Mô hình thống kê chuỗi thời gian giao dịch sàn chứng khoán

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 15 tháng 5 năm 2024

Tóm tắt nội dung: Bài báo này trình bày phương pháp dự đoán giá cổ phiếu của ba công ty lớn Samsung, LG và Sony trong 30 ngày tiếp theo. Cụ thể, chúng tôi sử dụng ba thuật toán là ARIMA, Linear Regression và VARMA để dự đoán. Dự đoán thị trường chứng khoán, dự đoán giá cổ phiếu, thống kê.

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.Là những công ty chủ chốt trong ngành, 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.

Nghiên cứu này nhằm mục đích dự đoán giá cổ phiếu của các công ty có ảnh hưởng bằng cách sử dụng 3 thuật toán ARIMA, Linear Regression, VARMA để đưa ra dựa 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 dầu tư trong việc định hướng bối cảnh năng động của lĩnh vực công nghệ.

2 Related word

Dự đoán giá cổ phiếu là một lĩnh vực nghiên cứu quan trọng trong lĩnh vực tài chính và khoa học máy. Có nhiều nghiên cứu đã tập trung vào việc áp dụng các thuật toán thống kê và học máy để dự đoán giá cổ phiếu của các công ty lớn như Samsung, LG và Sony. Trong phần này, chúng tôi xem xét một số công trình có liên quan về việc dự đoán giá cổ phiếu của các công ty này bằng cách sử dụng các phương pháp như VARMA, ARIMA và Linear Regression.

Agus Tri Haryono, Riyanarto Sarno và Kelly Rossa Sungkono (2023) đã tiến hành một nghiên cứu về hiệu suất của mô hình FEDformer, một biến thể của mô hình transformer, trong dự đoán giá cổ phiếu. Kết quả cho thấy rằng FEDformer vượt trội so với các mô hình khác như AutoFormer, Informer và Reformer. Đặc biệt, hiệu suất của FEDformer cũng được chứng minh cao hơn so với các mô hình truyền thống như ARIMA và LSTM.

Trong một nghiên cứu khác, Sasha S. Yamada và Ogulcan E. Orsel (2022) so sánh hiệu suất giữa bộ lọc Kalman tuyến tính và các biến thể khác của mạng nơ-ron hồi quy dài hạn (LSTM) trong dự đoán giá cổ phiếu. Kết quả cho thấy rằng hiệu suất của mỗi mô hình bị ảnh hưởng đáng kể bởi sự biến động của cổ phiếu được dự đoán. Đối với các cổ phiếu ít biến

động, bộ lọc Kalman tuyến tính có thể dự đoán giá cổ phiếu vào ngày tiếp theo với độ chính xác rất hợp lý. Tuy nhiên, sai số này tăng đáng kể đối với các cổ phiếu biến động lớn hơn, khiến cho các kiến trúc LSTM trở thành lựa chọn phù hợp hơn trong các tình huống như vậy.

Thuật toán VARMA (Vector Autoregressive Moving-Average) được sử dụng để dự đoán chuỗi thời gian đa biến. Bằng cách kết hợp tự hồi quy và trung bình di chuyển, VARMA có khả năng mô hình hóa mối quan hệ phức tạp giữa các biến tài chính. Điều này làm cho nó trở thành một công cụ hiệu quả cho việc dự đoán giá cổ phiếu trong môi trường thị trường đa biến.

Linear Regression là một thuật toán phổ biến trong lĩnh vực dự đoán giá cổ phiếu. Nó xây dựng một mô hình tuyến tính giữa các biến độc lập (ví dụ: các chỉ số kỹ thuật, dữ liệu cơ bản) và giá cổ phiếu. Mặc dù đơn giản, nhưng Linear Regression có thể cung cấp những hiểu biết quan trọng về mối quan hệ giữa các yếu tố ảnh hưởng đến giá cổ phiếu.

ARIMA (Autoregressive Integrated Moving Average) là một mô hình thống kê mạnh mẽ được sử dụng để dự đoán chuỗi thời gian. Bằng cách kết hợp tự hồi quy và trung bình di chuyển, ARIMA có thể mô hình hóa và dự đoán xu hướng và dao động trong giá cổ phiếu. Đặc biệt, ARIMA phù hợp cho các chuỗi thời gian có xu hướng và dao động không đồng nhất.

