# Đặt vấn đề

## Bối cảnh

### Định nghĩa việc phân tích dữ liệu chứng khoán

Phân tích dữ liệu chứng khoán là quá trình nghiên cứu, đánh giá và hiểu về các dữ liệu liên quan đến giá cổ phiếu, khối lượng giao dịch, chỉ số thị trường và các yếu tố khác liên quan đến thị trường chứng khoán.

Mục tiêu của phân tích dữ liệu chứng khoán là nắm bắt được xu hướng của thị trường chứng khoán và dự đoán giá cổ phiếu, đánh giá rủi ro và cơ hội đầu tư dựa trên thông tin đầu vào có sẵn.

### Phương pháp truyền thống và hạn chế

Các phương pháp truyền thống trong phân tích dữ liệu chứng khoán thường được sử dụng gồm có mô hình ARIMA, và các thuật toán học máy truyền thống (Traditional Machine Learning - ML). Các thuật toán này được ứng dụng rộng rãi trong lĩnh vực tài chính, tuy nhiên các phương pháp này lại có rất nhiều hạn chế như dễ bị nhiễu, gặp khó khăn trong việc xử lý chuỗi thời gian không dừng (non-stationary time series),  yêu cầu quá trình trích xuất các đặc trưng thủ công,…

## Deep learning trong dự báo giá chứng khoáng

### Tính mới của Deep learning

Học sâu (Deep learning) có khả năng học mạnh mẽ và phù hợp với các chuỗi thời gian phức tạp như chuỗi thời gian tài chính. Điều này đến từ cấu trúc liên kết phi tuyến sâu trong các mô hình học sâu, cũng như khả năng xuất sắc của chúng trong việc trích xuất thông tin liên quan từ chuỗi thời gian tài chính.

### Các phương pháp Deep learning phổ biến

Các mô hình học sâu như mạng neural hồi quy – RNN (Recurrent Neural Networks), mạng neural hồi quy dài hạn – LSTM (Long Short Term Memory) đã được áp dụng để phân tích dữ liệu chuỗi thời gian, đặc biệt trong lĩnh vực dự báo giá chứng khoán. RNN được thiết kế để duy trì và xử lý thông tin từ các sự kiện gần đây thông qua cấu trúc mạng có khả năng nhớ ngắn hạn, điều này làm cho RNN rất phù hợp với các nhiệm vụ dự báo chuỗi thời gian. Tuy nhiên, RNN gặp khó khăn với các vấn đề như gradient biến mất khi xử lý các chuỗi thời gian dài.

Để khắc phục những hạn chế này, LSTM, một biến thể tiên tiến của RNN, đã được phát triển. LSTM bao gồm các thành phần đặc biệt gọi là cổng (gates) bao gồm cổng đầu vào, cổng quên, và cổng đầu ra. Các cổng này giúp kiểm soát dòng chảy thông tin và lưu giữ thông tin cần thiết trong thời gian dài, do đó giải quyết được vấn đề của RNN. LSTM có khả năng xử lý không chỉ các điểm dữ liệu đơn lẻ mà còn cả các chuỗi dữ liệu dài và phức tạp.

Nhờ vào khả năng học hỏi từ lượng dữ liệu lớn và xử lý các đặc trưng phức tạp của chuỗi thời gian, các mô hình học sâu như RNN và LSTM đã nâng cao độ chính xác của dự báo tài chính. Sự phát triển này mở ra nhiều cơ hội mới trong việc cải thiện các chiến lược đầu tư và quản lý rủi ro tài chính, nhờ vào khả năng phân tích và dự đoán biến động của giá chứng khoán một cách hiệu quả.

## Mục tiêu phân tích

### Ý nghĩa bài phân tích

Bài phân tích nhằm nghiên cứu ứng dụng của 2 mô hình học sâu là RNN và LSTM trong việc dự báo xu hướng tăng hay giảm của giá chứng khoán trong tương lai. Nhóm nghiên cứu sẽ tiến hành xây dựng mô hình, đánh giá hiệu suất mô hình bằng các chỉ số như MSE (Mean Squared Error), RMSE (Mean Squared Error), MAPE (Mean Absolute Percentage Error), R-Square để tìm ra mô hình tốt nhất và dự báo giá chứng khoán của 3 ngân hàng khác nhau bằng mô hình đó.

