của mô hình VAR (p) và mô hình vector trung bình động (q) (VMA (q)). Bài báo này xác định 4 mô hình VARMA (p,q) lần lượt là (1,1), (2, 1), (3, 1), (4, 1). Việc lựa chọn mô hình tốt nhất được thực hiện bằng cách sử dụng một số tiêu chí thông tin (AICC, HQC, AIC, và SBC). Các giá trị nhỏ nhất của những tiêu chí này cho biết mô hình tốt nhất. Mô hình phù hợp nhất được chọn là VARMA(2, 1). Kết quả dự báo cho thấy sai số chuẩn tăng lên theo thời gian; sai số chuẩn trong tháng đầu tiên tương đối nhỏ so với dự đoán trung bình, nhưng tăng lên theo thời gian đến dự báo cho 12 tháng tiếp theo. Điều này cho thấy mô hình là hợp lý khi dự báo cho các khoảng thời gian ngắn, nhưng kết quả không ổn định (do sai số chuẩn cao hơn) khi dự báo cho các khoảng thời gian dài.

[5] Và để tìm ra tham số tối ưu cho mô hình VARMA, ta sử dụng auto_arima để thực hiện tìm các tham số tối ưu (p, q) cho mỗi biến. Cặp tham số của một biến có AIC nhỏ nhất là phù hợp nhất. Sau đó, huấn luyện với các cặp tham số (p, q) tìm được. Sau khi chạy, ta sử dụng chỉ số đánh giá RMSE để xem cặp tham số (p, q) nào phù hợp nhất với mô hình. RMSE càng nhỏ thì cặp tham số đó càng phù hợp.

[8] Bài báo đã đề xuất một mô hình kết hợp có tên HDFM để dự đoán giá chỉ số chứng khoán bằng cách sử dụng GRU và các phương pháp phân rã. Cụ thể, bài báo phân tách chuỗi thời gian giá cổ phiếu thành nhiều chuỗi phụ bằng phương pháp CEEMDAN. Để giảm thời gian tính toán, chuỗi con có entropy mẫu tương tự được hợp nhất bằng K-means. Co-IMF1 có tần suất cao nhất trong số tất cả các Co-IMF sẽ được phân tách thành chuỗi phụ để dự đoán dễ dàng hơn. Nghiên cứu cắt bỏ đã chứng minh tính hiệu quả của từng thành phần của mô hình, bao gồm GRU là mạng cơ sở, CEEMDAN là phương pháp phân rã và VMD là phương pháp phân tách lại. Mô hình của chúng tôi vượt trội hơn các phương pháp khác cho cả ba chỉ số thị trường chứng khoán.

[9] Bài báo đã đề xuất một mô hình kết hợp có tên HDFM để dự đoán giá chỉ số chứng khoán bằng cách sử dụng GRU và các phương pháp phân rã. Cụ thể, bài báo phân tách chuỗi thời gian giá cổ phiếu thành nhiều chuỗi phụ bằng phương pháp CEEMDAN. Để giảm thời gian tính toán, chuỗi con có entropy mẫu tương tự được hợp nhất bằng K-means. Co-IMF1 có tần suất cao nhất trong số tất cả các Co-IMF sẽ được phân tách thành chuỗi phụ để dự đoán dễ dàng hơn. Nghiên cứu cắt bỏ đã chứng minh tính hiệu quả của từng thành phần của mô hình, bao gồm GRU là mạng cơ sở, CEEMDAN là phương pháp phân rã và VMD là phương pháp phân tách lại. Mô hình của chúng tôi vượt trội hơn các phương pháp khác cho cả ba chỉ số thị trường chứng khoán.

[10] Bài báo đề xuất Hệ số tương quan song phương (BCORR) mới, có thể phát hiện một số chỉ báo sinh lời mà trước đây bị loại bỏ do CORR thấp. Để tạo ra một chỉ báo có BCORR cao với lợi nhuận, chúng tôi đề xuất một khung gọi là Khung giao dịch tự động song phương (BAF) dựa trên Tổn thất song phương để dự báo thứ hạng chéo của lợi nhuận cổ phiếu và dự đoán được sử dụng làm chỉ báo song phương để chọn cổ phiếu để đầu tư. Trong khi đó, vị thế của các cổ phiếu được chọn được tối ưu hóa theo định hướng Sharpe nhằm giảm thiểu rủi ro và cải thiện lợi nhuận. Các thử nghiệm trên bộ dữ liệu thi trường chứng khoán trong thế giới thực cho thấy BAF có thể cải thiện đáng kể hiệu suất đối với các phương pháp dự đoán chứng khoán chuyên sâu, chẳng hạn như Transformer, LSTM và NBEATS.

III. NGUYÊN LIỆU

A. DATASET

Trọng tâm của bài viết này là dự đoán giá cổ phiếu. Vì vậy, cả ba tập dữ liệu chúng tôi sử dụng trong nghiên cứu đều là giá cổ phiếu của các công ty lớn. Các công ty này là SONY, SAMSUNG và LG. Dữ liệu của SAMSUNG, LG, SONY được thu thập từ ngày 1 tháng 3 năm 2019 đến ngày 1 tháng 6 năm 2024.

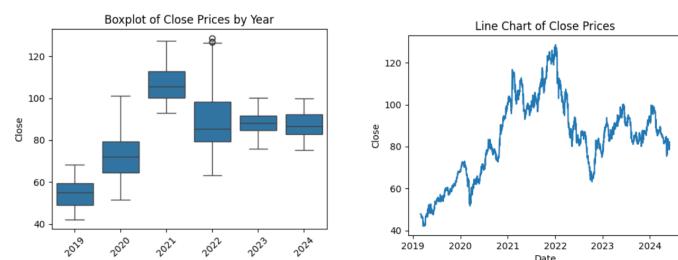
Mô tả từng cột trong dữ liệu :

- + Date : Ngày giao dịch.
- + Open: Giá mở cửa của cổ phiếu vào ngày giao dịch.
- + High: Giá cao nhất của cổ phiếu trong ngày giao dịch.
- + Low: Giá thấp nhất của cổ phiếu trong ngày giao dịch.
- + Close: Giá đóng cửa của cổ phiếu vào ngày giao dịch.

