

DỰ BÁO GIÁ CỔ PHIẾU BYDDY SỬ DỤNG LSTM VÀ GENETIC ALGORITHM

● TRẦN BÁ THUẤN

TÓM TẮT:

Bài nghiên cứu tập trung vào dự báo giá cổ phiếu BYDDY của Công ty BYD Company Limited, nhà sản xuất ô tô điện, pin sạc, tấm pin mặt trời hàng đầu Trung Quốc, sử dụng mạng Bộ nhớ ngắn hạn dài LSTM tối ưu. Kết quả nghiên cứu cho thấy, mô hình LSTM tối ưu nắm bắt hiệu quả những biến động phi tuyến của giá cổ phiếu BYDDY đóng cửa. Các số liệu hiệu suất chính của mô hình LSTM tối ưu trên tập dữ liệu kiểm tra như $STD = 0.015$, $MAE = 0.025$, $MSE = 0.001$, $RMSE = 0.031$ và mức độ phù hợp của mô hình là $R^2 = 0.92$. Nghiên cứu này không chỉ nêu bật tính hiệu quả của mạng LSTM được tối ưu hóa trong dự báo giá cổ phiếu, mà còn cung cấp khuôn khổ cho nghiên cứu trong tương lai trong việc áp dụng kỹ thuật học sâu vào dự đoán thị trường tài chính.

Từ khóa: dự báo giá cổ phiếu, giá cổ phiếu BYDDY, Genetic Algorithm, kỹ thuật học sâu, LSTM.

1. Đặt vấn đề

Nghiên cứu này tập trung vào dự báo giá cổ phiếu của BYD Company Limited (BYDDY), một nhà sản xuất nổi tiếng của Trung Quốc chuyên về xe điện, pin và giải pháp năng lượng tái tạo. Với vai trò quan trọng của BYD trong thị trường xe điện đang phát triển nhanh chóng, việc dự báo chính xác giá cổ phiếu BYDDY đóng cửa là mối quan tâm đáng kể của các nhà đầu tư và nhà phân tích tài chính. Các phương pháp dự báo giá cổ phiếu truyền thống, chẳng hạn như đường trung bình động và mô hình trung bình trượt tích hợp tự hồi quy ARIMA đã được sử dụng rộng rãi.

Tuy nhiên, những phương pháp này thường không thể nắm bắt được các mô hình dữ liệu phức tạp, phi tuyến tính và sự phụ thuộc có trong dữ liệu chuỗi thời gian. Để giải quyết những hạn chế này, những tiến bộ gần đây trong học sâu đã cho thấy nhiều hứa hẹn tích cực. Trong số này, mạng bộ nhớ ngắn hạn dài (LSTM), một loại mạng thần kinh tái

phát (RNN), đã thu hút được chú ý nhờ khả năng mô hình hóa các mẫu và phụ thuộc dài hạn trong dữ liệu tuần tự. Mạng LSTM được thiết kế để khắc phục vấn đề biến mất độ dốc vốn có trong RNN truyền thống, khiến chúng rất phù hợp để dự báo chuỗi thời gian. Bằng cách lưu giữ thông tin một cách hiệu quả trong thời gian dài, LSTM có thể nắm bắt được thông tin phức tạp theo thời gian của giá cổ phiếu.

Mục tiêu nghiên cứu là tìm mô hình LSTM tối ưu sử dụng thuật toán Di truyền GA, để dự báo giá cổ phiếu BYDDY đóng cửa và ra quyết định giao dịch, nhằm chứng minh hiệu suất vượt trội của mô hình LSTM tối ưu so với các nghiên cứu trước đây.

2. Tổng quan nghiên cứu

Tại Việt Nam, Phùng Đại Khánh & Nguyễn Xuân Huy (2022) sử dụng thuật toán di truyền huấn luyện mạng LSTM dự đoán khai thác dầu khí, hiệu suất của mô hình $MAE = 1.16$, $MSE = 1.39$, $RMSE = 1.18$, $R^2 = 0.93$. Kết quả là các ứng

dụng đổi mới của học máy trong quản lý và dự báo tài nguyên. Trần Đăng Tuyên (2023) đánh giá hiệu quả hoạt động của các mô hình kết hợp LSTM-GRU trong dự báo diễn biến chỉ số chứng khoán trên Sở Giao dịch chứng khoán TP. Hồ Chí Minh. Nghiên cứu nhấn mạnh tính hiệu quả của việc kết hợp các kiến trúc mạng thần kinh khác nhau để có hiệu suất dự báo tốt hơn. Vũ Thị Loan và các cộng sự (2023) đã cung cấp những kết quả thực tế cho các nhà đầu tư và các nhà hoạch định chính sách, hiệu suất của mô hình là $MAE = 0.88$, $MSE = 1.12$, $RMSE = 1.06$, $R^2 = 0.87$.

