Đánh giá hiệu quả của các phương pháp dự đoán giá cổ phiếu bằng các thuật toán học máy

|  |  |  |  |
| --- | --- | --- | --- |
| 1st Trần Văn Quang  Lớp: IS403.N21  Trường đại học Công nghệ thông tin  20520722@gm.uit.edu.vn | 2nd Lê Công Thành  Lớp: IS403.N21  Trường đại học Công nghệ thông tin  20520763@gm.uit.edu.vn | 3rd Võ Hoàng Phúc  Lớp: IS403.N21  Trường đại học Công nghệ thông tin  20520698@gm.uit.edu.vn | 4th Nguyễn Hoàng Nhật Lớp: IS403.N21  Trường đại học Công nghệ thông tin  20520673@gm.uit.edu.vn |

*Tóm tắt* — **Việc dự đoán giá cổ phiếu là một vấn đề quan trọng trong lĩnh vực tài chính và đầu tư. Trong nghiên cứu này, chúng tôi sử dụng các thuật toán học máy khác nhau để dự đoán giá cổ phiếu, bao gồm Linear Regression, ARIMA, SARIMAX, RNN, LSTM, GRU, CNN-LSTM, DLM, BDLM và GBT. Chúng tôi cũng sử dụng các kỹ thuật phân tích dữ liệu để phân tích chuỗi thời gian, phân tích tương quan và dự đoán đa biến. Mục tiêu của nghiên cứu này là so sánh hiệu suất của các thuật toán khác nhau và xác định phương pháp dự đoán giá cổ phiếu hiệu quả nhất. Kết quả cho thấy rằng các thuật toán RNN, LSTM và GRU đạt hiệu suất tốt nhất trong việc dự đoán giá cổ phiếu. Ngoài ra, kết quả cũng cho thấy rằng việc kết hợp CNN-LSTM có thể cải thiện độ chính xác và độ tin cậy của dự đoán.**

Từ khóa – dự đoán giá cổ phiếu, thuật toán, Linear Regression, ARIMA, SARIMAX, RNN, LSTM, GRU, CNN-LSTM, DLM, BDLM, GBT, phân tích dữ liệu, chuỗi thời gian, phân tích tương quan, hiệu suất, độ chính xác, độ tin cậy, kết hợp thuật toán, nghiên cứu, hướng phát triển.

# GIỚI THIỆU

Dự đoán giá cổ phiếu là một trong những vấn đề quan trọng trong lĩnh vực tài chính và đầu tư. Với sự phát triển của công nghệ và các thuật toán học máy, việc sử dụng các phương pháp dự đoán dựa trên dữ liệu đã trở nên phổ biến và đạt được kết quả khả quan trong thời gian gần đây. Trong nghiên cứu này, chúng tôi tập trung vào việc đánh giá và so sánh hiệu suất của một loạt các thuật toán học máy khác nhau để dự đoán giá cổ phiếu, bao gồm Linear Regression, ARIMA, SARIMAX, RNN, LSTM, GRU, CNN-LSTM, DLM, BDLM và GBT. Các thuật toán này đều có tính ứng dụng rộng trong việc dự đoán chuỗi thời gian và đã được chứng minh hiệu quả trong nhiều lĩnh vực khác nhau. Mục tiêu của nghiên cứu này là tìm ra phương pháp dự đoán giá cổ phiếu hiệu quả nhất và cung cấp cho các nhà đầu tư và các chuyên gia tài chính một cái nhìn tổng quan về hiệu suất của các phương pháp dự đoán bằng các thuật toán học máy. [1] Để đạt được mục tiêu này, chúng tôi sử dụng các kỹ thuật phân tích dữ liệu để phân tích chuỗi thời gian, phân tích tương quan và dự đoán đa biến. Các phần còn lại của bài báo bao gồm phân tích chi tiết về các thuật toán được sử dụng, kết quả thực nghiệm và phân tích, và cuối cùng là các kết luận và hướng phát triển tiếp theo. Kết quả của nghiên cứu này sẽ cung cấp cho các nhà đầu tư và các chuyên gia tài chính một phần cơ sở để đưa ra quyết định đầu tư thông minh và hiệu quả.

# NGHIÊN CỨU LIÊN QUAN

Trong phần này, chúng ta sẽ thảo luận về các nghiên cứu trước đây đã được thực hiện về dự đoán giá cổ phiếu bằng sử dụng các thuật toán học máy hồi quy và học sâu.

Trong bài báo của D. Bhuriya, G. Kaushal, A. Sharma và U. Singh, họ đã cách áp dụng hồi quy tuyến tính để dự báo hành vi của tập dữ liệu TCS, họ chứng minh rằng phương pháp đề xuất của họ là tốt nhất để so sánh với các phương pháp kỹ thuật hồi quy khác và các cổ đông có thể đầu tư một cách an toàn dựa trên phương pháp đó. [2]

Trong bài báo của F. V. Atabay, R. M. Pagkalinawan, S. D. Pajarillo, A. R. Villanueva và J. V. Taylar, các mô hình dự báo ARIMAX và SARIMAX truyền thống và các mô hình học sâu dựa trên RNN đã được sử dụng để lập mô hình dữ liệu lịch sử tiêu thụ điện của một ngôi nhà hai tầng ở Houston, Texas, Hoa Kỳ. Các tính năng được sử dụng trong quy trình lập mô hình bao gồm dữ liệu lịch sử mức tiêu thụ điện trung bình hàng ngày của ngôi nhà hai tầng, danh mục ngày (ngày trong tuần, cuối tuần, ngày nghỉ và thời gian phong tỏa do COVID) và các biến số liên quan đến thời tiết. Hiệu suất lỗi tương ứng của mỗi mô hình trên tập dữ liệu thử nghiệm được so sánh. Kết quả cho thấy các mô hình học sâu dựa trên RNN vượt trội so với các mô hình ARIMAX và SARIMAX truyền thống trong việc dự báo mức tiêu thụ điện trung bình hàng ngày của ngôi nhà hai tầng và hiệu suất của các mô hình học sâu dựa trên RNN không khác biệt đáng kể so với mỗi mô hình khác. [3]

Trong một đề xuất của S. Kwak và N. Geroliminis sử dụng các mô hình tuyến tính động (DLM) để tính gần đúng các trạng thái lưu lượng phi tuyến tính. Họ so sánh độ chính xác dự đoán về thời gian di chuyển của các đường cao tốc ở California (I210-E và I5-S) trong điều kiện giao thông tắc nghẽn nghiêm trọng với các phương pháp khác: thời gian di chuyển tức thời, k-hàng xóm gần nhất, hồi quy véc tơ hỗ trợ và mạng thần kinh nhân tạo. Kết quả cho thấy những cải tiến đáng kể về độ chính xác, đặc biệt là đối với dự đoán ngắn hạn. [4]