### Câu hỏi phân tích

Nhóm xây dựng bài phân tích hướng đến mục tiêu trên bằng cách trả lời được các câu hỏi sau:

*- Mô hình nào tối ưu nhất cho việc dự đoán xu hướng của dữ liệu chuỗi thời gian trong tài chính?*

*- Ảnh hưởng của các tham số và cấu trúc mô hình đối với hiệu suất của mỗi mô hình?*

*- Window size nào cho kết quả tối ưu nhất đối với mỗi mô hình?*

*- Tỉ lệ dữ liệu huấn luyện và kiểm tra nào là phù hợp nhất đối với mỗi mô hình?*

## Phạm vi phân tích

### Thời gian

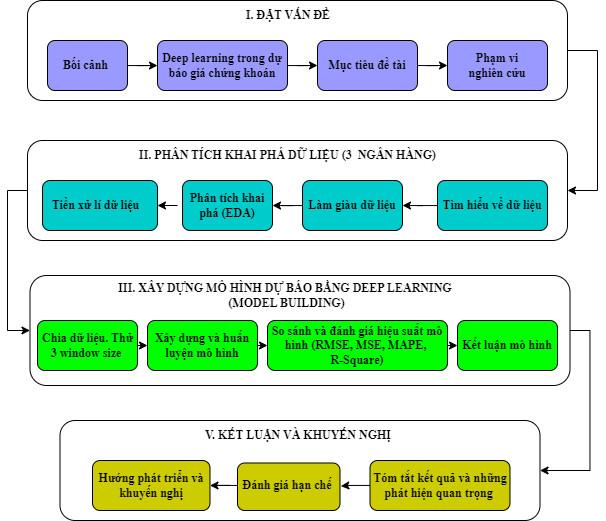
Bài phân tích sử dụng dữ liệu giá chứng khoán của 3 ngân hàng trong khoảng thời gian sau:

- Ngân hàng ACB, VCB: 2/1/2013 - 8/9/2023.

- Ngân hàng BIDV: 24/1/2014 - 8/9/2023.

### Không gian

Bài phân tích sử dụng tập dữ liệu giá chứng khoán của 3 ngân hàng ACB, VCB, BID bao gồm các thông tin như giá đóng, mở cửa, giá cao, thấp, khối lượng giao dịch, giá trị thay đổi và các cột giá trị hiệu chỉnh tương ứng của những cột trên.



*Hình 1.1. Sơ đồ quy trình phân tích*

# Phân tích khai phá dữ liệu (3 ngân hàng)

## 2.1. Tìm hiểu về dữ liệu (Data Understanding)

### 2.1.1. Giới thiệu về bộ dữ liệu

Nhóm sử dụng bộ dữ liệu gồm 3 tập dữ liệu về giá và các thông tin liên quan đến giao dịch cổ phiếu của 3 ngân hàng riêng biệt là Vietcombank (VCB), Asia Commercial Bank (ACB), Bank for Investment and Development of Vietnam (BIDV). Cả 3 tập dữ liệu đều gồm 25 cột và có 2740, 2737, 2470 dòng tương ứng lần lượt với VCB, ACB, BIDV.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

*Hình 2.1. Một phần của tập dữ liệu ACB ban đầu*

Vì cả 3 ngân hàng đều có 25 cột tức 25 biến giống nhau nên nhóm sẽ sử dụng một phần tập dữ liệu của ngân hàng ACB thể hiện trên hình 2.  để trình bày đại diện cho cả 3 tập.

### 2.1.2. Mô tả các biến

Sau đây là mô tả chi tiết các biến có trong tập dữ liệu của 3 ngân hàng trên.

- **code (Stock Code)** : Mã chứng khoán là mã định danh duy nhất của một cổ phiếu trên sàn giao dịch.

