

Phân tích, dự đoán giá chứng khoán của 3 công ty SAMSUNG, SONY, LG bằng mô hình thống kê, máy học và học sâu

Trần Thị Mỹ Xoan - 21522815
Lê Anh Tuấn Dũng - 21521974
Lê Thị Ánh Hồng - 21520245
Nguyễn Thị Mai Liên - 21522283
Đỗ Sĩ Đạt - 21521932

Ngày 6 tháng 6 năm 2024

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 thu về được lợi nhuận cao hơn khi đầu tư. 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.

1 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.

2 Related word

[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.

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) [4]. 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.

3 PHƯƠNG PHÁP

3.1 DATASET

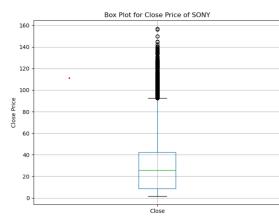
Trọng tâm của bài viết này là dự đoán giá cổ phiếu. Vì vậy, cả ba tập dữ liệu chúng tôi sử dụng trong nghiên cứu đều là giá cổ phiếu của các công ty lớn. Các công ty này là SONY, SAMSUNG và LG. Dữ liệu của SAMSUNG và LG được thu thập từ ngày 25 tháng 4 năm 2002 đến ngày 11 tháng 1 năm 2024. Dữ liệu của SONY được thu thập từ ngày 26 tháng 7 năm 1974.

Mô tả từng cột trong dữ liệu :

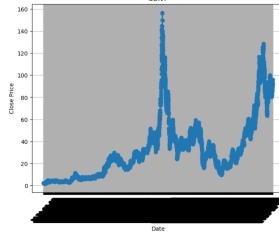
- + Date : Ngày giao dịch.
- + Open: Giá mở cửa của cổ phiếu vào ngày giao dịch.
- + High: Giá cao nhất của cổ phiếu trong ngày giao dịch.
- + Low: Giá thấp nhất của cổ phiếu trong ngày giao dịch.
- + Close: Giá đóng cửa của cổ phiếu vào ngày giao dịch.
- + Adj Close: Giá đóng cửa điều chỉnh của cổ phiếu. Giá này thường đã được điều chỉnh để phản ánh các biến động khác nhau như cổ tức, chia cổ tức, và phát hành cổ phiếu mới.
- + Volume: Khối lượng giao dịch của cổ phiếu trong ngày, hay số lượng cổ phiếu đã được giao dịch trong ngày đó.

Bảng 1: SONY, SAMSUNG, LG's Descriptive Statistics

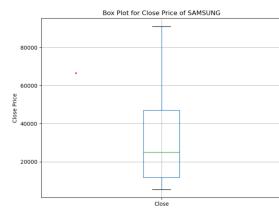
	SONY	SAMSUNG	LG
Count	1920	1920	1920
Mean	83.842011	83.842011	101952.316347
Std	15.828603	15.828603	26596.447698
Min	42.030000	42.030000	41850
25%	79.987500	79.987500	87200
50%	83.842011	83.842011	101952.316347
75%	90.442500	90.442500	114000
Max	128.59	128.590000	185000



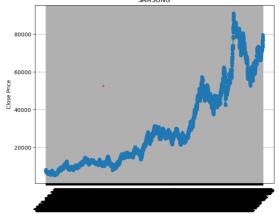
Hình 1: SONY stock price's boxplot



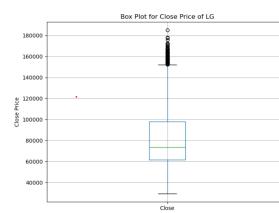
Hình 2: SONY stock price's histogram



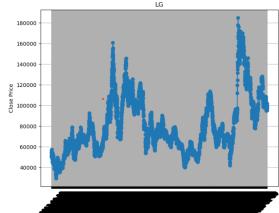
Hình 3: SAMSUNG stock price's boxplot



Hình 4: SAMSUNG stock price's histogram



Hình 5: LG stock price's boxplot



Hình 6: LG stock price's histogram

3.2 CHỈ SỐ ĐÁNH GIÁ MÔ HÌNH

3.2.1 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:

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

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, việc vẽ biểu đồ là một cách trực quan để xem xu hướng chung của dữ liệu. MSE cho bạn biết mức độ gần của đường hồi quy với tập hợp các điểm dữ liệu. Cụ thể, MSE đo khoảng cách từ các điểm dữ liệu đến đường hồi quy (các khoảng cách này là "sai số") và bình phương chúng. Việc bình phương sai số là rất quan trọng vì nó loại bỏ các dấu âm và làm tăng trọng số của các sai số lớn.