Trong nghiên cứu tìm hiểu sự phát triển của COVID-19 của S. Prakash, A. S. Jalal và P. Pathak, các thuật toán chuỗi thời gian nổi tiếng đã được triển khai bao gồm LSTM, LSTM hai chiều và LSTM xếp chồng và Tiên tri. Do thành công của các thuật toán lai trong các lĩnh vực vấn đề cụ thể - nghiên cứu hiện tại cũng tập trung vào các thuật toán như GRU-LSTM, CNN-LSTM và LSTM với sự chú ý. Tất cả các mô hình này đã được đào tạo trên tập dữ liệu chuỗi thời gian của COVID-19 cho Ấn Độ và các chỉ số hiệu suất được ghi lại. Trong tất cả các mô hình, các thuật toán đơn giản hoạt động tốt hơn các thuật toán phức tạp và kết hợp. Nhờ có kết quả tốt nhất này đã đạt được với Prophet, LSTM hai chiều và LSTM Vanilla [5]

Trong nghiên cứu của M. Afshar, M. Heydarzadeh và B. Akin với mục đích phát triển một hệ thống giám sát tình trạng thời gian thực đối với các lỗi của vòng bi và với mục đích này, một hệ thống cảm biến để đo rung động 3 trục được phát triển. Một mô hình chuỗi thời gian được đề xuất cho các tín hiệu rung động để trích xuất các tính năng một cách chính xác. Thuật toán này có thể chẩn đoán năm lỗi khác nhau với độ chính xác đầy hứa hẹn là 98,5%, được tính toán cho các điều kiện tải và tốc độ khác nhau. [6]

Trong nghiên cứu của Y. Liu đã điều tra dữ liệu lịch sử về giá cổ phiếu của các công ty niêm yết. Tôi sử dụng các thuật toán học máy như LSTM, GRU để dự đoán giá cổ phiếu trong tương lai của Tập đoàn Dầu khí Quốc gia Trung Quốc, bắt đầu với toàn bộ quá trình dự báo cổ phiếu và sau đó chuyển sang chạy quy trình này để có được biểu đồ dự đoán giá cổ phiếu của CNPC phù hợp với giá dự đoán và giá thực tế. giá. [7]

# KIẾN TRÚC MÔ HÌNH

## Linear Regression

Linear Regression là một kỹ thuật phân tích dữ liệu dự đoán giá trị của một biến dựa trên giá trị của một biến khác.

y = β0 + β1X +

Trong đó:

* y: giá trị dự đoán của biến phụ thuộc (y) cho bất kỳ giá trị nhất định nào của biến độc lập (x).
* β0: là hệ số chặn (intercept), giá trị dự đoán của y khi x bằng 0.
* β1: là hệ số hồi quy (regression coefficient) - cho biết mức độ dự đoán y thay đổi khi X tăng lên.
* X: là biến độc lập (independent variable) - biến mà chúng ta kỳ vọng có ảnh hưởng đến y.
* : là sai số ước lượng (error of the estimate), hoặc mức độ biến động trong ước lượng của chúng ta về hệ số hồi quy.

## ARIMA

Thuật toán ARIMA (Autoregressive Integrated Moving Average) là một phương pháp mô hình chuỗi thời gian phổ biến và mạnh mẽ. ARIMA kết hợp các thành phần autoregressive (AR), moving average (MA) và differencing (I) để xử lý và dự đoán các chuỗi thời gian không stationary.

Mô hình ARIMA có dạng ARIMA (p, d, q), trong đó:

* p là số lượng lags của autoregressive (AR), đại diện cho số lượng quan sát trước đó được sử dụng để dự đoán giá trị hiện tại.
* d là số lần differencing, đại diện cho số lần biến đổi chuỗi thời gian để đạt được tính stationary. Differencing giúp loại bỏ xu hướng và chu kỳ trong chuỗi.
* q là số lượng lags của moving average (MA), đại diện cho số lượng sai số trước đó được sử dụng để dự đoán giá trị hiện tại.

Công thức chung của mô hình ARIMA được biểu diễn như sau:

Y(t) = c + Σ(φ(i) \* Y(t-i)) + Σ(θ(j) \* ε(t-j)) + ε(t)

Trong đó:

* Y(t) là giá trị của chuỗi thời gian tại thời điểm t.
* c là hằng số chặn (intercept) của mô hình.
* φ(i) là hệ số autoregressive (AR) của thứ tự i.
* Y(t-i) là giá trị chuỗi thời gian tại thời điểm (t-i).
* θ(j) là hệ số moving average (MA) của thứ tự j.
* ε(t-j) là sai số dự đoán của thứ tự j.
* ε(t) là sai số ngẫu nhiên tại thời điểm t.

Thuật toán ARIMA kết hợp việc xác định các tham số p, d, q và ước lượng các hệ số φ(i) và θ(j) thông qua các phương pháp như hàm tự tương quan (autocorrelation function - ACF) và hàm tự tương quan riêng biệt (partial autocorrelation function - PACF) của chuỗi thời gian. Sau đó, mô hình ARIMA được xây dựng và sử dụng để dự đoán giá trị trong tương lai và đánh giá độ chính xác của dự đoán.

## SARIMAX

Thuật toán SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous Variables) là một phương pháp mô hình chuỗi thời gian mở rộng của ARIMA, cho phép mô hình hóa và dự đoán các chuỗi thời gian có yếu tố mùa vụ và kết hợp với các biến exogenous (biến không phụ thuộc vào chuỗi thời gian).

Mô hình SARIMAX có dạng SARIMAX (p, d, q) (P, D, Q, s), trong đó:

* p, d, q là các tham số tương tự như trong ARIMA, đại diện cho các thành phần autoregressive (AR), differencing (I) và moving average (MA).
* P, D, Q là các tham số tương tự như p, d, q nhưng được áp dụng cho thành phần mùa vụ của chuỗi thời gian.
* s là chu kỳ mùa vụ của chuỗi thời gian, cho biết sau mỗi s thời điểm, mẫu chuỗi thời gian bắt đầu lặp lại.