Trên thế giới, Bao, Yue & Rao (2017) đề xuất một khung học sâu kết hợp các bộ mã hóa tự động xếp chồng mạng bộ nhớ ngắn hạn dài (LSTM Stacked). Cách tiếp cận này nắm bắt một cách hiệu quả các mô hình phi tuyến tính và tuần tự trong dữ liệu tài chính, hiệu suất $MAE = 1.34$, $MSE = 2.87$, $RMSE = 1.69$, $R^2 = 0.97$ cải thiện đáng kể về độ chính xác của dự báo so với các mô hình truyền thống. Nghiên cứu của Fischer & Krauss (2018) cho kết quả là mô hình LSTM có thể hoạt động tốt hơn các phương pháp dự báo chuỗi thời gian truyền thống, mô hình LSTM có hiệu suất $MAE = 1.01$, $MSE = 1.48$, $RMSE = 1.21$, $R^2 = 0.92$. Bianmengyan & cộng sự (2023) đã thực hiện một nghiên cứu về dự báo giá cổ phiếu của BYD bằng mô hình LSTM và ARIMA; Kết quả đã nêu bật hiệu suất vượt trội của LSTM trong việc nắm bắt các động lực của thị trường phức tạp, hiệu suất tốt nhất là $MAE = 0.45$, $MSE = 0.58$, $RMSE = 0.76$, $R^2 = 0.85$. Qua đánh giá, một khoảng trống có thể đào sâu nghiên cứu là cần tìm ra các kiến trúc mô hình học sâu tốt nhất, phù hợp với bộ dữ liệu giá cổ phiếu BYDDY đóng cửa. Nghiên cứu này có thể dẫn đến các mô hình LSTM tối ưu dự đoán chính xác và mạnh mẽ hơn, mang lại lợi ích cho các nhà đầu tư, nhà giao dịch và nhà hoạch định chính sách.

3. Phương pháp nghiên cứu

- **Bài toán:** dự báo giá cổ phiếu BYDDY đóng cửa và đưa ra các quyết định giao dịch sử dụng LSTM và thuật toán di truyền GA.

Phần 1: sử dụng thuật toán Di truyền GA tìm bộ siêu tham số tốt nhất để xây dựng mô hình LSTM tối ưu phù hợp với dữ liệu lịch sử giá cổ phiếu BYDDY đóng cửa.

Phần 2: sử dụng mô hình LSTM tối ưu đã huấn luyện dự báo giá cổ phiếu BYDDY đóng cửa và ra quyết định giao dịch.

- **Dữ liệu:** sử dụng duy nhất lịch sử giá cổ phiếu BYDDY đóng cửa (Close) từ ngày 30/12/2009 đến ngày 09/7/2024, trích từ dữ liệu cổ phiếu BYDDY trên Yahoo Finance. Tập dữ liệu giá cổ phiếu BYDDY đóng cửa được sử dụng để huấn luyện và đánh giá mô hình LSTM tối ưu, gồm 3.653 bản ghi.

- **Mô hình LSTM:** thiết kế đặc biệt để xác định các mẫu phức tạp trong dữ liệu chuỗi thời gian như dự đoán giá cổ phiếu. LSTM có các kết nối phản hồi cho phép chúng xử lý không chỉ các điểm dữ liệu đơn lẻ mà còn toàn bộ chuỗi dữ liệu. Đối với mô hình dự đoán giá cổ phiếu đóng cửa, LSTM xem xét trình tự lịch sử của giá cổ phiếu đóng cửa để đưa ra dự đoán về giá đóng cửa trong tương lai, từ đó dự báo các giao dịch tương ứng. Một đơn vị LSTM được thiết kế để theo dõi các phụ thuộc trong chuỗi đầu vào thông qua các cấu trúc gọi là cổng điều chỉnh luồng thông tin. Các cổng này quyết định những gì cần giữ lại trong bộ nhớ và những gì cần quên dựa trên đầu vào hiện tại và trạng thái trước đó. Vai trò các hàm số trong mô hình LSTM như sau:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (2)$$

$$C_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * C_t \quad (4)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

Cổng quên (1) quyết định thông tin nào từ nút trạng thái nên được loại bỏ hoặc giữ lại. Cổng đầu vào (2) cập nhật thông tin mới cho nút trạng thái. Nút đầu vào (3) tạo ra một vectơ các giá trị ứng cử viên mới có thể được thêm vào nút trạng thái. Nút cập nhật trạng thái (4) đây là bộ nhớ của LSTM được cập nhật bằng cách quên đi những gì không cần thiết và thêm các giá trị ứng cử viên mới theo mức độ quan trọng của chúng. Cổng đầu ra (5) nút trạng thái được chuyển qua Tanh để đẩy các giá trị nằm trong khoảng từ $[-1;1]$ đem nhân với đầu ra của cổng Sigmoid và xuất ra những quyết định. Trạng thái đầu ra ẩn (6) là đầu ra dựa trên nút trạng thái nhưng chỉ một số phần nhất định được phép xuất ra, được điều khiển bởi .

Trọng số W và phương sai b là 2 ma trận tham số và véc tơ tương ứng. Trạng thái ẩn trước đó và đầu vào hiện tại được huấn luyện trong quá trình học. Khả năng duy trì nhớ dài hạn của LSTM do nút trạng thái và cách các cổng tương tác để sửa đổi bộ nhớ này. Hàm Tanh $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ là hàm kích hoạt và Adam là thuật toán tối ưu hóa việc học thích hợp với bài toán có độ biến thiên không ổn định.

- **Genetic Algorithm:** là thuật toán Di truyền (GA) sử dụng tìm kiếm bộ siêu tham số tốt nhất giúp xây dựng mô hình LSTM tối ưu tương thích với bộ dữ liệu lịch sử cổ phiếu BYDDY đóng cửa. Quy trình tối ưu hóa các siêu tham số bằng GA bao gồm chọn các siêu tham số muốn tối ưu hóa, định nghĩa hàm mục tiêu là hiệu suất của mô hình trên tập kiểm tra, tạo ra một quần thể ban đầu với các cá thể đại diện cho các bộ siêu tham số ngẫu nhiên, huấn luyện mô hình với từng bộ siêu tham số và tính toán giá trị hàm mục tiêu, chọn các cá thể tốt nhất thực hiện lai ghép và đột biến để tạo ra các thế hệ tiếp theo. Cuối cùng, lặp lại các bước trên cho đến khi đạt được giá trị hàm mục tiêu mong muốn.