B. THỐNG KÊ MÔ TẢ

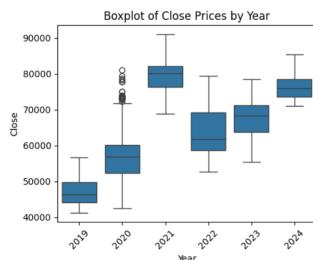
	SONY	SAMSUNG	LG
Count	1920	1920	1920
Mean	83.842011	83.842011	101952.316347
Std	15.828603	15.828603	26596.447698
Min	42.030000	42.030000	41850
25%	79.987500	79.987500	87200
50%	83.842011	83.842011	101952.316347
75%	90.442500	90.442500	114000
Max	128.59	128.590000	185000

Bảng I: SONY, SAMSUNG, LG's Descriptive Statistics

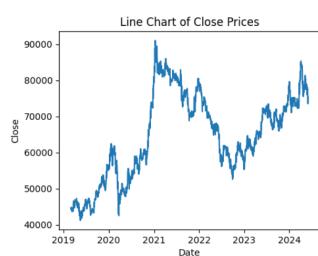


Hình 1: SONY stock price's boxplot

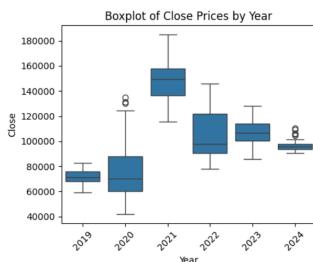
Hình 2: SONY stock price's histogram



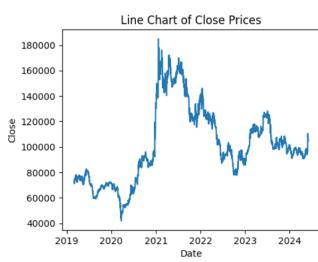
Hình 3: SAMSUNG stock price's boxplot



Hình 4: SAMSUNG stock price's histogram



Hình 5: LG stock price's boxplot



Hình 6: LG stock price's histogram

C. TỶ LỆ PHÂN CHIA TẬP DỮ LIỆU

Tỷ lệ phân chia dữ liệu là một quyết định quan trọng trong việc xây dựng mô hình học máy. Trong bài nghiên cứu này, chúng tôi sẽ chia dữ liệu theo ba tỷ lệ khác nhau: 6:4, 7:3 và 8:2. Tỷ lệ 6:4, cho phép chúng ta có một tập kiểm thử tương đối lớn, giúp đánh giá mô hình một cách toàn diện hơn. Tuy nhiên, điều này cũng có nghĩa là chúng ta có ít dữ liệu hơn để huấn luyện mô hình, có thể ảnh hưởng đến khả năng học của mô hình. Tỷ lệ 7:3 cho phép chúng ta huấn luyện mô hình trên một lượng dữ liệu lớn hơn, trong khi vẫn đảm bảo có đủ dữ liệu để kiểm thử mô hình. Điều này giúp cải thiện hiệu suất của mô hình và khả năng tổng quát hóa. Tỷ lệ 8:2 cung cấp một sự cân nhắc giữa việc huấn luyện mô hình và đánh giá toàn diện trên tập dữ liệu kiểm thử, cung cấp một lựa chọn linh hoạt cho nhiều tình huống khác nhau. Nghiên cứu của chúng tôi nhấn mạnh tầm quan trọng của việc xem xét tỷ lệ phân chia dữ liệu để đảm bảo hiệu quả trong cả việc huấn luyện và đánh giá mô hình. Phương pháp tiếp cận tỉ mỉ này giúp đảm bảo rằng mô hình của chúng tôi có khả năng cung cấp dự đoán chính xác và đáng tin cậy trong các tình huống thực tế. Đây không chỉ là một phần quan trọng của quá trình nghiên cứu của chúng tôi mà còn là một bước quan trọng để đảm bảo rằng kết quả của chúng tôi có thể được áp dụng và triển khai hiệu quả trong thực tế.

D. CHỈ SỐ ĐÁNH GIÁ MÔ HÌNH

1) Mean Squared Error(MSE)

Mean Squared Error (MSE) là một số liệu phổ biến được sử dụng trong các bài toán hồi quy. Về cơ bản, MSE tính sai số bình phương trung bình giữa các giá trị được dự đoán và giá trị thực tế. Đây là một thước đo chất lượng của mô hình dự đoán, luôn không âm và các giá trị càng gần 0 càng tốt.

Công thức tính MSE:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

trong đó:

- n là số điểm dữ liệu
- y_i là giá trị quan sát
- \hat{y}_i là giá trị dự đoán

Trong phân tích hồi quy, biểu đồ giúp trực quan hóa xu hướng dữ liệu. MSE (Mean Squared Error) đo mức độ gần của đường hồi quy với các điểm dữ liệu bằng cách tính khoảng cách từ các điểm đến đường hồi quy, rồi bình phương các khoảng cách này. Bình phương giúp loại bỏ dấu âm và tăng trọng số cho các sai số lớn.

Để giảm thiểu MSE, mô hình cần dự đoán chính xác hơn, tức gần với dữ liệu thực tế hơn. Ví dụ, hồi quy tuyến tính sử dụng phương pháp bình phương nhỏ nhất để đánh giá sự phù hợp của mô hình với dữ liệu hai biến. Tuy nhiên, phương pháp này có giới hạn nếu phân phối dữ liệu không phù hợp. MSE càng thấp thì dự báo càng tốt.

2) Root Mean Square Error(RMSE)

Root Mean Square Error (RMSE) là căn bậc hai của mức trung bình các sai số bình phương. RMSE là độ lệch chuẩn của các phần dư (sai số dự đoán). Phần dư đo khoảng cách từ các điểm dữ liệu đến đường hồi quy, RMSE đo mức độ phân tán của các phần dư này, tức là mức độ tập trung của dữ liệu xung quanh đường hồi quy.