Để giảm thiểu MSE, mô hình cần dự đoán chính xác hơn, tức là gần với dữ liệu thực tế hơn. Một ví dụ về hồi quy tuyến tính sử dụng phương pháp này là phương pháp bình phương nhỏ nhất. Phương pháp này đánh giá sự phù hợp của mô hình hồi quy tuyến tính với tập dữ liệu hai biến. Tuy nhiên, nó có giới hạn liên quan đến phân phối dữ liệu đã biết.

MSE càng thấp thì dự báo càng tốt.

3.2.2 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:

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

trong đó:

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

Ảnh hưởng của mỗi lỗi đối với RMSE tỷ lệ với kích thước của lỗi bình phương, do đó các sai số lớn hơn có ảnh hưởng lớn đến RMSE một cách không cân xứng. Vì lý do này, RMSE nhạy cảm với các yếu tố ngoại lai. RMSE thường được sử dụng trong khí hậu học, dự báo, và phân tích hồi quy để xác minh kết quả thực nghiệm.

Khi các quan sát và dự báo được chuẩn hóa, RMSE có mối quan hệ trực tiếp với hệ số tương quan. Ví dụ, nếu hệ số tương quan là 1, RMSE sẽ bằng 0 vì tất cả các điểm nằm trên đường hồi quy (không có sai số).

RMSE luôn không âm, và giá trị 0 (hầu như không bao giờ đạt được trong thực tế) chỉ ra sự phù hợp hoàn hảo với dữ liệu. Nói chung, RMSE thấp hơn cho thấy mô hình tốt hơn so với RMSE cao hơn.

3.2.3 Mean Absolute Percentage Error(MAPE)

MAPE (Mean Absolute Percentage Error) 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:

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

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.

3.3 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.

3.3.1 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 $x^{(t)}$.

3.3.2 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.

3.4 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.

3.5 THUẬT TOÁN VARMA

- Thuật toán VARMA mô tả mối quan hệ giữa các biến thông qua cả khía cạnh tự hồi quy (AR) và trung bình di chuyển (MA).

- Autoregressive (AR): Trong mô hình AR, giá trị của biến tại một thời điểm được dự đoán dựa trên các giá trị trước đó của chính nó và của các biến khác. Một mô hình AR(p) sử dụng p giá trị trước đó để dự đoán giá trị hiện tại của biến dựa trên p giá trị trước đó của chính nó. Cụ thể, một mô hình AR(p) có dạng:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \epsilon_t$$

Trong đó:

+ Y_t : Biến phụ thuộc tại thời điểm t

+ c là hằng số.

+ $\phi_1, \phi_2, \dots, \phi_p$ là các tham số của mô hình, thể hiện mức độ tương quan giữa giá trị hiện tại và các giá trị trước đó.

+ $Y_{t-1}, Y_{t-2}, \dots, Y_{t-p}$ là các giá trị của biến tại các thời điểm trước đó.

+ ϵ_t : Sai số dự báo tại thời điểm t

- Moving-average (MA): Trong mô hình MA, giá trị của biến tại một thời điểm được dự đoán dựa trên các sai số dự báo trước đó. Một mô hình MA(q) sử dụng q sai số trước đó để dự đoán giá trị hiện tại. Cụ thể, một mô hình MA(q) có dạng:

$$Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q}$$

Rõ ràng:

+ Y_t : là giá trị của biến tại thời điểm t .

+ μ là giá trị kỳ vọng của biến.

+ ϵ_t là sai số ngẫu nhiên (error term) tại thời điểm t thường được giả định có phân phối chuẩn với mean bằng 0 và độ lệch chuẩn không đổi.

+ $\theta_1, \theta_2, \dots, \theta_q$ là các tham số của mô hình, thể hiện mức độ ảnh hưởng của các sai số dự báo trước đó lên giá trị hiện tại.

+ $\epsilon_{t-1}, \epsilon_{t-2}, \dots, \epsilon_{t-q}$ là các sai số dự báo trước đó.