- **Hồi quy tuyến tính:** dự báo giá cổ phiếu BYDDY đóng cửa dẫn đến phương trình hồi quy tuyến tính $y = a \cdot X + b$, trong đó X là yếu tố dự đoán, y là giá trị dự đoán được tính từ a , b và X . Mục tiêu của hồi quy tuyến tính là tìm ra đường phù hợp nhất cho dữ liệu BYDDY, điều này sẽ dẫn đến giá trị $\hat{y}$ dự đoán gần với giá trị y lịch sử. Hệ số b là giá trị dự đoán của biến phụ thuộc khi các biến độc lập bằng 0, a là độ dốc, cả hai đều là giá trị ước tính từ hồi quy.

- Đánh giá mô hình LSTM tối ưu:

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

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}, \quad R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$$\text{và } STD = \sqrt{\frac{1}{n} \sum_{i=1}^n (\text{error}_i - \overline{\text{error}})^2}$$

4. Phân tích thực nghiệm

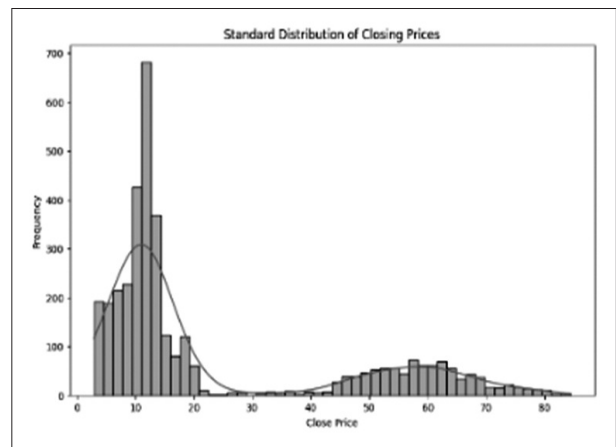
4.1. Tiền xử lý dữ liệu

Tập dữ liệu cổ phiếu BYDDY đóng cửa gồm 3.653 bản ghi được chuẩn hóa về khoảng (0;1) giúp mô hình học tốt hơn. Chia tập dữ liệu huấn luyện là 80% và tập dữ liệu kiểm tra là 20%. Sự chênh lệch lớn giữa giá trị nhỏ nhất \$2.91 và giá trị lớn nhất \$84.42 cho thấy cổ phiếu BYDDY đóng cửa có sự biến động mạnh trong thời gian dài. Độ lệch chuẩn cao \$21.60 cũng chỉ ra rằng giá cổ phiếu BYDDY có sự biến động lớn xung quanh giá trị trung bình \$23.25.

4.2. Trực quan hóa dữ liệu

Biểu đồ Histogram hỗ trợ quan sát giá cổ phiếu BYDDY đóng cửa được phân bố theo tần suất. Chia biểu đồ thành 50 bins để thấy rõ các dao động nhỏ trong phân phối giá. Biểu đồ có dạng hình chuông và tập trung vào một giá trị trung tâm, chứng tỏ cổ phiếu BYDDY thường dao động quanh giá trị trung bình, có sự ổn định tương đối. Đường KDE xác định phân phối xác suất của giá đóng cửa tuân theo phân phối chuẩn và dạng tổng quát của phân phối so với biểu đồ. (Hình 1)

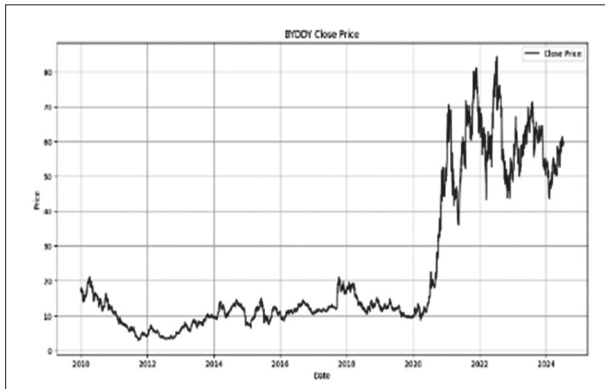
Hình 1: Phân phối chuẩn giá BYDDY đóng cửa



Nguồn: <https://www.kaggle.com/code/thuntnb/lstm-fine-tuned-best-model-byddy-pred?scriptVersionId=187689470>

Đồ thị giá cổ phiếu BYDDY đóng cửa cung cấp thông tin trực quan về biến động giá đóng cửa của cổ phiếu BYDDY trong khoảng thời gian 15 năm. Từ đồ thị, nhà đầu tư có thể quan sát xu hướng chung của giá cổ phiếu, nhận diện các đợt tăng