Công thức tính RMSE:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}$$

trong đó:

- n là số điểm dữ liệu
- y_i là giá trị thực tế
- \hat{y}_i là giá trị dự đoán

RMSE (Root Mean Squared Error) chịu ảnh hưởng mạnh bởi các sai số lớn, khiến nó nhạy cảm với các yếu tố ngoại lai. RMSE được dùng trong khí hậu học, dự báo và phân tích hồi quy để kiểm tra kết quả. Khi dữ liệu được chuẩn hóa, RMSE liên quan trực tiếp đến hệ số tương quan; hệ số tương quan bằng 1 thì RMSE bằng 0, chỉ ra không có sai số. RMSE luôn không âm và càng gần 0 thì mô hình càng tốt.

3) Mean Absolute Percentage Error(MAPE)

Mean Absolute Percentage Error (MAPE) là một số liệu được sử dụng để đo lường độ chính xác của các dự đoán trong các mô hình dự báo. MAPE tính toán sai số tuyệt đối trung bình dưới dạng phần trăm của giá trị thực tế, giúp đánh giá mức độ chính xác của dự đoán một cách dễ hiểu và so sánh dễ dàng giữa các tập dữ liệu khác nhau.

Công thức tính MAPE:

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

trong đó:

- n là số điểm dữ liệu
- y_i là giá trị thực tế
- \hat{y}_i là giá trị dự đoán

Ưu điểm của MAPE là nó biểu diễn sai số dưới dạng phần trăm, làm cho việc diễn giải và so sánh dễ dàng hơn. Tuy nhiên, MAPE có một số hạn chế, chẳng hạn như việc không xác định khi giá trị thực tế y_i bằng 0 và có thể bị ảnh hưởng mạnh bởi các giá trị cực nhỏ của y_i , làm cho sai số phần trăm trở nên rất lớn.

MAPE thường được sử dụng trong các lĩnh vực như dự báo kinh tế, tài chính và quản lý chuỗi cung ứng để đánh giá độ chính xác của các mô hình dự báo.

IV. PHƯƠNG PHÁP

A. THUẬT TOÁN RNN

Mạng nơ-ron tái phát (RNN) là một loại mạng nơ-ron nhân tạo được thiết kế đặc biệt để xử lý các chuỗi dữ liệu có thứ tự, như chuỗi thời gian hoặc văn bản. Điểm đặc biệt của RNN là khả năng ghi nhớ và sử dụng thông tin từ các bước trước đó trong chuỗi, làm cho chúng rất hữu ích trong các bài toán như dịch máy, nhận dạng giọng nói và phân tích chuỗi thời gian.

1) Cấu trúc và hoạt động của RNN

RNN có cấu trúc đặc biệt với các kết nối hồi quy (recurrent connections), cho phép truyền thông tin ngược trở lại từ bước hiện tại về các bước trước đó. Điều này có nghĩa là tại mỗi bước thời gian t , trạng thái ẩn $a^{(t)}$ được tính toán dựa trên trạng thái ẩn từ bước trước đó $a^{(t-1)}$ và đầu vào hiện tại $x^{(t)}$.

2) Công thức của RNN

Công thức tính toán trong RNN có thể được biểu diễn như sau:

1. Trạng thái ẩn $a^{(t)}$:

$$a^{(t)} = g_1 \left(W_{aa} a^{(t-1)} + W_{ax} x^{(t)} + b_a \right)$$

2. Đầu ra $y^{(t)}$:

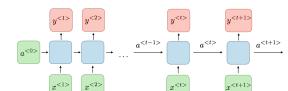
$$y^{(t)} = g_2 \left(W_{ya} a^{(t)} + b_y \right)$$

Trong đó:

- W_{ax} và W_{aa} là ma trận trọng số cho đầu vào và trạng thái ẩn.
- W_{ya} là ma trận trọng số cho đầu ra.
- b_a và b_y là các bias tương ứng.
- g_1 và g_2 là các hàm kích hoạt, thường là hàm sigmoid hoặc hàm tanh cho trạng thái ẩn và hàm softmax cho đầu ra.

RNN rất mạnh mẽ nhưng cũng gặp khó khăn trong việc xử lý các chuỗi dài do vấn đề vanishing gradient. Để giải quyết vấn đề này, các biến thể của RNN như LSTM (Long Short-Term Memory) và GRU (Gated Recurrent Unit) đã được phát triển.

RNN là một công cụ quan trọng trong học máy và trí tuệ nhân tạo, đặc biệt khi làm việc với dữ liệu tuần tự và có cấu trúc thời gian.



Hình 7: Mô hình RNN đơn giản

B. THUẬT TOÁN LIGHTGBM

LightGBM là một thuật toán học máy dựa trên cây quyết định, nó là một phần của họ các thuật toán tăng cường gradient.

Công thức cơ bản của LightGBM là:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), \quad f_k \in \mathcal{F}$$

Trong đó:

\hat{y}_i là giá trị dự đoán cho mẫu i

K là số lượng cây (trees) được sử dụng

f_k là cây thứ k trong tập hợp các cây \mathcal{F}

x_i là vector đặc trưng cho mẫu i

LightGBM tối ưu hóa hàm mất mát bằng cách sử dụng thuật toán Gradient Boosting (tăng cường gradient).

Hàm mất mát thường được định nghĩa như sau:

$$\text{loss}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{i=1}^n L(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

Trong đó:

\mathbf{y} là vector các giá trị thực của các mẫu

$\hat{\mathbf{y}}$ là vector các giá trị dự đoán

n là số lượng mẫu

L là hàm mất mát cho dự đoán và giá trị thực

Ω là hàm điều chỉnh (regularization function) cho cây

LightGBM sử dụng các kỹ thuật như Gradient Boosting Decision Trees (GBDT) và histogram-based algorithm để tối ưu hóa hiệu suất tính toán và bộ nhớ.