Tóm lại, các nghiên cứu trên cung cấp cái nhìn tổng quan về hiệu suất của các thuật toán và mô hình khác nhau trong dự đoán giá cổ phiếu của các công ty lớn như Samsung, LG và Sony. Sự đa dạng của các phương pháp này cung cấp cơ sở cho sự phát triển và cải thiện của các phương pháp dự đoán trong tương lai.

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.

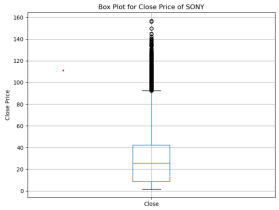
+ Close: Giá đóng cửa của cổ phiếu vào ngày giao dịch.

+ 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.

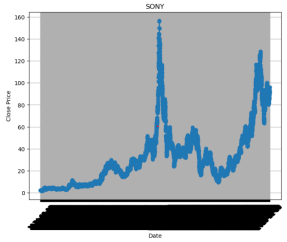
+ Volume: Khối lượng giao dịch của cổ phiếu trong ngày, hay số lượng cổ phiếu đã được giao dịch trong ngày đó.

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

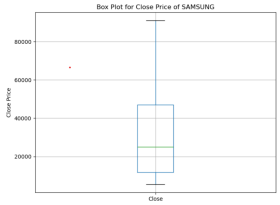
	SONY	SAMSUNG	LG
Count	12473	5502	5493
Mean	32,000373	30510,212650	80578,825496
Std	26,751183	21808,668213	26952,453202
Min	1,590909	5390	29190,132813
25%	9,659091	11820	61618,164063
50%	26,389999	25000	73400
75%	42,843750	47000	97900
Max	156,75	91000	185000



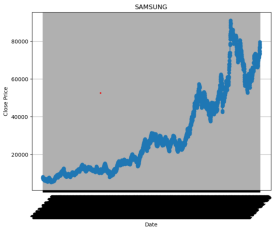
Hình 1: SONY stock price's boxplot



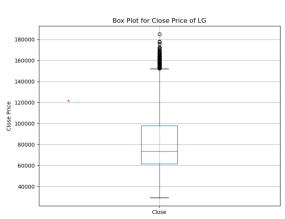
Hình 2: SONY stock price's histogram



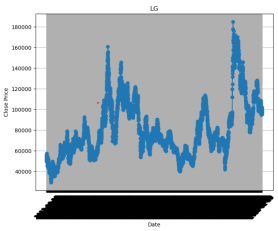
Hình 3: SAMSUNG stock price's boxplot



Hình 4: SAMSUNG stock price's histogram



Hình 5: LG stock price's boxplot



Hình 6: LG stock price's histogram

3.2 THUẬT TOÁN ARIMA

- ARIMA tích hợp của 2 quá trình: quá trình tự hồi quy bậc p–AR(p) và quá trình trung bình trượt bậc q–MA(q). Mặt khác, cần phải dùng tích hợp sai phân I(d) (hay còn gọi là toán tử trễ) để làm cho chuỗi thời gian trở thành chuỗi dừng.

*Chuỗi dừng: một chuỗi thời gian có tính dừng là một chuỗi có các giá trị như trung bình, phương sai và các giá trị tương quan (autocorrelation) của quá trình không thay đổi theo thời gian và không bao hàm yếu tố xu thế.

* Để kiểm tra tính dừng của chuỗi thời gian, ta có hai phương pháp kiểm định phổ biến: Kiểm định Dickey Fuller3 (DF) và Dickey Fuller cải tiến (ADF4).

+ Quá trình tự hồi quy bậc p – AR(p): quá trình tìm mối quan hệ giữa dữ liệu hiện tại và p dữ liệu trước đó (lag).
+ Quá trình trung bình vượt bậc q – MA(q): quá trình tìm mối quan hệ giữa dữ liệu hiện tại và d phần lỗi trước đó.
+ Tích hợp sai phân I(d): so sánh sự khác nhau giữa d quan sát (Hiệu giữa giá trị hiện tại và d giá trị trước đó) để biến một chuỗi thành chuỗi dừng.
- Với các giá trị không âm của 3 tham số trên cho biết mô hình ARIMA cụ thể nào được sử dụng:

• p : số lượng giá trị trong quá khứ có trong mô hình AR
$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \epsilon_t$

Trong đó :
+ c là hằng số
+ $\phi_1, \ldots, \phi_p$: là các tham số
+ ϵ_t là nhiễu trắng

• d : số lần chuỗi thời gian bị sai khác.