Công thức chung của mô hình SARIMAX được biểu diễn như sau:

Y(t) = c + Σ(φ(i) \* Y(t-i)) + Σ(θ(j) \* ε(t-j)) + Σ(Φ(k) \* Y(t-ks)) + Σ(Θ(l) \* ε(t-ls)) + β' \* X(t) + ε(t)

Trong đó:

* Y(t) là giá trị của chuỗi thời gian tại thời điểm t.
* c là hằng số chặn (intercept) của mô hình.
* φ(i), θ(j) là hệ số autoregressive (AR) và moving average (MA) tương ứng.
* Y(t-i), ε(t-j) là giá trị chuỗi thời gian và sai số dự đoán của thứ tự i và j.
* Φ(k), Θ(l) là hệ số autoregressive (AR) và moving average (MA) cho thành phần mùa vụ của chuỗi thời gian.
* Y(t-ks), ε(t-ls) là giá trị chuỗi thời gian và sai số dự đoán của thứ tự ks và ls cho thành phần mùa vụ.
* β' là vector hệ số cho các biến exogenous (X(t)) không phụ thuộc vào chuỗi thời gian.
* ε(t) là sai số ngẫu nhiên tại thời điểm t.

Thuật toán SARIMAX sử dụng các phương pháp ước lượng tương tự như trong ARIMA để xác định các tham số và hệ số của mô hình. Nó kết hợp mô hình của ARIMA với khả năng mô hình hóa yếu tố mùa vụ và tích hợp các biến exogenous để giải quyết các yếu tố bên ngoài ảnh hưởng đến chuỗi thời gian. Thuật toán SARIMAX cho phép dự đoán và phân tích các chuỗi thời gian phức tạp hơn, có sự biến đổi theo mùa vụ và bị ảnh hưởng bởi các biến không phụ thuộc vào chuỗi thời gian. Nó đã được sử dụng rộng rãi trong nhiều lĩnh vực như dự báo doanh thu bán lẻ, lưu lượng giao thông, tiêu thụ năng lượng, và dự báo thời tiết.

## DLM

Mô hình tuyến tính động (Dynamic Linear Model - DLM) là một thuật toán thống kê được sử dụng để mô hình hóa và dự đoán các chuỗi dữ liệu thời gian. Nó là một phần của lĩnh vực học máy thống kê và thường được sử dụng trong việc phân tích dữ liệu thời gian, dự báo, và điều khiển. Một DLM bao gồm hai thành phần chính:

* State component: Đây là một mô hình tuyến tính động mà mô tả sự phát triển của hệ thống theo thời gian. Thông thường, mô hình trạng thái sử dụng một số biến trạng thái để mô hình hóa sự biến đổi và sự phụ thuộc của dữ liệu thời gian. Mô hình trạng thái thường được mô tả bằng một phương trình đệ quy.
* Observation component: Đây là mô hình mô tả cách các biến quan sát được tạo ra từ các biến trạng thái. Thông thường, mô hình quan sát sử dụng một mô hình tuyến tính để liên kết giữa các biến trạng thái và các biến quan sát.

Phương trình trạng thái (State equation):

xt ​= Gt​ xt−1​+ wt​

Trong đó:

* xt ​ là vector trạng thái tại thời điểm t.
* Gt là ma trận chuyển đổi trạng thái, mô tả cách trạng thái tại thời điểm t phụ thuộc vào trạng thái tại thời điểm t−1.
* wt​ ​là vector nhiễu trạng thái tại thời điểm t.

Phương trình quan sát (Observation equation):

yt​ = Ft ​xt​ + vt

Trong đó:

* yt ​là vector quan sát tại thời điểm t.
* Ft là ma trận quan sát, mô tả cách quan sát tại thời điểm t phụ thuộc vào trạng thái tại thời điểm t.
* vt ​là vector nhiễu quan sát tại thời điểm t.

## BDLM

Thuật toán Bayesian Dynamic Linear Model (BDLM) là một phương pháp được sử dụng để mô hình hóa và dự đoán chuỗi dữ liệu thời gian. Nó kết hợp cả yếu tố học máy Bayesian (Bayesian machine learning) và mô hình tuyến tính động (dynamic linear model) để xác định mô hình tốt nhất cho dữ liệu thời gian và cung cấp dự đoán chính xác.

BDLM sử dụng một mô hình tuyến tính động để mô tả sự biến đổi của dữ liệu thời gian theo thời gian. Mô hình tuyến tính động gồm hai thành phần chính là quá trình ẩn (hidden process) và quá trình quan sát (observation process). Quá trình ẩn là một quá trình Markov ẩn, mô tả sự biến đổi của dữ liệu thời gian không nhìn thấy được, trong khi quá trình quan sát liên kết giữa quá trình ẩn và dữ liệu quan sát được.

Phương trình trạng thái (State equation):

**x**(*t*+1) =**Ax**(*t*)+**n**(*t*) với n ~ N(µ, 2n )

Trong đó:

* x: Trạng thái ở thời điểm t.
* A: Ma trận giá trị trạng thái.
* n: Sai số trạng thái.
* t: Tại thời điểm t.

Phương trình quan sát (Observation equation):

**y**(*t*)=**Bx**(*t*)+**v**(*t*) với v ~ N(µ, 2v )

Trong đó:

* y: Giá trị đo đạc tại thời điểm quan sát t.
* x: Trạng thái tại thời điểm t.
* B: Ma trận giá trị quan sát.
* v: Sai số giá trị đo đạc.
* t: Tại thời điểm t.

## GBT

Thuật toán GBT (Gradient Boosting Trees) là một phương pháp học máy được sử dụng trong việc xây dựng các mô hình dự đoán. GBT kết hợp cả hai yếu tố quan trọng: Gradient Boosting và Decision Trees (cây quyết định). GBT hoạt động bằng cách xây dựng các cây quyết định tuần tự theo kiểu học tập tập trung. Mỗi cây quyết định đóng vai trò như một bộ phân loại hoặc một bộ hồi quy và được xây dựng dựa trên các cây quyết định trước đó. GBT kết hợp cả hai yếu tố quan trọng: Gradient Boosting và Decision Trees (cây quyết định). Quá trình xây dựng mô hình GBT bao gồm các bước sau:

* Xây dựng cây quyết định ban đầu dựa trên dữ liệu huấn luyện và đầu ra dự đoán ban đầu.
* Tính toán đạo hàm riêng cho mô hình hiện tại dựa trên dữ liệu huấn luyện và đầu ra dự đoán.
* Xây dựng cây quyết định mới để dự đoán sai số giữa đầu ra thực tế và đầu ra dự đoán hiện tại.
* Cộng dồn cây quyết định mới vào mô hình hiện tại và cập nhật đầu ra dự đoán.

Quá trình này được lặp lại cho đến khi đạt được số lượng cây quyết định mong muốn hoặc đạt được điều kiện dừng nhất định.GBT có khả năng xử lý các biến độc lập và tạo ra các dự đoán chính xác trên dữ liệu mới. Nó cũng cung cấp độ quan trọng của các biến đầu vào, giúp phân tích và đánh giá yếu tố ảnh hưởng đến giá cổ phiếu.