Hình 2: Xu hướng giá cổ phiếu BYDDY đóng cửa

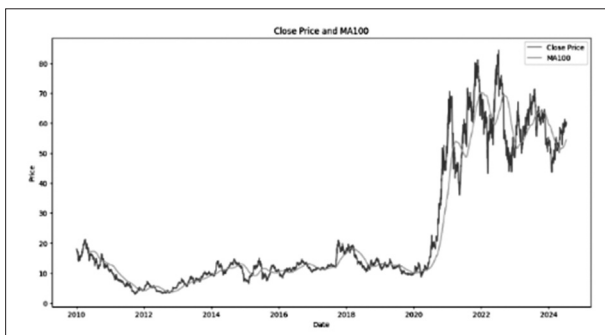


Nguồn: <https://www.kaggle.com/code/thuntrnb/lstm-fine-tuned-best-model-byddy-pred?scriptVersionId=187689470>

giảm và đánh giá biến động của giá cổ phiếu BYDDY. (Hình 2)

Đường trung bình động MA100 dựa trên 100 phiên giao dịch gần nhất và so sánh với đồ thị lịch sử giá cổ phiếu BYDDY đóng cửa, khi đường giá đóng cửa cao hơn đường trung bình động MA100, báo hiệu giá cổ phiếu đóng cửa tăng hoặc đà tăng, nhà đầu tư xem đây là tín hiệu mua vào. (Hình 3)

Hình 3: Đường trung bình động MA100



Nguồn: <https://www.kaggle.com/code/thuntrnb/lstm-fine-tuned-best-model-byddy-pred?scriptVersionId=187689470>

Khi đường giá cổ phiếu BYDDY đóng cửa dưới MA100, đây là xu hướng giảm hoặc đà giảm, nhà đầu tư nên bán hoặc tránh mua.

4.3. Tìm bộ siêu tham số tối ưu của mô hình LSTM sử dụng GA

Thuật toán Di truyền (GA) là một phương pháp tối ưu hóa dựa trên nguyên lý tiến hóa tự nhiên, bao gồm chọn lọc tự nhiên, lai ghép (crossover), và đột

biến (mutation). Sử dụng GA để tối ưu hóa một tập các siêu tham số cho 1 mô hình LSTM.

Hàm fitness_function là một hàm mục tiêu mà GA sẽ tối ưu hóa. Trong bài toán, hàm fitness được sử dụng để tối thiểu hóa tổng bình phương của các thành phần trong solution (mỗi solution sẽ là một tập hợp các giá trị đại diện cho các siêu tham số). Mục tiêu của GA là tìm ra giải pháp (tức là tập hợp các siêu tham số) làm cho giá trị của hàm fitness là nhỏ nhất. (Bảng 1)

Bảng 1. Hàm mục tiêu fitness_function

```
def fitness_function(solution):
    # Example fitness function: minimize the sum
    # of squares
    fitness = sum(x ** 2 for x in solution)
    return fitness
```

Nguồn: Tác giả

Bảng 2. Mã Python cài đặt thuật toán GA

```
algorithm_param = {'max_num_iteration': 50,
                   'population_size': 10,
                   'mutation_probability': 0.1,
                   'elit_ratio': 0.01,
                   'crossover_probability': 0.5,
                   'parents_portion': 0.3,
                   'crossover_type': 'uniform',
                   'max_iteration_without_improv':
                   None}
```

Nguồn: Tác giả

Số lượng vòng lặp tối đa mà GA sẽ chạy là 50 vòng, số lượng cá thể trong quần thể mỗi thế hệ 10 cá thể, xác suất một cá thể bị đột biến 10%, tỷ lệ các cá thể tốt nhất được giữ lại qua các thế hệ 1%, xác suất lai ghép giữa hai cá thể 50%, tỷ lệ các cá thể được chọn làm cha mẹ 30%, kiểu lai ghép được sử dụng là lai ghép đều (uniform) và số vòng lặp tối đa không có cải thiện trước khi dừng (để None tức là không giới hạn). (Bảng 3)

Ở đây, fitness_function là hàm mục tiêu cần tối ưu, Dimension số lượng siêu tham số cần tối ưu là 5, variable_type cho biết kiểu dữ liệu của các siêu tham số (real cho các giá trị thực), variable_boundaries giới hạn cho các siêu tham số

Bảng 3. Mã Python chạy thuật toán GA

```
model = ga(function = fitness_function,
dimension = 5, variable_type = 'real',
variable_boundaries = varbound,
algorithm_parameters = algorithm_param)
model.run()
```

Nguồn: Tác giả

và algorithm_parameters gồm các tham số của thuật toán GA được định nghĩa. Thuật toán di truyền GA sẽ tối ưu hóa một tập hợp các siêu tham số cho mô hình LSTM, dựa trên hàm mục tiêu (fitness function). Quá trình bao gồm việc khởi tạo quần thể ban đầu, chọn lọc, lai ghép, đột biến, và tìm ra giải pháp tốt nhất thông qua nhiều thế hệ. (Bảng 4)

Bảng 4. Siêu tham số tối ưu tìm thấy

Num_layers	1
Units	51
Dropout_rate	0.0077
Learning_rate	0.0008
Activation	Tanh
N_lag	60