C. THUẬT TOÁN VARMA

Mô hình VARMA thường được sử dụng để dự báo dữ liệu chuỗi thời gian đa biến. Mô hình VARMA là một mở rộng của mô hình ARMA trong chuỗi thời gian đơn biến (Lutkepohl, 2005; Wei, 1990) và được sử dụng với điều kiện dữ liệu phải là dừng theo thời gian (Lutkepohl, 2005). Mô hình VARMA (p,q) là sự kết hợp của mô hình VAR (p) và mô hình trung bình trượt vector (q) (VMA (q)).

- Công thức:

$$\mathbf{y}_t = \mathbf{c} + \sum_{i=1}^p \Phi_i \mathbf{y}_{t-i} + \sum_{j=1}^q \Theta_j \epsilon_{t-j} + \epsilon_t \quad (1)$$

Trong đó:

- \mathbf{y}_t là vector của các biến tại thời điểm t .
- \mathbf{c} là một vector hằng số.
- Φ_i là ma trận hệ số của chuỗi thời gian (độ trễ, lag) i của mô hình AR.
- Θ_j là ma trận hệ số của chuỗi thời gian (độ trễ, lag) j của mô hình MA.
- ϵ_t là vector của các nhiễu ngẫu nhiên tại thời điểm t .

D. ARIMA

ARIMA model là viết tắt của cụm từ Autoregressive Intergrated Moving Average. Mô hình sẽ biểu diễn phương trình hồi qui tuyến tính đa biến (multiple linear regression) của các biến đầu vào (còn gọi là biến phụ thuộc trong thống kê) là 2 thành phần chính:

- Auto regression: Kí hiệu là AR. Đây là thành phần tự hồi qui bao gồm tổng hợp các độ trễ của biến hiện tại. Độ trễ bậc p chính là giá trị lùi về quá khứ bước thời gian của chuỗi. Độ trễ dài hoặc ngắn trong quá trình AR phụ thuộc vào tham số trễ.
- Moving average: Quá trình trung bình trượt được hiểu là quá trình dịch chuyển hoặc thay đổi giá trị trung bình của chuỗi theo thời gian. Do chuỗi của chúng ta được giả định là dừng nên quá trình thay đổi trung bình dường như là một chuỗi nhiễu trắng. Quá trình moving average sẽ tìm mối liên hệ về mặt tuyến tính giữa các phần tử ngẫu nhiên(stochastic term).
- Intergrated: Là quá trình đồng tích hợp hoặc lấy sai phân. Yêu cầu chung của các thuật toán trong time series là chuỗi phải đảm bảo tính dừng. Hầu hết các chuỗi đều tăng hoặc giảm theo thời gian. Do đó yếu tố tương quan giữa chúng chưa chắc là thực sự mà là do chúng cùng tương quan theo thời gian. Khi biến đổi sang chuỗi dừng, các nhân tố ảnh hưởng thời gian được loại bỏ và chuỗi sẽ dễ dự báo hơn. Để tạo thành chuỗi dừng, một phương pháp đơn giản nhất là chúng ta sẽ lấy sai phân. Một số chuỗi tài chính còn qui đổi sang logarit hoặc lợi suất. Bậc của sai phân để tạo thành chuỗi dừng còn gọi là bậc của quá trình đồng tích hợp (order of intergration).

- Công thức:

$$\mathbf{y}_t = \mathbf{c} + \epsilon_t + \sum_{i=1}^p \Phi_i \mathbf{y}_{t-i} + \sum_{j=1}^q \Theta_j \epsilon_{t-j} \quad (2)$$

Trong đó:

- \mathbf{y}_t là giá trị của chuỗi thời gian tại thời điểm t .
- \mathbf{c} là hằng số.
- ϵ_t là thành phần lỗi tại thời điểm thời điểm t .
- Φ_i là hệ số của thành phần tự hồi quy (AR).
- Θ_j là hệ số của thành phần trung bình trượt (MA).

E. THUẬT TOÁN GRADIENT BOOSTING REGRESSION

Gradient Boosting Regressor là một thuật toán học máy mạnh mẽ được sử dụng cho các bài toán hồi quy. Nó kết hợp nhiều mô hình hồi quy đơn giản (thường là các cây quyết định) để tạo ra một mô hình dự đoán mạnh mẽ hơn. Thuật toán này

dựa trên phương pháp boosting, một kỹ thuật học tập ensemble, trong đó các mô hình con được xây dựng liên tiếp nhau, với mỗi mô hình mới nhằm mục đích sửa lỗi của mô hình trước đó. Gradient Boosting là xây dựng mô hình dần dần bằng cách thêm vào các mô hình con mà mỗi mô hình mới tập trung vào việc sửa lỗi (residual) của các dự đoán trước đó. Điều này được thực hiện bằng cách tối ưu hóa gradient của hàm lỗi.

Các bước thực hiện

Khởi tạo mô hình ban đầu:

Bắt đầu với một mô hình đơn giản, thường là giá trị trung bình của biến mục tiêu y .

Tính toán residuals:

Residuals là sai số giữa giá trị thực tế và giá trị dự đoán của mô hình hiện tại.

Huấn luyện mô hình con:

Huấn luyện một mô hình con (thường là một cây quyết định) để dự đoán residuals.

Cập nhật mô hình hiện tại:

Cập nhật mô hình hiện tại bằng cách thêm mô hình con mới với một trọng số (learning rate).

Lặp lại:

Lặp lại quá trình cho đến khi đạt được số lượng mô hình con mong muốn hoặc sai số giảm đến mức chấp nhận được.

Ưu điểm và nhược điểm

Ưu điểm:

Hiệu quả cao trong việc giảm sai số và dự đoán chính xác. Linh hoạt với khả năng xử lý nhiều loại dữ liệu khác nhau. Có thể điều chỉnh nhiều tham số để tối ưu hóa hiệu suất.

Nhược điểm:

Tốn nhiều thời gian và tài nguyên tính toán. Dễ bị overfitting nếu không điều chỉnh cẩn thận các tham số. Khó khăn trong việc giải thích mô hình do tính phức tạp cao.