Chúng tôi sử dụng các thư viện phổ biến như XGBoost, LightGBM và CatBoost để triển khai GBT trong việc dự đoán giá cổ phiếu.Kết quả thực nghiệm cho thấy rằng GBT là một công cụ mạnh mẽ để dự đoán giá cổ phiếu, mang lại hiệu suất cao và khả năng khả diễn giải. Nghiên cứu này có thể đóng góp vào việc phân tích thị trường tài chính và hỗ trợ các nhà đầu tư trong việc ra quyết định mua bán cổ phiếu.

## RNN

Trong lĩnh vực dự đoán giá cổ phiếu, mô hình Recurrent Neural Network (RNN) là một công cụ mạnh mẽ được sử dụng để mô hình hóa và dự đoán các chuỗi thời gian phức tạp. Với khả năng xử lý dữ liệu chuỗi theo thứ tự và ghi nhớ thông tin lịch sử, RNN cho phép ứng dụng dự đoán giá cổ phiếu với tính linh hoạt và hiệu suất cao. Kiến trúc mô hình RNN thông thường bao gồm các thành phần sau:

* Input Layer: Nhận dữ liệu đầu vào, chẳng hạn như các biến kỹ thuật (ví dụ: giá đóng cửa, khối lượng giao dịch, chỉ số kỹ thuật, vv).
* Hidden Layers: Gồm một số lớp trạng thái ẩn (hidden states) kết nối với nhau theo thứ tự thời gian. Mỗi lớp trạng thái ẩn chứa một số cell RNN, cho phép mô hình ghi nhớ thông tin lịch sử và học cấu trúc dữ liệu chuỗi.
* Output Layer: Tạo ra dự đoán về giá cổ phiếu dựa trên trạng thái ẩn cuối cùng của mô hình.

A picture containing diagram, sketch, line, white

Description automatically generated

Hình Kiến trúc mô hình RNN

## LSTM

Trong lĩnh vực dự đoán giá cổ phiếu, mô hình Long Short-Term Memory (LSTM) đã được áp dụng và đạt được kết quả tốt. Kiến trúc LSTM là một loại mạng neuron nhân tạo đặc biệt được thiết kế để xử lý và mô hình hóa dữ liệu chuỗi dài và phức tạp. Mô hình LSTM cũng có cấu trúc giống với mô hình RNN, tuy nhiên nó sử dụng các đơn vị lặp lại (cell) phức tạp hơn, gọi là "memory cell", giúp mô hình lưu trữ và truy xuất thông tin lịch sử một cách hiệu quả hơn. Kiến trúc mô hình LSTM thông thường bao gồm các thành phần sau:

* Input Layer: Nhận dữ liệu đầu vào, chẳng hạn như các biến kỹ thuật (ví dụ: giá đóng cửa, khối lượng giao dịch, chỉ số kỹ thuật, vv).
* LSTM Layer: Gồm một số memory cell kết nối với nhau theo thứ tự thời gian. Mỗi memory cell bao gồm ba cổng (gate): cổng quên (forget gate), cổng đầu vào (input gate) và cổng đầu ra (output gate), giúp mô hình quyết định lưu trữ hay loại bỏ thông tin lịch sử.
* Output Layer: Tạo ra dự đoán về giá cổ phiếu dựa trên trạng thái ẩn cuối cùng của mô hình.

A screenshot of a computer

Description automatically generated with low confidence

Hình Kiến trúc mô hình LSTM

Quá trình huấn luyện mô hình LSTM cũng tương tự như mô hình RNN, bao gồm cung cấp dữ liệu huấn luyện, xác định hàm mất mát và tối ưu hóa các tham số mô hình. Một khi mô hình được huấn luyện, nó có thể sử dụng để dự đoán giá cổ phiếu trong tương lai dựa trên dữ liệu mới.

## CNN-LSTM

Mô hình CNN-LSTM là một kiến trúc mạng hỗn hợp được áp dụng trong lĩnh vực dự đoán giá cổ phiếu, kết hợp giữa Convolutional Neural Network (CNN) và Long Short-Term Memory (LSTM). Kiến trúc này kết hợp khả năng trích xuất đặc trưng của CNN và khả năng mô hình hóa dữ liệu chuỗi của LSTM để tăng cường khả năng dự đoán chính xác của mô hình. Kiến trúc mô hình CNN-LSTM bao gồm các thành phần sau:

* Input Layer: Nhận dữ liệu đầu vào, chẳng hạn như các biến kỹ thuật (ví dụ: giá đóng cửa, khối lượng giao dịch, chỉ số kỹ thuật, vv).
* CNN Layer: Sử dụng các lớp tích chập để trích xuất đặc trưng từ dữ liệu chuỗi. Các bộ lọc tích chập được áp dụng để nhận diện các mẫu và cấu trúc trong dữ liệu chuỗi, giúp mô hình học được những đặc trưng quan trọng cho dự đoán giá cổ phiếu.
* LSTM Layer: Nhận các đặc trưng đã được trích xuất từ CNN và sử dụng các memory cell để xử lý và mô hình hóa dữ liệu chuỗi. LSTM giúp mô hình ghi nhớ thông tin lịch sử quan trọng và dự đoán dựa trên quá khứ.
* Output Layer: Tạo ra dự đoán về giá cổ phiếu dựa trên trạng thái ẩn cuối cùng của mô hình.