Nguồn: <https://www.kaggle.com/code/thuntnb/ga-for-fine-tuning-lstm>

4.4. Huấn luyện mô hình LSTM tối ưu

Mô hình LSTM một lớp có số đơn vị (units) trong lớp LSTM là 51 đơn vị ẩn no-ron. Số lượng đơn vị này quyết định khả năng học và lưu trữ thông tin của mô hình. Kỹ thuật dropout được sử dụng với tỷ lệ là 0.0077, mô hình không sử dụng nhiều dropout vì không có vấn đề học lệch. Tốc độ học learning rate quyết định bước nhảy trong quá trình tối ưu hóa là 0.0008, mô hình LSTM sẽ học từ từ và cẩn thận, giúp tránh bỏ qua các cực tiểu cục bộ và đảm bảo sự hội tụ ổn định. Hàm kích hoạt được sử dụng trong LSTM tối ưu là Tanh, một hàm phi tuyến có đầu ra trong khoảng $[-1, 1]$, giúp mô hình có thể học các mẫu phức tạp hơn. Số bước thời gian 60 lags, mô hình sử dụng thông tin của 60 ngày trước đó để dự báo giá trị ngày hiện tại. Một lớp

Dense được thêm vào cuối để đưa ra một dự đoán duy nhất. (Bảng 5)

Bảng 5. Mã Python xây dựng mô hình LSTM tối ưu

```
# Xây dựng mô hình LSTM tối ưu
model = Sequential()
model.add(LSTM(units = 51, input_shape =
(X_train.shape[1], X_train.shape[2]),
activation = 'tanh'))
model.add(Dropout(0.0077))
model.add(Dense(1))
model.compile(optimizer = Adam(0.0008),
loss = 'mse')
```

Nguồn: Tác giả

Tổng số tham số của mô hình LSTM tối ưu là 32,594, số lượng tham số huấn luyện là 10,864 và tham số tối ưu hóa là 21,730 mô hình có thể học và điều chỉnh để cải thiện hiệu suất. Tất cả các tham số đều có thể huấn luyện, chứng tỏ mô hình linh hoạt và có khả năng học tập từ dữ liệu. Tổng kích thước bộ nhớ 127.32 KB, điều này thuận tiện cho việc triển khai mô hình. (Bảng 6)

Bảng 6. Mã Python huấn luyện mô hình LSTM tối ưu

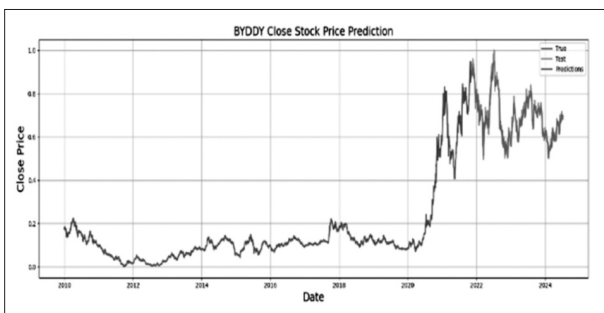
```
# Huấn luyện mô hình LSTM tối ưu
history = model.fit(X_train, y_train, epochs =
10, batch_size = 16, validation_data = (X_test,
y_test), verbose = 1)
```

Nguồn: Tác giả

Mô hình LSTM tối ưu là một mạng nơ-ron sử dụng hàm model.fit() là phương thức dùng để huấn luyện mô hình, giúp cập nhật trọng số của mô hình dựa trên dữ liệu huấn luyện và tối ưu hóa hàm mất mát. X_train dữ liệu đầu vào (features) của tập huấn luyện. Mô hình sẽ học từ dữ liệu này để tìm ra các mẫu và mối quan hệ trong dữ liệu. y_train giá trị mục tiêu tương ứng với dữ liệu đầu vào X_train. Mô hình học cách ánh xạ từ X_train đến y_train. Số lượng epoch = 10 là số lần toàn bộ dữ liệu huấn luyện được duyệt qua mô hình. Kích thước batch_size = 16 mỗi lần cập nhật trọng số sẽ dựa

trên 16 mẫu dữ liệu. Dữ liệu xác thực $\text{validation_data} = (X_{\text{test}}, y_{\text{test}})$ dùng để đánh giá mô hình trong quá trình huấn luyện. Đây là dữ liệu không được sử dụng trong quá trình huấn luyện, nhưng được dùng để kiểm tra hiệu suất của mô hình sau mỗi epoch. X_{test} và y_{test} là dữ liệu đầu vào và nhãn của tập dữ liệu kiểm tra (validation set). Với $\text{verbose} = 1$, thông tin về quá trình huấn luyện sẽ được in ra sau mỗi epoch, cho thấy sự tiến triển của mô hình. (Hình 4)

Hình 4: Mô hình LSTM tối ưu

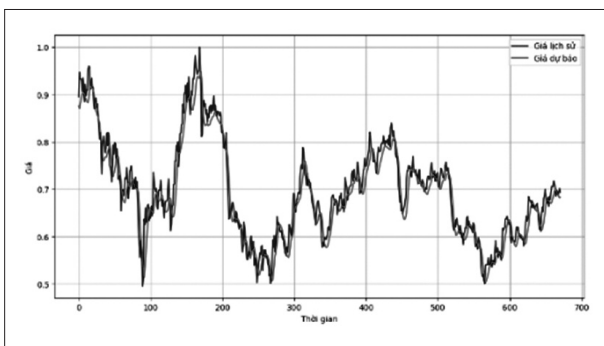


Nguồn: <https://www.kaggle.com/code/thuntrnb/lstm-fine-tuned-best-model-byddy-pred?scriptId=187689470>