- **date**: Ngày giao dịch

- **time:** Thời gian giao dịch

- **floor**: Sàn giao dịch

- **type**: Loại giao dịch

- **basicPrice:** Giá cơ bản (giá tham chiếu hoặc giá định mức)

- **ceilingPrice**: Giá trần là mức giá cao nhất mà cổ phiếu có thể đạt được trong ngày giao dịch.

- **floorPrice**: Giá sàn là mức giá thấp nhất mà cổ phiếu có thể đạt được trong ngày giao dịch.

- **open (Opening Price)**: Giá mở cửa là giá cổ phiếu tại thời điểm mở cửa phiên giao dịch.

- **high (Highest Price):** Giá cao nhất của cổ phiếu trong phiên giao dịch.

- **low (Lowest Price):** Giá thấp nhất của cổ phiếu trong phiên giao dịch.

- **close (Closing Price)**: Giá đóng cửa là giá cổ phiếu tại thời điểm đóng cửa phiên giao dịch.

- **average (Average Price)**: Giá trung bình của cổ phiếu trong phiên giao dịch, thường tính bằng cách lấy tổng giá trị giao dịch chia cho tổng khối lượng giao dịch.

- **adOpen (Adjusted Opening)**: Giá điều chỉnh mở cửa là giá mở cửa của cổ phiếu đã được điều chỉnh sau các sự kiện như chia cổ tức, phát hành thêm cổ phiếu.

- **adHigh (Adjusted Highest Price):** Giá điều chỉnh cao nhất là giá cao nhất của cổ phiếu đã được điều chỉnh.

- **adLow (Adjusted Lowest Price)**: Giá điều chỉnh thấp nhất là giá  thấp nhất của cổ phiếu đã được điều chỉnh.

- **adClose (Adjusted Closing Price)**: Giá điều chỉnh đóng cửa là giá đóng cửa của cổ phiếu đã được điều chỉnh.

- **adAverage (Adjusted Average Price)**: Giá điều chỉnh trung bình là giá trung bình của cổ phiếu đã được điều chỉnh.

- **nmVolume (Normal Trading Volume)**: Khối lượng giao dịch thường là số lượng cổ phiếu được giao dịch thông qua các lệnh khớp lệnh thông thường trên sàn.

- **nmValue (Normal Trading Value)**: Giá trị giao dịch thường là tổng giá trị giao dịch của các lệnh khớp lệnh thông thường trên sàn.

- **ptVolume (Put-through Trading Volume):** Khối lượng giao dịch thỏa thuận Số lượng cổ phiếu được giao dịch thông qua các lệnh thỏa thuận trực tiếp giữa các bên.

**- ptValue (Put-through Trading Value):** Giá trị giao dịch thỏa thuận là tổng giá trị giao dịch của các lệnh thỏa thuận trực tiếp giữa các bên.

**- change (Price Change):** Thay đổi giá là mức thay đổi giá cổ phiếu so với phiên giao dịch trước đó.

**- adChange (Adjusted Price Change):** Thay đổi giá điều chỉnh là  mức thay đổi giá cổ phiếu đã được điều chỉnh so với phiên giao dịch trước đó.

**- pctChange (Percentage Price Change)**: Phần trăm thay đổi giá là tỷ lệ phần trăm thay đổi giá cổ phiếu so với phiên giao dịch trước đó.

Biến "code", "date", "time", "floor", "type" có kiểu dữ liệu là "chr" (chuỗi ký tự), các biến còn lại có kiểu kiểu dữ liệu là num (số).

## Điều chỉnh các cột

Khi tiến hành phân tích và dự đoán giá cổ phiếu trong tương lai, nhóm cần tập trung vào những biến có liên quan trực tiếp đến biến động giá và khối lượng giao dịch. Đồng thời, bộ dữ liệu chứa rất nhiều biến nên các biến chứa các chỉ số khác cũng nên được loại bỏ để đơn giản hóa bộ dữ liệu. Vì vậy, nhóm sẽ tiến hành tách các cột dữ liệu ra thành một DataFrame mới có dạng như hình 2. bên dưới bao gồm các cột: date, open, high, low, close, adClose, và nmVolume để dễ dàng theo dõi và xử lý các biến này hơn.