• q : số lỗi dự báo trong quá khứ có trong mô hình MA hoặc kích thước của cửa sổ trung bình động. Nó được đặt tên là mô hình MA vì mỗi y t có thể được coi là trung bình động có trọng số của các lỗi dự báo trong quá khứ

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

⇒ Phương trình mô hình đầy đủ của ARIMA(p, d, q) là:

$$\nabla Y_t = c + \phi_1 \nabla Y_{t-1} + ... + \phi_p \nabla Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q}$$

Trong đó ∇Y_t là chuỗi thời gian khác biệt, có thể chênh lệch nhiều lần.
- Mô hình ARIMA không phải là mô hình dự báo hoàn hảo ứng với bất kỳ dữ liệu chuỗi thời gian nào và chỉ hoạt động tốt nhất nếu dữ liệu phụ thuộc nhiều vào thời gian và dự báo dạng điểm thời gian những dữ liệu ngẫu nhiên thường ít hoạt động với mô hình ARIMA.

3.3 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:
$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \epsilon_t$
Trong đó:
+ Y_t: Biến phụ thuộc tại thời điểm t
+ c là hằng số.
+ $\phi_1, \phi_2, ..., \phi_p$ là các tham số của mô hình, thể hiện mức độ tương quan giữa giá trị hiện tại và các giá trị trước đó.
+ $Y_{t-1}, Y_{t-}, ..., Y_{t-p}$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:

$Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_q \epsilon_{t-q}$ TRong đó:

+ 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.
+ $\theta_1, \theta_2, ..., \theta_q$ 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.
+ $\epsilon_{t-1}, \epsilon_{t-2}, ..., \epsilon_{t-q}$ 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.

3.4 THUẬT TOÁN LINEAR REGRESSION

Thuật toán Linear Regression thường được sử dụng trong các bài toán dự đoán và hồi quy. Quá trình huấn luyện của thuật toán này là tìm ra các tham số tốt nhất cho đường thẳng hoặc siêu phẳng sao cho tổng bình phương sai số giữa giá trị dự đoán và giá trị thực tế là nhỏ nhất. Phương pháp này thường được sử dụng cho các bài toán có dữ liệu liên tục.

Đặc điểm của Linear Regression bao gồm:

+ Đơn giản và dễ hiểu.

+ Dễ triển khai và tính toán.

+ Phù hợp cho dữ liệu có mối quan hệ tuyến tính đơn giản.

+ Có thể được mở rộng để xử lý các mô hình phức tạp hơn thông qua các biến phụ thuộc phi tuyến tính hoặc sử dụng kỹ thuật biến đổi dữ liệu.

Công thức :

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$

Trong đó :

+ Y : Đây là biến phụ thuộc (hoặc biến mục tiêu), là giá trị chúng ta muốn dự đoán hoặc giải thích bằng các biến độc lập khác

+ $X_1, X_2, \ldots, X_p$: Là các biến độc lập, là các đặc trưng hoặc biến giải thích mà chúng ta sử dụng để dự đoán giá trị của Y

+ β_0 : Là hệ số chặn của đường hồi quy, thường được gọi là hệ số chặn hoặc hệ số độc lập. Nó cho biết giá trị của Y khi tất cả các biến độc lập đều bằng 0.

+ $\beta_1, \beta_2, \ldots, \beta_p$: Là các hệ số hồi quy của các biến độc lập tương ứng $X_1, X_2, \ldots, X_p$. húng cho biết mức độ ảnh hưởng của mỗi biến độc lập lên biến phụ thuộc Y

+ ϵ : Là sai số ngẫu nhiên, biểu thị những yếu tố không được mô hình hóa hoặc không biết đến có thể ảnh hưởng đến biến phụ thuộc Y

3.5 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.

3.6 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:

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

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.

3.7 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.

Porecast: 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