4.5. Đánh giá mô hình LSTM tối ưu

Mô hình LSTM tối ưu đã huấn luyện sử dụng hàm $y_{\text{pred}} = \text{model.predict}(X_{\text{test}})$ để dự đoán các giá trị đầu ra (y_{pred}) dựa trên dữ liệu đầu vào X_{test} từ tập dữ liệu kiểm tra. Kết quả là các giá trị dự đoán được mô hình tính toán, đồ thị giá cổ phiếu BYDDY đóng cửa được dự báo bám sát đồ thị giá đóng cửa thực tế. (Hình 5)

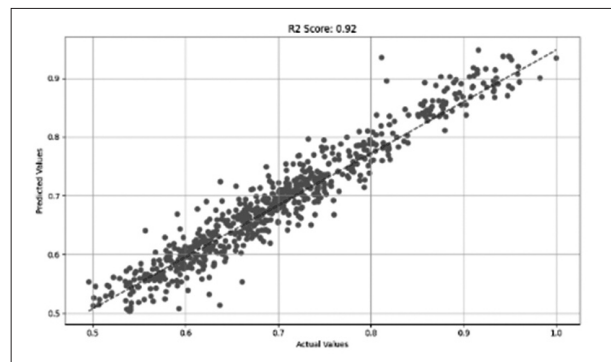
Hình 5: So sánh giá thực và giá dự báo



Nguồn: <https://www.kaggle.com/code/thuntrnb/lstm-fine-tuned-best-model-byddy-pred?scriptId=187689470>

Mô hình LSTM sau khi tối ưu hóa đã đạt được mức độ phù hợp rất cao với hệ số xác định là 0.92, nghĩa là mô hình có khả năng giải thích được 92% sự biến thiên trong dữ liệu, chỉ còn lại 8% là do nhiễu. Mô hình LSTM tối ưu đã chứng tỏ sự hiệu quả vượt trội trong việc dự báo, mang lại niềm tin vững chắc cho các quyết định dựa trên kết quả từ mô hình. (Hình 6)

Hình 6: Độ phù hợp của mô hình LSTM



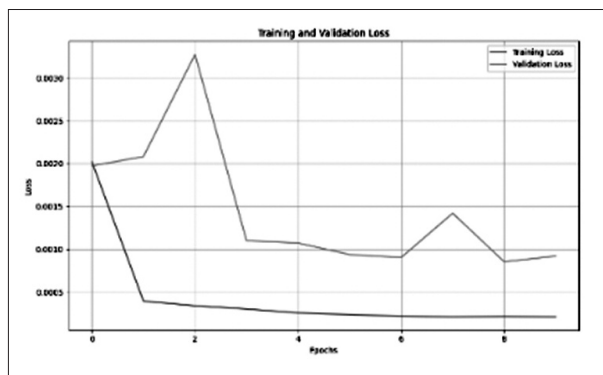
Nguồn: <https://www.kaggle.com/code/thuntrnb/lstm-fine-tuned-best-model-byddy-pred?scriptId=187689470>

Hiệu suất của mô hình LSTM tối ưu còn được thể hiện qua độ lệch chuẩn thấp ($\text{STD} = 0.015$), cho thấy các giá trị dự đoán ít biến động và rất sát với giá trị trung bình, chứng tỏ mô hình có tính ổn định cao. Với $\text{MAE} = 0.025$, mô hình đưa ra các dự đoán gần với thực tế, hỗ trợ nhà đầu tư dự đoán chính xác hơn giá đóng cửa của cổ phiếu BYDDY. Chỉ số $\text{MSE} = 0.001$ thể hiện mức độ chính xác của dự đoán và giúp giảm thiểu rủi ro do sai số lớn. $\text{RMSE} = 0.031$ khẳng định độ chính xác cao của mô hình trong việc dự đoán giá.

Hàm lỗi của mô hình LSTM tối ưu có xu hướng hội tụ về 0. Khi hàm lỗi của mô hình LSTM có xu hướng hội tụ về 0, mô hình đang dần học được cách dự đoán một cách chính xác hơn và dự đoán của mô hình càng gần với giá trị thực tế. Việc lỗi hội tụ về 0 là dấu hiệu quá trình tối ưu hóa đang diễn ra thành công. (Hình 7)

Sự hội tụ ổn định của hàm lỗi là dấu hiệu mô hình đã đạt đến điểm tối ưu trong không gian tham số của nó, mô hình không còn cải thiện nhiều nữa và đã đạt được hiệu suất tốt nhất có thể với tập dữ liệu.

Hình 7: Hàm lỗi hội tụ về 0



Nguồn: <https://www.kaggle.com/code/thuntrnb/lstm-fine-tuned-best-model-byddy-pred?scriptId=187689470>

4.6. Ứng dụng mô hình LSTM tối ưu trong dự báo

Bảng 7. Mã Python dự báo giá từng ngày trong tương lai

```
future_predictions = []
for i in range(future_days):
    next_pred = model.predict(last_data)
    future_predictions.append(next_pred[0, 0])
    last_data = np.roll(last_data, -1)
    last_data[0, -1, 0] = next_pred
```

Nguồn: Tác giả

Mô hình LSTM tối ưu đã huấn luyện được sử dụng mô hình để dự báo giá cổ phiếu trong 5 ngày tiếp theo bắt đầu từ ngày 10/07/2024 và ra quyết định mua, bán, giữ. (Bảng 7, 8)

Bảng 8. Mô hình LSTM tối ưu dự báo

Mô hình	Ngày dự báo	Giá dự báo	Hành động
LSTM	11-07-2024	59.62 \$	Hold
LSTM	12-07-2024	59.63\$	Buy
LSTM	15-07-2024	59.62\$	Sell
LSTM	16-07-2024	59.61\$	Sell
LSTM	17-07-2024	59.60\$	Sell