- Mô hình VARMA kết hợp cả hai phương trình tự hồi quy và trung bình di chuyển để mô hình hóa dữ liệu chuỗi thời gian đa biến. Các hệ số ϕ và θ được ước lượng từ dữ liệu và được sử dụng để dự đoán giá trị tiếp theo của chuỗi thời gian.

3.6 THUẬT TOÁN Gradient Boosting egressor

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.

3.7 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 khái quát 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:

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

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.

3.8 THUẬT TOÁN NBEAT

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

4 THỰC NGHIỆM

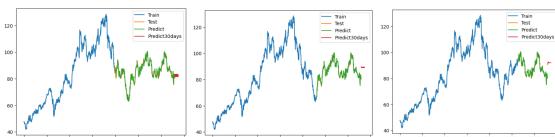
4.1 MÔ HÌNH DỰ ĐOÁN

4.1.1 RNN + LIGHTGBM

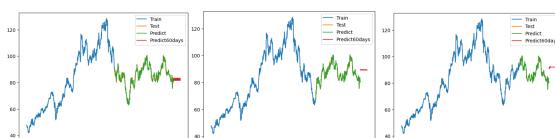
4.1.2 VARMA

4.1.3 GRADIENT BOOSTING

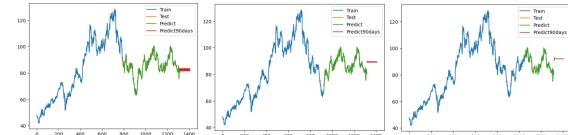
4.1.4 XGBOOST



Hình 8: SONY 30 DAYS 6:4, 7:3, 8:2



Hình 9: SONY 60 DAYS 6:4, 7:3, 8:2



Hình 10: SONY 90 DAYS 6:4, 7:3, 8:2

4.1.5 NBEAT



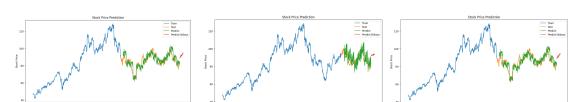
Hình 11: LG 30 DAYS 6:4, 7:3, 8:2



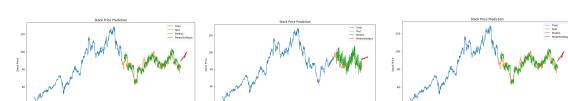
Hình 12: LG 60 DAYS 6:4, 7:3, 8:2



Hình 13: LG 90 DAYS 6:4, 7:3, 8:2



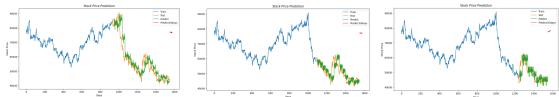
Hình 14: SONY 30 DAYS 6:4, 7:3, 8:2



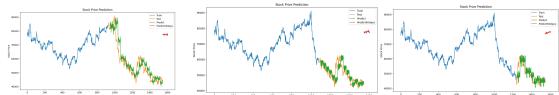
Hình 15: SONY 60 DAYS 6:4, 7:3, 8:2



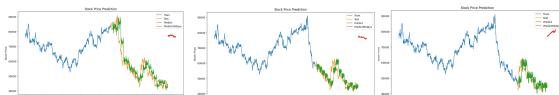
Hình 16: SONY 90 DAYS 6:4, 7:3, 8:2



Hình 17: SAMSUNG 30 DAYS 6:4, 7:3, 8:2



Hình 18: SAMSUNG 60 DAYS 6:4, 7: 3, 8:2



Hình 19: SAMSUNG 90 DAYS 6:4, 7: 3, 8:2

4.2 THANG ĐO VÀ KẾT QUẢ

4.2.1 ĐÁNH GIÁ MÔ HÌNH VỚI DATASET LG

4.2.2 ĐÁNH GIÁ MÔ HÌNH VỚI DATASET SONY

4.2.3 ĐÁNH GIÁ MÔ HÌNH VỚI DATASET SAMSUNG

5 REFERENCE

[1]: "Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis"

Link: <https://ieeexplore.ieee.org/document/9165760>

[2]: "Augmented Textual Features-Based Stock Market Prediction"

Link: <https://ieeexplore.ieee.org/document/9016182>

[3]: "Predicting the Direction of US Stock Prices Using Effective Transfer Entropy and Machine Learning Techniques"

Link: <https://ieeexplore.ieee.org/document/9119388>

[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)"