F. THUẬT TOÁN XGBoost

Thuật toán XGBoost là một thuật toán máy học thuộc loại ensemble learning, cụ thể là Gradient Boosting Framework. Nó sử dụng cây quyết định làm người học cơ sở và sử dụng các kỹ thuật chính quy hóa để nâng cao khả năng khai thác của mô hình. XGBoost được sử dụng rộng rãi cho các tác vụ như hồi quy, phân loại và xếp hạng.

XGBoost là một thuật toán máy học theo phương pháp học tập tổng hợp. Nó là xu hướng cho các nhiệm vụ học tập có giám sát, chẳng hạn như hồi quy và phân loại. XGBoost xây dựng mô hình dự đoán bằng cách kết hợp các dự đoán của nhiều mô hình riêng lẻ, thường là cây quyết định.

XGBoost tối ưu hàm mục tiêu tại lần lặp t là:

$$\text{Obj}(t) = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (3)$$

Trong đó:

- l là hàm mất mát (loss function, ví dụ như MSE cho regression, logistic loss cho classification).
- y_i là giá trị thực tế của quan sát thứ i .
- $\hat{y}_i^{(t-1)}$ là dự đoán tại vòng lặp thứ $t - 1$.
- f_t là cây quyết định được thêm vào tại vòng lặp thứ t .
- $\Omega(f_t)$ là hàm phạt (regularization term) giúp ngăn ngừa overfitting.

Các lợi ích và đặc điểm của mô hình XGBoost:

- Độ chính xác cao: Bộ phân loại XGBoost nổi tiếng với độ chính xác cao và đã được chứng minh vượt trội hơn so với các thuật toán máy học khác trong nhiều nhiệm vụ dự đoán.
- Khả năng mở rộng: XGBoost có khả năng mở rộng cao và có thể xử lý các bộ dữ liệu lớn với hàng triệu hàng và cột.
- Hiệu quả: Nó được thiết kế để tính toán hiệu quả và có thể nhanh chóng huấn luyện các mô hình trên các bộ dữ liệu lớn.
- Linh hoạt: XGBoost hỗ trợ nhiều loại dữ liệu và mục tiêu khác nhau, bao gồm hồi quy, phân loại và các vấn đề xếp hạng.
- Chính quy hóa: Nó tích hợp các kỹ thuật chính quy để tránh hiện tượng quá khớp và cải thiện hiệu suất tổng quát.

G. THUẬT TOÁN N-BEAT

Thuật toán N-BEATS là một mô hình dự đoán chuỗi thời gian mạnh mẽ dựa trên một cấu trúc kiến trúc mạng nơ-ron đa tầng sâu. N-BEATS được thiết kế để có khả năng học được các biến đổi không tuyến tính và phức tạp trong dữ liệu chuỗi thời gian. Điểm nổi bật của N-BEATS là khả năng mở rộng và linh hoạt trong việc tùy chỉnh cấu trúc mạng.

Các đặc điểm chính:

+ Cấu trúc mô hình : Mô hình N-BEATS thường bao gồm một hoặc nhiều khối có thể lặp lại. Mỗi khối bao gồm hai mạng nơ-ron: một mạng BackcastNet và một mạng ForecastNet. Cả hai mạng này thường có cấu trúc tương tự như Feedforward Neural Networks hoặc Convolutional Neural Networks.

+ Hàm mất mát : N-BEATS thường sử dụng hàm mất mát như Mean Absolute Error (MAE) hoặc Mean Squared Error (MSE) giữa dự đoán và giá trị thực tế của chuỗi thời gian.

+ Quá trình huấn luyện : Mô hình N-BEATS được huấn luyện thông qua việc tối ưu hóa hàm mất mát bằng các phương pháp tối ưu hóa như gradient descent hoặc các biến thể của nó, như Adam hoặc RMSprop.

+ Backcast và Forecast:

Backcast : Quá trình dự đoán các giá trị trong quá khứ của chuỗi thời gian. Trong quá trình này, mô hình sẽ sử dụng các giá trị quan sát được trong quá khứ để dự đoán giá trị tại một thời điểm cụ thể trong quá khứ. Quá trình này giúp mô hình học được cách biến đổi và ảnh hưởng của dữ liệu đầu vào đối với các giá trị trong quá khứ, từ đó cung cấp thông tin cần thiết để dự đoán tương lai.

Forecast: Quá trình dự đoán các giá trị trong tương lai của chuỗi thời gian. Trong quá trình này, mô hình sẽ sử dụng các

giá trị quan sát được trong hiện tại và quá khứ để dự đoán giá trị tại một thời điểm cụ thể trong tương lai. Quá trình này cho phép mô hình học được cách dữ liệu hiện tại và quá khứ ảnh hưởng đến các giá trị trong tương lai, từ đó cung cấp dự đoán cho các giá trị trong tương lai của chuỗi thời gian