Nguồn: <https://www.kaggle.com/code/thuntrnb/lstm-fine-tuned-best-model-byddy-pred?scriptId=187689470>

Kết quả dự báo chính xác của mô hình giúp chúng ta hiểu đúng đắn và hành động kịp thời, từ đó các nhà đầu tư có thể tối đa hóa lợi nhuận và giảm thiểu rủi ro trên thị trường chứng khoán. Mô hình tối ưu đã huấn luyện dự báo tín hiệu giữ (hold) trong ngày 11/07/2024 và dự báo mua (buy) trong ngày 12/07/2024, nghĩa là tín hiệu giá cổ phiếu tăng. (Hình 8)

Tín hiệu bán (sell) trong 3 ngày 15,16 và 17/07/2024 cho thấy giá cổ phiếu đóng cửa dự kiến sẽ giảm. Các kết quả này phù hợp với dự báo của BYD Company Limited được thực hiện bởi 2 nhà phân tích dự báo của Wall Street, link kiểm tra: “<https://www.tipranks.com/stocks/byddy/forecast>”.

4. Kết quả và thảo luận

Thuật toán di truyền GA đã tìm ra bộ siêu tham số tốt nhất để xây dựng mô hình LSTM tối ưu phù hợp với bộ dữ liệu BYDDY. Công tác huấn luyện mô hình LSTM tối ưu đã cho những kết quả về hiệu suất đáng khích lệ với $STD=0.015$, $MAE=0.025$, $MSE=0.001$, $RMSE=0.031$ và $=0.92$. Kết quả này khẳng định mô hình LSTM được tối ưu hóa bởi thuật toán GA trên dữ liệu lịch sử của BYDDY đóng cửa cho dự báo chính xác và tính ổn định cao. Mức độ phù hợp của mô hình LSTM tối ưu là $=0.92$, dự đoán của mô hình bám sát giá cổ phiếu thực tế trong thời gian thử nghiệm, nắm bắt được cả những biến động ngắn hạn và xu hướng dài hạn. Mô hình đã dự báo thành công giá cổ phiếu BYDDY đóng cửa trong tương lai với hiệu suất cao. So sánh với phương pháp truyền thống, sử dụng thuật toán di truyền GA để xây dựng mô hình ARIMA tối ưu dự báo giá cổ phiếu BYDDY đóng cửa và ra quyết định giao dịch. Kết quả, mô hình ARIMA tối ưu đã huấn luyện có các đánh giá về hiệu suất như Bảng 9.

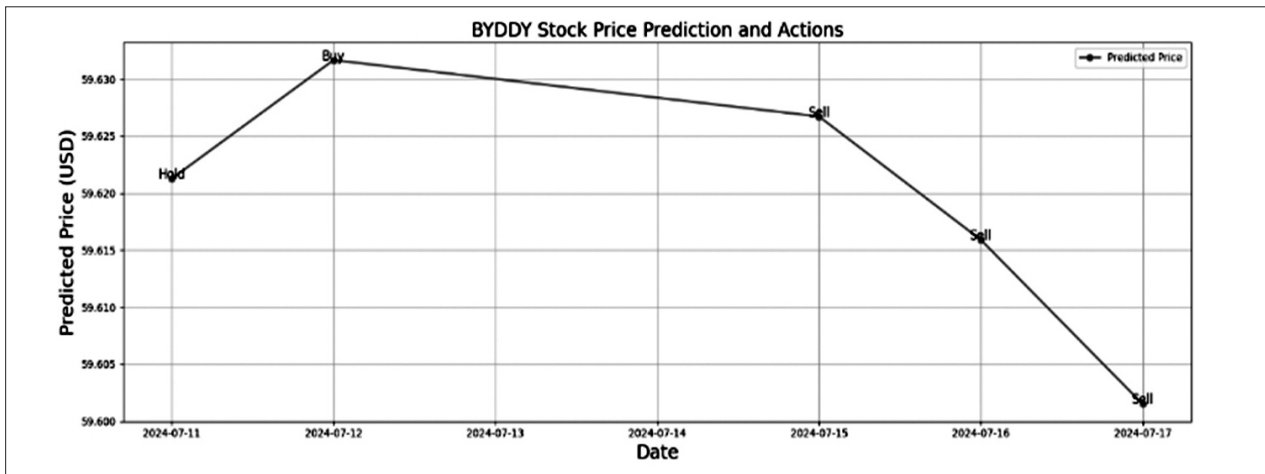
Kết quả này chứng tỏ mô hình ARIMA cho dù được tối ưu hóa vẫn không phù hợp với nguồn dữ

Bảng 9. Đánh giá hiệu suất của ARIMA tối ưu

MAE	MSE	RMSE	R ² Score	STD
40.67	1920.51	43.82	-5.82	16.43

Nguồn: <https://www.kaggle.com/code/thuntrnb/lstm-fine-tuned-best-model-byddy-pred?scriptId=187689470>

Hình 8: Dự báo hành động mua bán giữ trong 5 ngày tới



Nguồn: <https://www.kaggle.com/code/thuntrnb/lstm-fine-tuned-best-model-byddy-pred?scriptVersionId=187689470>

liệu lớn phi tuyến của giá cổ phiếu BYDDY đóng cửa, các dự đoán sai lệch lớn. So sánh với các kết quả nghiên cứu của tác giả Bianmengian Ji & cộng sự (2023), trong bài viết có tựa “The Stock Price of BYD Prediction using LSTM and ARIMA”, hiệu suất tốt nhất mô hình trong nghiên cứu của tác giả là $MAE=0.45$, $MSE=0.58$, $RMSE=0.76$, $R^2=0.85$. Vậy, mô hình LSTM tối ưu bởi GA trong nghiên cứu này chứng minh sự hiệu quả trong việc xử lý và học hỏi từ dữ liệu lớn chuỗi thời gian dài, có khả năng nắm bắt các mối quan hệ phi tuyến tính giữa các điểm dữ liệu đã góp phần mang lại hiệu suất dự đoán cao.