A screenshot of a computer

Description automatically generated

*Hình 2.2. Một phần của tập dữ liệu BID sau khi trích xuất các cột cần thiết*

### Biến dữ liệu cột close (và các cột cần thiết khác) thành chuỗi thời gian

Vì trong bộ dữ liệu, cả 3 tập dữ liệu từ 3 ngân hàng đều chỉ chứa một cột date (thời gian giao dịch) chứ vẫn chưa có định dạng chuỗi thời gian nên nhóm sẽ tiến hành biến đổi 3 tập dữ liệu trên thành dạng chuỗi thời gian bằng cách dùng cột “date” làm chỉ số cho tập dữ liệu tương ứng.

### Phân tích khám phá chuỗi thời gian

Trong phân tích chuỗi thời gian, việc phân tách và hiểu rõ các thành phần xu hướng (trend) và thời vụ (seasonality) là cực kỳ quan trọng.

*+ Xu hướng (Trend)*: Đây là thành phần thể hiện sự biến động dài hạn của tập dữ liệu. Phân tích xu hướng giúp ta thấy được liệu giá trị của chuỗi thời gian có xu hướng tăng, giảm, hay duy trì ổn định theo thời gian. Bằng cách xác định xu hướng, ta có thể đưa ra các dự đoán dài hạn và hiểu được xu hướng chung của tập dữ liệu.

*+ Thời vụ (Seasonality)*: Đây là thành phần phản ánh các biến động định kỳ trong chuỗi thời gian, thường xảy ra trong một chu kỳ cố định, chẳng hạn như hàng năm, hàng tháng, hay hàng tuần. Phân tích thời vụ giúp nhận diện các mẫu (pattern) lặp lại và hiểu được ảnh hưởng của các yếu tố theo mùa lên dữ liệu. Điều này sẽ giúp cho doanh nghiệp trong việc dự đoán xu hướng trong ngắn hạn và lập các kế hoạch kinh doanh theo chu kỳ.

Bằng cách phân tách chuỗi thời gian thành các thành phần xu hướng và thời vụ, ta có thể đưa ra các nhận định chính xác hơn và ứng dụng những hiểu biết này vào việc dự báo và ra quyết định kinh doanh. Vì vậy tiếp theo, nhóm sẽ tiến hành phân tích chi tiết các thành phần này của chuỗi thời gian giá đóng cửa của cổ phiếu các ngân hàng.

**a) Nội suy chuỗi thời gian còn thiếu**

Do tính chất của sàn chứng khoán chỉ giao dịch từ thứ Hai đến thứ Sáu (nghỉ thứ Bảy, Chủ Nhật) nên chuỗi thời gian trong tập dữ liệu đã giới thiệu ở trên không thể ghi nhận các giá trị liên tiếp làm cho khoảng cách giữa các giá trị ghi nhận không đều nhau. Điều này dẫn đến việc không thể phân rã chuỗi thời gian ra các thành phần như xu hướng thời vụ được nên nhóm sẽ sử dụng phương pháp Nội suy tuyến tính (Linear Interpolation) để làm đều khoảng cách thời gian giữa các giá trị trong chuỗi thời gian.

Phương pháp Nội suy tuyến tính này là một phương pháp nội suy trong đó các giá trị giữa hai điểm dữ liệu được ước tính bằng cách sử dụng một đường thẳng nối liền hai điểm đó, giả định rằng sự thay đổi giữa hai điểm dữ liệu là tuyến tính. Nhóm sẽ tạo ra một chuỗi thời gian cách nhau hàng ngày có giá trị liên tục từ ngày đầu đến ngày cuối của tập dữ liệu. Sau đó sử dụng phương pháp này để lấp đầy các giá trị của chuỗi thời gian vào những ngày không có giá trị ghi nhận do sàn chứng khoán không giao dịch thứ Bảy, Chủ Nhật.