A picture containing diagram, text, plan, rectangle

Description automatically generated

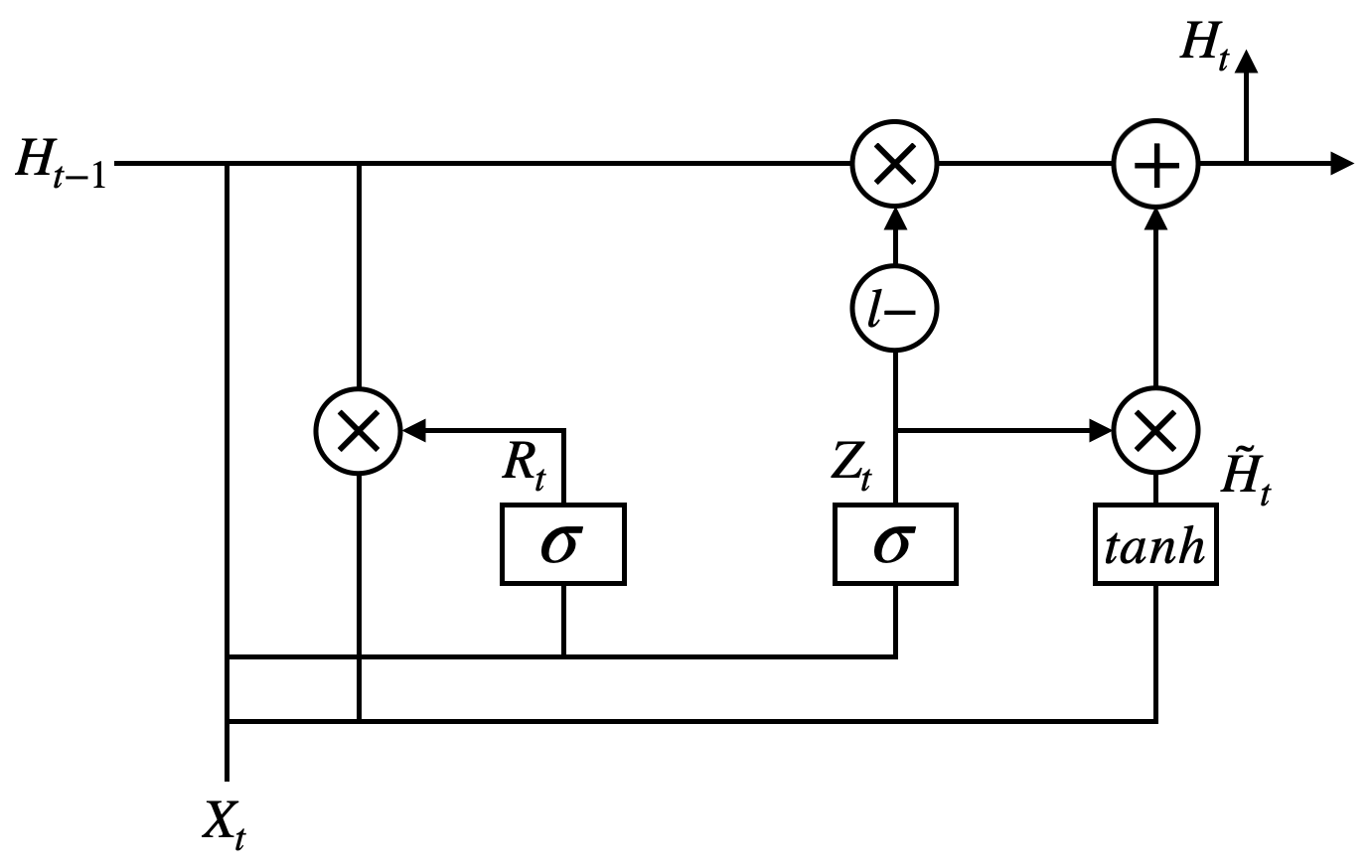
Hình Kiến trúc mô hình CNN-LSTM

Quá trình huấn luyện mô hình CNN-LSTM tương tự như các mô hình khác, bao gồm cung cấp dữ liệu huấn luyện, xác định hàm mất mát và tối ưu hóa các tham số mô hình. Mô hình CNN-LSTM có khả năng học được sự phụ thuộc không gian và thời gian trong dữ liệu chuỗi, giúp cải thiện hiệu suất dự đoán giá cổ phiếu.

## GRU

Mô hình Gated Recurrent Unit (GRU) là một kiến trúc mạng neuron nhân tạo được sử dụng trong lĩnh vực dự đoán giá cổ phiếu. GRU là một biến thể của mô hình Long Short-Term Memory (LSTM) nhưng có cấu trúc đơn giản hơn và hiệu suất tính toán cao hơn. Kiến trúc mô hình GRU bao gồm các thành phần sau:

* Input Layer: Nhận dữ liệu đầu vào, bao gồm các biến kỹ thuật (ví dụ: giá đóng cửa, khối lượng giao dịch, chỉ số kỹ thuật, vv).
* GRU Layer: Là lớp trung gian giữa input layer và output layer. GRU sử dụng các cổng (gates) để kiểm soát việc lưu trữ và truy xuất thông tin lịch sử. Nó bao gồm hai cổng chính: cổng cập nhật (update gate) và cổng khôi phục (reset gate), giúp mô hình quyết định thông tin nào cần lưu giữ và thông tin nào cần bỏ qua.
* Output Layer: Tạo ra dự đoán về giá cổ phiếu dựa trên trạng thái ẩn cuối cùng của mô hình GRU.



Hình Kiến trúc mô hình GRU

Quá trình huấn luyện mô hình GRU bao gồm cung cấp dữ liệu huấn luyện, xác định hàm mất mát và tối ưu hóa các tham số mô hình. Mô hình GRU có khả năng học được mối quan hệ phức tạp trong dữ liệu chuỗi, giúp cải thiện khả năng dự đoán giá cổ phiếu.

Tóm lại, mô hình GRU là một kiến trúc mạng neuron nhân tạo đơn giản và hiệu quả trong ứng dụng dự đoán giá cổ phiếu. Với khả năng kiểm soát thông tin lịch sử bằng các cổng, GRU cho phép mô hình học được sự phụ thuộc trong dữ liệu chuỗi và đưa ra dự đoán chính xác về giá cổ phiếu.

# DỮ LIỆU VÀ CÔNG CỤ

## Dữ liệu

Yahoo Finance là một dịch vụ trực tuyến cung cấp thông tin tài chính, tin tức và công cụ phân tích cho các nhà đầu tư và người quan tâm đến thị trường tài chính. Nó được cung cấp bởi công ty công nghệ và truyền thông đa phương tiện Yahoo. Các tập dữ liệu được dùng trong nghiên cứu này dựa trên dữ liệu lịch sử do trang Yahoo Finance cung cấp bao gồm giá cổ phiếu các công ty HBAN, KEY, ERIC từ ngày 4/1/2010 đến ngày 31/5/2023 Tập dữ liệu bao gồm các thông tin liên quan đến chứng khoán như sau:

* Open: giá đầu tiên mà một cổ phiếu được giao dịch khi sàn giao dịch mở cửa vào một ngày giao dịch.
* Close Price: giá cuối cùng mà một cổ phiếu được giao dịch khi sàn giao dịch đóng cửa ngày giao dịch.
* High: giá cao nhất của cổ phiếu.
* Low: giá thấp nhất của cổ phiếu.
* Volume: số lượng cổ phiếu được giao dịch trong một mã cổ phiếu cụ thể.

Tiến hành chia tập dữ liệu thành 3 phần: train, test, validate. Trong đó các phần dữ liệu train, test, validate được chia theo các thứ tự tỉ lệ tương ứng sau: 60%-20%-20%, 70%-20%-10%, 80%-10%-10%.