5. Kết luận

Mô hình LSTM tối ưu nắm bắt được sự phụ thuộc phức tạp về giá cổ phiếu BYDDY đóng cửa trong tập dữ liệu lịch sử. Dự đoán giá cổ phiếu BYDDY đóng cửa và các giao dịch của LSTM tối ưu bởi GA, do đó chính xác và đáng tin cậy hơn. Nghiên cứu phát hiện chính là mô hình LSTM tối

ưu dự báo giá cổ phiếu BYDDY đóng cửa và các quyết định giao dịch đạt hiệu suất cao hơn so với kết quả nghiên cứu dự báo giá BYD sử dụng mô hình LSTM không được tối ưu hóa và ARIMA truyền thống của các nghiên cứu trước đây. Mô hình LSTM tối ưu có thể hỗ trợ các nhà đầu tư phát triển các chiến lược giao dịch hiệu quả hơn, cho phép quản lý những biến động giá bất lợi. Phương pháp nghiên cứu có thể được áp dụng cho các cổ phiếu và công cụ tài chính khác. Tuy nhiên, vẫn còn những hạn chế, mô hình LSTM tối ưu cần nguồn dữ liệu lớn, thời gian huấn luyện dài. Việc sử dụng mô hình học sâu vào phân tích dự báo tài chính đòi hỏi các nhà hoạch định chính sách phải cân nhắc kỹ lưỡng, để đảm bảo công nghệ được khai thác một cách hiệu quả và có đạo đức.

Hướng nghiên cứu mới: sử dụng mô hình Support Vector Regression và LSTM tối ưu hóa bởi thuật toán GA phân tích dữ liệu cảm xúc và dự báo giá cổ phiếu ■

TÀI LIỆU THAM KHẢO:

- Đỗ Văn Thành (2017). Mô hình hóa dự báo giá cổ phiếu trong ngữ cảnh dữ liệu có số chiều cao. Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017.
- Phùng Đại Khánh, Nguyễn Xuân Huy (2022). Dự báo khai thác dầu khí sử dụng giải thuật di truyền (GA) dựa trên việc huấn luyện mạng nơ ron hồi quy có bộ nhớ ngắn hạn định hướng dài hạn (LSTM). Tạp chí Khí tượng Thủy văn, EME4, 355-361

3. Trần Đăng Tuyên (2023). Đánh giá hiệu suất mô hình phức hợp LSTM-GRU: Nghiên cứu điển hình về dự báo chỉ số đo lường xu hướng biến động giá cổ phiếu trên Sàn Giao dịch chứng khoán Hồ Chí Minh. Tạp chí Khoa học Đại học Cần Thơ, 60(1), 235-249.
4. Vũ Thị Loan, Nguyễn Thị Anh Thu, Nông Thị Hương Ly và Vũ Mai Linh (2023). Dự báo giá cổ phiếu trước các thông tin tài chính và phi tài chính: Áp dụng phân tích Event Study và Random Forest. VNU Journal of Economics and Business, 3(3), 94-104.
5. Bao, W. & Yue, J. & Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. PLoS ONE, 12(7), e0180944.
6. Bianmengyan J. & Xiyuan T. & Xiaohan W. (2023). The stock price of BYD prediction using LSTM and ARIMA. Financial Engineering and Risk Management, 6, 73-79.
7. Fischer, T. & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions, European Journal of Operational Research, 270(2), 654-669.

Ngày nhận bài: 18/7/2024

Ngày phản biện đánh giá và sửa chữa: 26/7/2024

Ngày chấp nhận đăng bài: 18/8/2024

Thông tin tác giả:

TRẦN BÁ THUẤN

Trường Đại học Kinh tế, Đại học Huế

Email: tbthuan@hce.edu.vn

FORECASTING BYDDY STOCK PRICE BY USING THE OPTIMAL LSTM MODEL AND GENETIC ALGORITHM

● TRAN BA THUAN

University of Economics, Hue University

ABSTRACT:

This study focused on the BYDDY stock price forecast of BYD Company Limited, China's leading manufacturer of electric cars, rechargeable batteries, and solar panels, through using the optimal long-sshort-time memory model (LSTM). The study's results show that the optimal LSTM effectively captures the nonlinear fluctuations of the closing BYDDY stock price. The main performance metrics of the optimal LSTM model on the test data set are $STD=0.015$, $MAE=0.025$, $MSE=0.001$, and $RMSE=0.031$, and the model fit is $= 0.92$. The study's results not only highlight the effectiveness of using the optimal LSTM in forecasting stock prices but also provide a framework for future research in applying deep learning techniques to financial market prediction.

Keywords: stock price prediction, BYDDY stock price, Deep learning, Genetic Algorithm, LSTM.