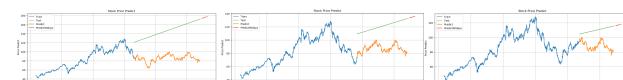
V. THỰC NGHIỆM

A. MÔ HÌNH DỰ ĐOÁN

1) LINEAR REGRESSION



Hình 8: Kết quả với dataset LGLG

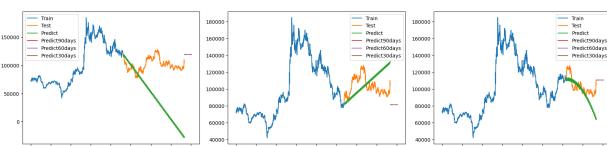


Hình 9: Kết quả của dataset SONY

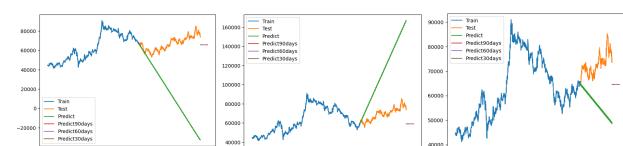


Hình 10: Kết quả với dataset SAMSUNG

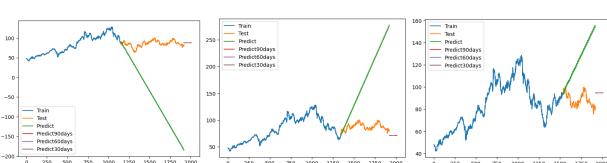
2) ARIMA



Hình 11: ARIMA model's result with LG dataset

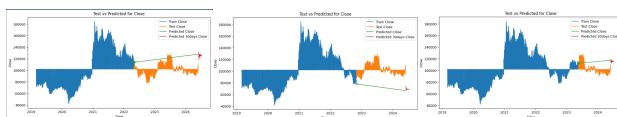


Hình 12: ARIMA model's result with SAMSUNG dataset

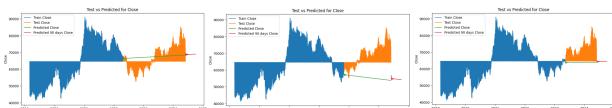


Hình 13: ARIMA model's result with SONY dataset

3) VARMA



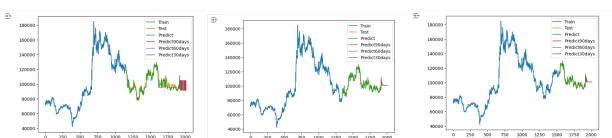
Hình 14: LG 30 DAYS 6:4, 7:3, 8:2



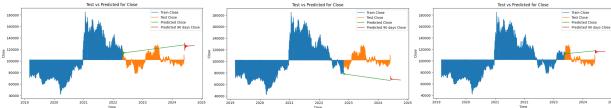
Hình 22: SAMSUNG 90 DAYS 6:4, 7:3, 8:2



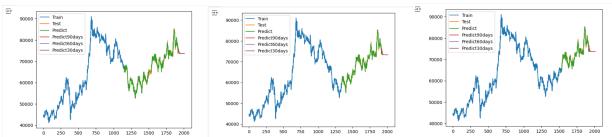
Hình 15: LG 60 DAYS 6:4, 7:3, 8:2



Hình 23: Gradient Boosting Regressor model's result with LG dataset



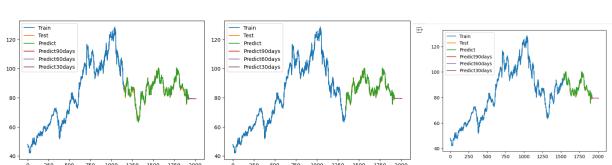
Hình 16: LG 90 DAYS 6:4, 7:3, 8:2



Hình 24: Gradient Boosting Regressor model's result with SAMSUNG dataset



Hình 17: SONY 30 DAYS 6:4, 7:3, 8:2



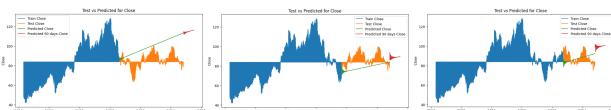
Hình 25: Gradient Boosting Regressor model's result with SONY dataset



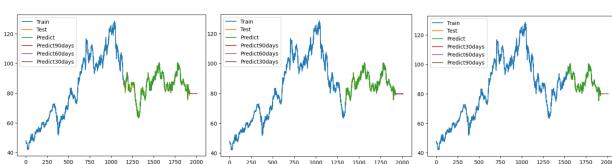
Hình 18: SONY 60 DAYS 6:4, 7:3, 8:2



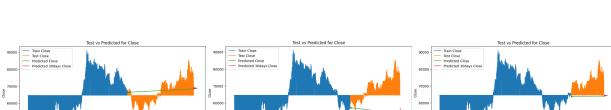
Hình 25: Gradient Boosting Regressor model's result with SONY dataset



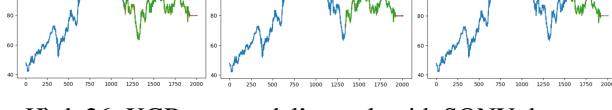
Hình 19: SONY 90 DAYS 6:4, 7:3, 8:2



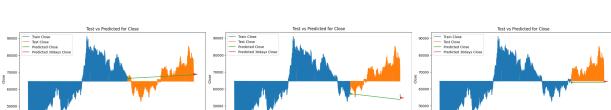
Hình 26: XGBoost model's result with SONY dataset