Trong nghiên cứu này, chúng tôi sử dụng giá trị Close làm giá trị cần dự đoán. Các giá trị trung bình, độ lệch chuẩn, giá trị lớn nhất và giá trị nhỏ nhất, số lượng giá trị của dữ liệu được thể hiện trong Bảng 1.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **HBAN** | **KEY** | **ERIC** |
| **Count** | 3375 | 3375 | 3375 |
| **Mean** | 10.746607 | 14.278945 | 9.728969 |
| **Min** | 3.68 | 5.71 | 4.85 |
| **Max** | 17.63 | 27.01 | 15.38 |
| **25%** | 7.49 | 9.885 | 7.91 |
| **50%** | 10.56 | 13.64 | 9.75 |
| **75%** | 13.855 | 18.29 | 11.76 |
| **Box Plot** |  |  |  |
| **Histogram** |  |  |  |
| **Skewness** | -0.091 | 0.165 | -0.127 |
| **Kurtosis** | -1.271 | -0.99 | -0.914 |

Bảng 1 Thống kê mô tả dữ liệu

## Công cụ

Trong nghiên cứu này, chúng tôi sử dụng ngôn ngữ Python để đánh giá các mô hình.

Đối với việc triển khai thuật toán Linear Regression, chúng tôi sử dụng thư viện sklearn.linear model.Linear-Regression. Đối với việc triển khai thuật toán ARIMA và SARIMAX, chúng tôi sử dụng thư viện pmdarima và statsmodels. tsa.arima. Đối với việc triển khai thuật toán Gradient Boosting (GBT), chúng tôi sử dụng thư viện sklearn.ensemble.GradientBoostingRegressor. Đối với việc triển khai thuật toán Dynamic Linear Model (DLM), chúng tôi sử dụng thư viện pydlm. Đối với việc triển khai thuật toán Bayesian Dynamic Linear Model (BDLM), chúng tôi sử dụng thư viện pybats. Đối với việc triển khai thuật toán Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM) và Gated Recurrent Unit (GRU), chúng tôi sử dụng thư viện keras. Đối với việc triển khai thuật toán CNN-LSTM, chúng tôi sử dụng thư viện keras kết hợp với keras.layers.Conv-1D, MaxPooling.

## Các độ đo đánh giá

Trong nghiên cứu này, chúng tôi sử dụng 3 độ đo hiệu suất khác nhau cho từng mô hình: RMSE (Root mean squared error), MAE (Mean Absolute Error), MAPE (Mean absolute percentage error).

Root mean squared error (RMSE) là một độ đo thường được sử dụng để đánh giá độ chính xác của dự đoán được thu được bằng một mô hình. Nó đo lường sự khác biệt hoặc sai số giữa giá trị thực tế và giá trị dự đoán. Độ đo này so sánh các sai số dự đoán của các mô hình khác nhau cho một tập dữ liệu cụ thể và không so sánh giữa các tập dữ liệu khác nhau. Công thức tính RMSE như sau:

RMSE =

Mean Absolute Error (MAE) đo độ lớn trung bình của các lỗi trong một tập hợp các dự đoán mà không cần xem xét hướng của chúng. Đó là giá trị trung bình trên mẫu thử nghiệm về sự khác biệt tuyệt đối giữa dự đoán và quan sát thực tế, trong đó tất cả các khác biệt riêng lẻ có trọng số bằng nhau. [8] Công thức tính MAE như sau:

MAE =

Mean absolute percentage error (MAPE) là phần trăm sai số trung bình tuyệt đối. Lỗi phần trăm tuyệt đối trung bình (MAPE) là thước đo thống kê mức độ chính xác của một hệ thống dự báo. Nó đo độ chính xác này theo phần trăm và có thể được tính là sai số phần trăm tuyệt đối trung bình cho mỗi khoảng thời gian trừ đi các giá trị thực chia cho các giá trị thực. Công thức tính MAPE như sau:

MAPE =

Trong đó, At là giá trị thực tế và Ft là giá trị dự báo. Sai số giữa chúng được chia cho giá trị thực tế At. Giá trị tuyệt đối của tỷ lệ này được tổng hợp cho mỗi điểm dự báo trong thời gian và chia cho số điểm phù hợp n.

# KẾT QUẢ

Chúng tôi tính toán RMSE (Root mean squared error), MAE (Mean Absolute Error), MAPE (Mean absolute percentage error) giữa Giá trị dự đoán và Giá trị thực tế để đánh giá và so sánh giữa các mô hình.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **HBAN** | | | | |
| **Mô hình** | **Train -Test - Validate** | **MAPE** | **RMSE** | **MAE** |
| Linear Regression | 6-2-2 | 23.153 | 3.382 | 2.441 |
| 7-2-1 | 30.385 | 4.059 | 3.251 |
| 8-1-1 | 10.787 | 1.898 | 1.304 |
| ARIMA | 6-2-2 | 39.460 | 5.403 | 4.271 |
| 7-2-1 | 19.132 | 2.617 | 2.145 |
| 8-1-1 | 38.461 | 6.075 | 5.755 |
| SARIMAX | 6-2-2 | 44.185 | 5.934 | 4.852 |
| 7-2-1 | 19.188 | 2.614 | 2.181 |
| 8-1-1 | 38.911 | 6.171 | 5.830 |
| DLM | 6-2-2 | 22.064 | 3.414 | 2.581 |
| 7-2-1 | 24.564 | 3.706 | 2.888 |
| 8-1-1 | 15.662 | 2.796 | 2.067 |
| BDLM | 6-2-2 | 40.471 | 5.471 | 4.592 |
| 6-3-1 | 48.072 | 6.814 | 5.814 |
| 7-2-1 | 21.169 | 2.980 | 2.661 |
| GBT | 6-2-2 | 2.987 | 0.348 | 0.279 |
| 7-2-1 | 2.66 | 0.453 | 0.364 |
| 8-1-1 | 2.67 | 0.59 | 0.435 |
| **RNN** | 6-2-2 | 1.86 | 0.295 | 0.212 |
| 7-2-1 | 2.403 | 0.387 | 0.281 |
| **8-1-1** | **1.552** | **0.313** | **0.237** |
| LSTM | 6-2-2 | 2.177 | 0.355 | 0.246 |
| 7-2-1 | 2.517 | 0.416 | 0.293 |
| **8-1-1** | **1.6** | **0.317** | **0.246** |
| CNN-LSTM | 6-2-2 | 8.741 | 1.279 | 1.155 |
| **7-2-1** | **8.141** | **1.251** | **1.111** |
| 8-1-1 | 10.068 | 1.545 | 1.387 |
| **GRU** | **6-2-2** | **1.858** | **0.29** | **0.213** |
| 7-2-1 | 2.119 | 0.342 | 0.25 |
| 8-1-1 | 1.525 | 0.306 | 0.233 |

Bảng 2 Bảng thống kê giá trị MAPE, RMSE, MAE của các mô hình trên tập dữ liệu HBAN

Kết quả 3 thuật toán tốt nhất trên các tập dữ liệu HBAN:

A picture containing text, plot, line, font

Description automatically generated

Hình Kết quả mô hình RNN trên tập dữ liệu HBAN

A picture containing text, plot, line, font

Description automatically generated

Hình Kết quả mô hình LSTM trên tập dữ liệu HBAN

A picture containing text, plot, line, font

Description automatically generated

Hình Kết quả mô hình GRU trên tập dữ liệu HBAN

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **KEY** | | | | |
| **Mô hình** | **Train -Test - Validate** | **MAPE** | **RMSE** | **MAE** |
| Linear Regression | 6-2-2 | 25.347 | 4.633 | 3.525 |
| 7-2-1 | 28.650 | 5.170 | 3.898 |
| 8-1-1 | 13.637 | 3.079 | 2.424 |
| ARIMA | 6-2-2 | 43.558 | 7.567 | 6.175 |
| 7-2-1 | 21.700 | 4.197 | 3.578 |
| 8-1-1 | 40.342 | 9.144 | 8.471 |
| SARIMAX | 6-2-2 | 48.286 | 8.266 | 6.929 |
| 7-2-1 | 21.709 | 4.214 | 3.594 |
| 8-1-1 | 39.690 | 9.052 | 8.354 |
| DLM | 6-2-2 | 23.447 | 4.612 | 3.616 |
| 7-2-1 | 28.955 | 5.77 | 4.642 |
| 8-1-1 | 20.953 | 4.909 | 3.825 |
| BDLM | 6-2-2 | 49.510 | 8.495 | 7.390 |
| 6-3-1 | 56.217 | 10.318 | 8.914 |
| 7-2-1 | 27.819 | 6.129 | 5.244 |
| GBT | 6-2-2 | 2.319 | 0.338 | 0.282 |
| 7-2-1 | 4.471 | 1.457 | 0.932 |
| 8-1-1 | 11.598 | 3.102 | 2.824 |
| **RNN** | 6-2-2 | 2.018 | 0.423 | 0.293 |
| 7-2-1 | 2.405 | 0.513 | 0.383 |
| **8-1-1** | **1.566** | **0.425** | **0.339** |
| LSTM | 6-2-2 | 2.432 | 0.525 | 0.355 |
| 7-2-1 | 2.399 | 0.533 | 0.37 |
| **8-1-1** | **1.48** | **0.4092** | **0.318** |
| CNN-LSTM | **6-2-2** | **5.707** | **1.132** | **0.952** |
| 7-2-1 | 7.849 | 1.748 | 1.462 |
| 8-1-1 | 7.658 | 1.885 | 1.711 |
| **GRU** | 6-2-2 | 2.164 | 0.451 | 0.319 |
| 7-2-1 | 2.326 | 0.498 | 0.37 |
| **8-1-1** | **1.494** | **0.412** | **0.321** |

Bảng 3 Bảng thống kê giá trị MAPE, RMSE, MAE của các mô hình trên tập dữ liệu KEY

Kết quả 3 thuật toán tốt nhất trên các tập dữ liệu KEY:

A picture containing text, plot, line, diagram

Description automatically generated

Hình Kết quả mô hình RNN trên tập dữ liệu KEY

A picture containing text, plot, line, diagram

Description automatically generated

Hình Kết quả mô hình LSTM trên tập dữ liệu KEY

A picture containing text, plot, line, diagram

Description automatically generated

Hình Kết quả mô hình GRU trên tập dữ liệu KEY

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ERIC** | | | | |
| **Mô hình** | **Train -Test - Validate** | **MAPE** | **RMSE** | **MAE** |
| Linear Regression | 6-2-2 | 22.679 | 2.966 | 2.773 |
| 7-2-1 | 33.245 | 4.332 | 3.776 |
| 8-1-1 | 37.808 | 4.706 | 4.601 |
| ARIMA | 6-2-2 | 16.438 | 1.805 | 1.510 |
| 7-2-1 | 15.798 | 1.950 | 1.693 |
| 8-1-1 | 10.938 | 1.684 | 1.386 |
| SARIMAX | 6-2-2 | 15.918 | 1.764 | 1.465 |
| 7-2-1 | 15.794 | 1.983 | 1.706 |
| 8-1-1 | 11.964 | 1.807 | 1.511 |
| DLM | 6-2-2 | 15.088 | 1.659 | 1.281 |
| 7-2-1 | 20.709 | 2.597 | 2.107 |
| 8-1-1 | 9.828 | 1.456 | 1.179 |
| BDLM | 6-2-2 | 34.347 | 3.314 | 3.075 |
| 6-3-1 | 44.113 | 4.963 | 4.527 |
| 7-2-1 | 26.069 | 3.011 | 2.466 |
| GBT | 6-2-2 | 3.39 | 0.439 | 0.376 |
| 7-2-1 | 1.43 | 0.25 | 0.169 |
| 8-1-1 | 1.219 | 0.183 | 0.13 |
| **RNN** | 6-2-2 | 1.66 | 0.209 | 0.146 |
| 7-2-1 | 1.61 | 0.244 | 0.167 |
| **8-1-1** | **1.242** | **0.213** | **0.15** |
| LSTM | 6-2-2 | 1.78 | 0.223 | 0.158 |
| 7-2-1 | 1.522 | 0.235 | 0.159 |
| **8-1-1** | **1.28** | **0.216** | **0.154** |
| CNN-LSTM | **6-2-2** | **3.049** | **0.357** | **0.263** |
| 7-2-1 | 2.814 | 0.407 | 0.302 |
| 8-1-1 | 5.627 | 0.767 | 0.698 |
| **GRU** | 6-2-2 | 1.875 | 0.222 | 0.165 |
| 7-2-1 | 1.497 | 0.231 | 0.156 |
| **8-1-1** | **1.193** | **0.2** | **0.144** |

Bảng 4 Bảng thống kê giá trị MAPE, RMSE, MAE của các mô hình trên tập dữ liệu ERIC

Kết quả 3 thuật toán tốt nhất trên các tập dữ liệu ERIC:

A picture containing text, plot, line, diagram

Description automatically generated

Hình Kết quả mô hình RNN trên tập dữ liệu ERIC

A picture containing screenshot, text, plot, line

Description automatically generated

Hình Kết quả mô hình LSTM trên tập dữ liệu ERIC

A picture containing text, plot, line, diagram

Description automatically generated

Hình Kết quả mô hình GRU trên tập dữ liệu ERIC

Trong Bảng 5 dưới đây, chúng tôi đưa ra kết quả dự đoán giá đóng cửa của các tập dữ liệu chứng khoán trong vòng 30 ngày tiếp theo dựa trên 2 mô hình có kết quả tốt nhất.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **RNN** | | | **GRU** | | |
| **HBAN** | **KEY** | **ERIC** | **HBAN** | **KEY** | **ERIC** |
| **1/6** | 10.51 | 9.55 | 5.163 | 10.3 | 9.516 | 5.18 |
| **2/6** | 10.52 | 9.58 | 5.177 | 10.25 | 9.517 | 5.19 |
| **3/6** | 10.66 | 9.80 | 5.203 | 10.2 | 9.58 | 5.199 |
| **4/6** | 10.82 | 9.94 | 5.256 | 10.15 | 9.63 | 5.2 |
| **5/6** | 10.88 | 9.90 | 5.245 | 10.1 | 9.69 | 5.21 |
| **6/6** | 10.86 | 9.77 | 5.253 | 10.05 | 9.74 | 5.22 |
| **7/6** | 10.94 | 9.75 | 5.253 | 10 | 9.79 | 5.23 |
| **8/6** | 10.96 | 9.80 | 5.259 | 9.95 | 9.85 | 5.238 |
| **9/6** | 10.94 | 9.71 | 5.340 | 9.9 | 9.9 | 5.245 |
| **10/6** | 11.09 | 9.78 | 5.325 | 9.86 | 9.96 | 5.252 |
| **11/6** | 11.11 | 9.78 | 5.340 | 9.8 | 10.01 | 5.259 |
| **12/6** | 11.10 | 9.72 | 5.347 | 9.75 | 10.07 | 5.266 |
| **13/6** | 11.09 | 9.69 | 5.386 | 9.71 | 10.12 | 5.273 |
| **14/6** | 11.09 | 9.72 | 5.447 | 9.66 | 10.18 | 5.28 |
| **15/6** | 11.10 | 9.77 | 5.454 | 9.61 | 10.24 | 5.287 |
| **16/6** | 11.19 | 9.93 | 5.498 | 9.56 | 10.29 | 5.294 |
| **17/6** | 11.21 | 9.91 | 5.481 | 9.51 | 10.35 | 5.301 |
| **18/6** | 11.36 | 10.12 | 5.517 | 9.46 | 10.4 | 5.307 |
| **19/6** | 11.42 | 9.87 | 5.546 | 9.41 | 10.46 | 5.314 |
| **20/6** | 11.31 | 9.64 | 5.531 | 9.37 | 10.52 | 5.321 |
| **21/6** | 11.34 | 9.58 | 5.567 | 9.32 | 10.58 | 5.327 |
| **22/6** | 11.47 | 9.82 | 5.567 | 9.27 | 10.64 | 5.334 |
| **23/6** | 11.56 | 10.03 | 5.585 | 9.22 | 10.69 | 5.34 |
| **24/6** | 11.67 | 10.29 | 5.649 | 9.17 | 10.75 | 5.347 |
| **25/6** | 11.76 | 10.33 | 5.660 | 9.13 | 10.81 | 5.353 |
| **26/6** | 11.79 | 10.31 | 5.687 | 9.08 | 10.87 | 5.36 |
| **27/6** | 11.81 | 10.20 | 5.678 | 9.04 | 10.93 | 5.366 |
| **28/6** | 11.91 | 10.25 | 5.722 | 8.99 | 10.98 | 5.372 |
| **29/6** | 12.02 | 10.55 | 5.730 | 8.94 | 11.04 | 5.379 |
| **30/6** | 12.13 | 10.81 | 5.778 | 9.9 | 11.1 | 5.385 |

Bảng 5 Dự đoán giá cổ phiếu 30 ngày tiếp theo của tháng 6 năm 2023 dựa trên 2 mô hình có kết quả tốt nhất

# KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã nghiên cứu và so sánh hiệu suất của các thuật toán dự đoán giá cổ phiếu, bao gồm Linear Regression, ARIMA, SARIMAX, GBT, DLM, BDLM, RNN, LSTM, CNN-LSTM và GRU. Kết quả cho thấy, trong số các thuật toán này, RNN và GRU đã cho ra kết quả tốt nhất. Cả hai thuật toán RNN và GRU đã chứng minh khả năng dự đoán giá cổ phiếu hiệu quả. Tuy nhiên, cần lưu ý rằng các thuật toán dự đoán giá cổ phiếu chỉ dựa trên các thông tin lịch sử và các yếu tố khác nhau có thể ảnh hưởng đến giá cổ phiếu. Do đó, việc dự đoán giá cổ phiếu vẫn có một mức độ không chắc chắn. Cần có sự cân nhắc và phân tích kỹ lưỡng khi sử dụng dự đoán giá cổ phiếu trong thực tế. Tóm lại, bài báo này đã thể hiện hiệu suất của một loạt các thuật toán dự đoán giá cổ phiếu. RNN và GRU đã cho ra kết quả tốt nhất trong số các thuật toán được nghiên cứu. Các kết quả này có thể cung cấp một cơ sở quan trọng cho việc phát triển và áp dụng các mô hình dự đoán giá cổ phiếu trong thực tế.

# TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | Jigar Patel, Sahil Shah, Priyank Thakkar, "Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques," *Expert Systems with Applications,* vol. 42, no. 1, pp. 259-268, January 2015. |
| [2] | D. Bhuriya, G. Kaushal, A. Sharma và U. Singh, "Stock market predication using a linear regression," *International conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2017,* pp. 510-513, 2017. |
| [3] | F. V. Atabay, R. M. Pagkalinawan, S. D. Pajarillo, A. R. Villanueva và J. V. Taylar, "Multivariate Time Series Forecasting using ARIMAX, SARIMAX, and RNN-based Deep Learning Models on Electricity Consumption," *International Informatics and Software Engineering Conference (IISEC), Ankara, Turkey,* pp. 1-6, 2022. |
| [4] | S. Kwak và N. Geroliminis, "Travel Time Prediction for Congested Freeways With a Dynamic Linear Model," *Transactions on Intelligent Transportation Systems,* vol. 22, no. 12, pp. 7667-7677, 2021. |
| [5] | S. Prakash, A. S. Jalal và P. Pathak, "Forecasting COVID-19 Pandemic using Prophet, LSTM, hybrid GRU-LSTM, CNN-LSTM, Bi-LSTM and Stacked-LSTM for India," *2023 6th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India,* pp. 1-6, 2023. |
| [6] | M. Afshar, M. Heydarzadeh và B. Akin, "Generalized Roughness Bearing Fault Diagnosis Using Time Series Analysis and Gradient Boosted Tree," *2022 IEEE Energy Conversion Congress and Exposition (ECCE), Detroit, MI, USA,* pp. 1-4, 2022. |
| [7] | Y. Liu, "Stock Prediction Using LSTM and GRU," *6th Annual International Conference on Data Science and Business Analytics (ICDSBA), Changsha, China,* pp. 206-211, 2022. |
| [8] | T. Chai, R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?," *Methods for assessment of models,* vol. 7, no. 3, p. 1247–1250, 2014. |