Hình 20: SAMSUNG 30 DAYS 6:4, 7:3, 8:2

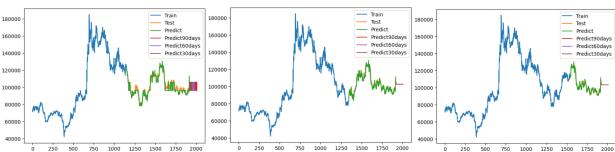


Hình 27: XGBoost model's result with SAMSUNG dataset



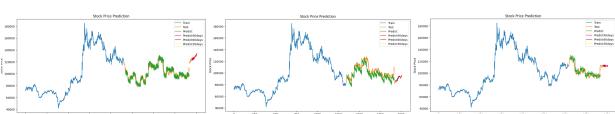
Hình 21: SAMSUNG 60 DAYS 6:4, 7:3, 8:2





Hình 28: XGBoost model's result with LG dataset

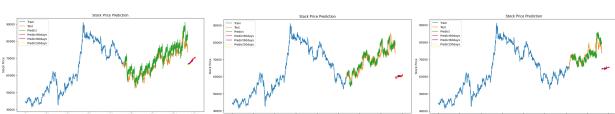
6) N-BEAT



Hình 29: Kết quả với dataset LG

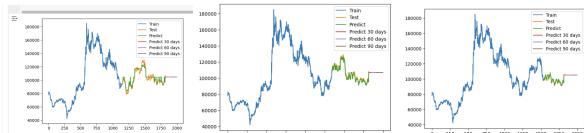


Hình 30: kết quả với dataset SONY

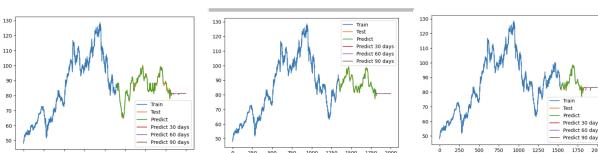


Hình 31: Kết quả với dataset SAMSUNG

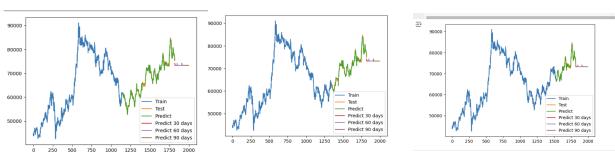
7) RNN + LIGHT GBM



Hình 32: LG dataset 6:4, 7:3, 8:2



Hình 33: Sony dataset 6:4, 7:3, 8:2



Hình 34: SAMSUNG dataset 6:4, 7:3, 8:2

B. ĐÁNH GIÁ VÀ KẾT QUẢ

1) MÔ HÌNH ĐÁNH GIÁ VỚI TẬP DỮ LIỆU LG

Model	Train test 6-4			Train test 7-3			Train test 8-2		
	RMSE	MSE	MAPE	RMSE	MSE	MAPE	RMSE	MSE	MAPE
Varma	101777	10359460720	0,9996	103032	10761010445	0,9999	1040460	10828539410	0,9999
XGBoost	2673,95	7150001,506	1,74%	1267,94	16076469,635	0,85%	887,543	79777,7929	0,62%
Gradient Boosting	2589,68	6706425,803	1,75%	1098,57	1206865,824	0,79%	989,435	978981,5028	0,68%
NBeats	7488,15	56072379,91	5,69%	8198,36	67213054,23	6,44%	6212,06	38599716	4,10%
RNN	0,03191	0,001017948	6,73%	0,01531	0,000234325	2,38%	0,01332	0,0001777509	2,21%
LightGBM	2204,35	8895180,596	1,70%	2004,18	4097321,9	1,36%	2131,28	4714665,1783	1,75%
LSTM	101749	10352788600	246893	104459	10911619739	236914	100794	10159507213	254341
Linear Regression	0,59781	0,357370885	1,47039	0,33286	0,110796822	0,79841	0,23409	0,054798288	0,54162
ARIMA	71218,81	5072111055	57,35%	19704,8	388277402,9	16,79%	13139,6	172649827,8	9,96%

2) MÔ HÌNH ĐÁNH GIÁ VỚI TẬP DỮ LIỆU SONY

Model	Train test 6-4			Train test 7-3			Train test 8-2		
	RMSE	MSE	MAPE	RMSE	MSE	MAPE	RMSE	MSE	MAPE
Varma	95,9466	7386,9247	0,9996	87,6592	7684,1325	0,9998	88,7191	7871,0743	0,9997
XGBoost	0,71893	0,516553494	0,63%	0,43119	0,185924811	0,38%	0,36364	0,132235006	0,32%
Gradient Boosting	0,66987	0,448719472	0,59%	0,45624	0,208157128	0,42%	0,43111	0,185958354	0,39%
NBeats	4,1997	24,96997073	4,52%	4,52845	20,50684615	4,09%	5,05112	25,51380042	4,72%
RNN	0,0216	0,00046664	3,50%	0,01823	0,000332423	2,47%	0,03879	0,001504276	5,63%
LightGBM	1,70956	2,922593726	1,57%	1,45923	2,129363742	1,22%	1,41142	1,992102775	1,18%
RNN	0,72179	0,520980232	1,45665	0,39742	0,157939188	0,76044	0,26178	0,06852931	0,48473
Arima	156,1131	24376,8089	152,97%	104,69	10959,94322	99,65%	41,6921	1738,22823	42,17%
LSTM	86,5793	7495,9837	174,713	87,6018	7674,0832	163,707	86,8366	7540,5961	167,732

3) MÔ HÌNH ĐÁNH GIÁ VỚI TẬP DỮ LIỆU SAMSUNG

Model	Train test 6-4			Train test 7-3			Train test 8-2		
	RMSE	MSE	MAPE	RMSE	MSE	MAPE	RMSE	MSE	MAPE
Varma	67316,5	4531050285	0,9997	69470,9	482620690	0,9999	72979,9	5326068095	0,9998
XGBoost	435,492	189652,8507	0,45%	457,289	209113,339	0,49%	435,492	189652,8507	0,45%
Gradient Boosting	459,053	210729,636	0,50%	487,65	237802,136	0,52%	475,713	226302,9847	0,48%
NBeats	4050,63	16407620,96	5,18%	3553,16	12624920,3	3,93%	3196,83	10219720,75	3,25%
RNN	0,02114	0,000447047	3,17%	0,02108	0,000444548	2,54%	0,02279	0,00051952	2,51%
LightGBM	1003,66	1007324,53	1,15%	1120,56	125568,257	1,10%	1134,23	1727189,462	1,30%
Arima	61234,6	3749681296	71,29%	51034,3	2604497386	61,62%	18099,2	3275914487	21,68%
Linear Regression	0,60938	0,370128443	1,27229	0,27523	0,07575102	0,53034	0,08421	0,007091839	0,11978
LSTM	67935,5	4615232856	134645	71130,6	5056718944	121259	73551,4	5409808100	113401

VI. KẾT LUẬN

A. KẾT LUẬN TỔNG THỂ

Dự đoán giá cổ phiếu vẫn là một nhiệm vụ phức tạp và đầy thách thức, và việc tìm ra một mô hình chính xác toàn diện vẫn còn khó khăn. Nghiên cứu này đã khám phá các phương pháp thống kê, học máy, và học sâu để dự đoán giá cổ phiếu của các gã khổng lồ công nghệ hàng đầu: LG, Samsung và Sony. Mỗi phương pháp mang lại những hiểu biết và lợi thế độc đáo, nhấn mạnh tầm quan trọng của việc xem xét các phương pháp đa dạng cho phân tích toàn diện.

Nhìn chung, nghiên cứu này chứng minh rằng mặc dù không có mô hình nào có thể dự đoán chính xác giá cổ phiếu, việc sử dụng kết hợp các phương pháp thống kê, học máy và học sâu có thể cung cấp những hiểu biết quý báu và cải thiện độ chính xác của dự báo. Tuy nhiên, điều quan trọng là phải xem xét các hạn chế của từng phương pháp và diễn giải kết quả một cách thận trọng, thừa nhận sự bất định vốn có liên quan đến dự đoán thị trường chứng khoán.

B. THÁCH THỨC

Dữ liệu là yếu tố quan trọng nhất trong phân tích thống kê và dự đoán. Một tập dữ liệu phù hợp cần phải đáp ứng các yêu cầu sau: Đầu đủ và chính xác cao, đủ lớn để mô hình có thể học và dự đoán chính xác, liên quan chặt chẽ đến đối tượng cần dự đoán. Tuy nhiên, việc tìm kiếm một tập dữ liệu phù hợp thường không dễ dàng. Dữ liệu công khai thường không đáp ứng được các yêu cầu trên hoặc không phù hợp với chủ đề của dự án.

Sau khi thu thập dữ liệu, tôi bắt đầu mã hóa các dự đoán bằng các mô hình. Tôi đã chọn một số mô hình phổ biến như Gradient

Boosting, VARMA, XGBoost, và mô hình Deep Learning như NBEATS, Light GBM kết hợp RNN.

Tuy nhiên, việc xử lý các mô hình này không đơn giản. Nhóm cần phải có kiến thức về lý thuyết thống kê, lý thuyết học máy và kỹ năng lập trình. Chúng em đã dành nhiều thời gian nghiên cứu cách các mô hình này hoạt động.

C. KẾ HOẠCH TƯƠNG LAI

Để nâng cao độ chính xác của dự đoán, chúng em dự định sẽ:

- Tinh chỉnh dữ liệu: Nâng cao kỹ năng thu thập và xử lý dữ liệu thông qua các phương pháp mới nhất, cung cấp cho các mô hình của chúng tôi thông tin sạch hơn và đáng tin cậy hơn.
- Nâng cấp thuật toán: Khám phá các phương pháp tiên tiến như Học sâu (Deep Learning) và Học tăng cường (Reinforcement Learning) để có các mô hình dự đoán tinh vi và chính xác hơn.
- Đánh giá thông minh hơn: Áp dụng các chỉ số tiêu chuẩn trong ngành như MASE, MAPE và SMAPE để đánh giá toàn diện hiệu quả của các mô hình.
- Hợp tác: Tích cực hợp tác với các chuyên gia trong các cộng đồng chuyên môn để chia sẻ kinh nghiệm và mở rộng nền tảng kiến thức của chúng tôi.

NHÌN NHẬN

Nhóm chúng em xin gửi lời cảm ơn sâu sắc đến thầy Nguyễn Đình Thuân và thầy Nguyễn Minh Nhựt đã trang bị cho chúng em có được những kiến thức căn bản vững chắc để có thể thực hiện đồ án lần này. Ngoài ra, xin cảm ơn đến tất cả sự đóng góp của các thành viên trong nhóm, những người đã chăm chỉ và hoàn thành nhiệm vụ của mình đúng hạn để có thể hoàn thiện một cách đầy đủ nhất. Trong quá trình nghiên cứu và thực hiện đồ án, nhóm chúng em đã kết hợp giữa những kiến thức căn bản và những gì được thầy trao đổi và truyền đạt trên lớp để cố gắng hoàn thiện đồ án tốt nhất có thể. Tuy nhiên, trong đồ án của chúng em vẫn còn một vài thiếu sót nhưng nó là kết quả của sự nỗ lực và cố gắng của các thành viên trong nhóm cũng như sự giúp đỡ từ bạn bè và Thầy. Nhóm rất mong nhận được sự góp ý từ phía Thầy nhằm rút ra được những kinh nghiệm quý báu và hoàn thiện vốn kiến thức của mình để có thể tiếp tục hoàn thành được những đồ án khác ở trong tương lai. Lời cuối cùng, nhóm chúng em xin chúc quý Thầy thật nhiều sức khỏe và niềm vui để có thể tiếp tục giảng dạy và truyền đạt thật nhiều kiến thức bổ ích đến cho những sinh viên khác. Chúng em xin chân thành cảm ơn!

TÀI LIỆU THAM KHẢO

[1] : Mojtaba Nabipour; Pooyan Nayyeri; Hamed Jabani; Shahab S.; Amir Mosavi , "Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis", 2020.

[2]: Salah Bouktif; Ali Fiaz; Mamoun Awad , "Augmented Textual Features-Based Stock Market Prediction", 2020

[3] : Sondo Kim; Seungmo Ku; Woojin Chang; Jae Wook Song , "Augmented Textual Features-Based Stock Market Prediction", 2020

[4]: Warsono, Edwin Russel, Wamiliana*, Widiarti, Mustofa Usman, "Modeling and Forecasting by the Vector Autoregressive Moving Average Model for Export of Coal and Oil Data

(Case Study from Indonesia over the Years 2002-2017)". International Journal of Energy Economics and Policy, 2019

[5]: Yugesh Verma, "A Guide to VARMA with Auto ARIMA in Time Series Modelling". analyticsindiamag.com, 2021

[6]: Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", 31st Conference on Neural Information Processing Systems (NIPS 2017).

[8]: Yi Li, Lei Chen, Cuiping Sun, Guoxu Liu, Chunlei Chen, "Accurate Stock Price Forecasting Based on Deep Learning and Hierarchical Frequency Decomposition", 2024.

[9]: Yankai Sheng, Yuanyu Qu, Ding Ma, "Stock price crash prediction based on multimodal data machine learning models", 2024

[10]: Qifei Zhou; Hucheng Liu; Weiping Li; Tong Mo; Bo Wu, "Bilateral Autotrading Framework for Stock Prediction", 2021