**b) Phân rã chuỗi thời gian**

A graph of different types of lines

Description automatically generated with medium confidence

*Hình 2.3. Đồ thị phân rã chuỗi thời gian giá đóng cửa của ngân hàng VCB*

Dựa vào đồ thị trên có thể thấy giá đóng cửa của ngân hàng VCB thể hiện xu hướng tăng trưởng mạnh mẽ từ từ năm 2013 đến 2023, điều này thể hiện khả năng là xu hướng tăng của giá vẫn sẽ tiếp tục trong tương lai. Về tính mùa vụ, giá đóng cửa thể hiện chu kỳ biến động theo năm, bắt đầu tăng cao và đạt đỉnh vào khoảng tháng 3, 4 sau đó giảm mạnh vào khoảng giữa năm rồi lại tăng nhẹ vào khoảng tháng 8, 9, sau đó là giảm chạm đáy vào thời điểm cuối năm.

*A graph of different types of lines

Description automatically generated with medium confidence*

*Hình 2.4. Đồ thị phân rã chuỗi thời gian giá đóng cửa của ngân hàng ACB*

Đồ thị trên cho thấy giá đóng cửa của ngân hàng ACB có sự biến động mạnh qua các năm, vào năm 2019 giá đã tăng đạt đỉnh với mốc hơn 50, nhưng sau đó lại giảm mạnh, tuy vậy nhìn chung giá đóng cửa vẫn có xu hướng tăng nhẹ so với thời điểm năm 2013. Tính mùa vụ cũng thể hiện qua chu kỳ biến động lặp lại định kỳ là 1 năm, giá sẽ tăng mạnh và đạt đỉnh vào khoảng tháng 3 sau đó biến động lên xuống nhẹ rồi giảm mạnh và chạm đáy vào tháng 12.

A graph of different types of lines

Description automatically generated with medium confidence

*Hình 2.5. Đồ thị phân rã chuỗi thời gian giá đóng cửa của ngân hàng BIDV*

Đồ thị trên cho thấy giá đóng cửa của ngân hàng BIDV cũng có đà tăng trưởng mạnh mẽ qua các năm, tuy nhiên vào những năm gần đây thì xu hướng tăng có phần chững lại nhưng vẫn có khả năng giá sẽ tiếp tục tăng. Giá đóng cửa cũng có tính mùa vụ lặp lại theo năm giống 2 ngân hàng trên. Khác là giá đóng cửa của ngân hàng BIDV cũng tăng và đạt đỉnh vào khoảng tháng 3 tuy nhiên lại chạm đáy vào giữa năm tầm tháng 7, sau đó giá cũng chỉ dao động quanh mức đáy cho đến cuối năm.

**b) Phân tích đồ thị ACF và PACF**

*ACF (AutoCorrelation Function):* Biểu đồ ACF thể hiện mối tương quan giữa các giá trị của chuỗi thời gian tại các khoảng thời gian trễ (lags) khác nhau. Mỗi thanh trong biểu đồ ACF cho biết mức độ tương quan của chuỗi với chính nó tại một khoảng thời gian trễ cụ thể.

*PACF (Partial AutoCorrelation Function):* Biểu đồ PACF thể hiện mối tương quan giữa các giá trị của chuỗi thời gian tại các khoảng thời gian trễ khác nhau sau khi đã loại bỏ ảnh hưởng của các khoảng thời gian trễ trung gian. PACF giúp xác định số lượng lags cần thiết để giải thích đầy đủ mối tương quan trong dữ liệu.

Thông thường, hai đồ thị này thường được dùng để nhận diện tính chất của chuỗi thời gian và lựa chọn mô hình cũng như tham số của mô hình phù hợp với chuỗi  thời gian đó. Vì cả hai đồ thị ACF và PACF của 3 ngân hàng được thể hiện ở hình 2. bên dưới đều giống nhau nên nhóm sẽ rút ra nhận xét chung từ 2 đồ thị này cho cả 3 tập dữ liệu.

A graph with numbers and lines

Description automatically generatedA graph with numbers and lines

Description automatically generatedA graph of a bar chart

Description automatically generated with medium confidenceA graph of a price

Description automatically generated with medium confidenceA graph of a bar chart

Description automatically generated with medium confidenceA graph with numbers and a line

Description automatically generated

*Hình 2.6. Đồ thị ACF, PACF của chuỗi thời gian giá đóng cửa của cả 3 ngân hàng*

Đồ thị trên cho thấy giá trị ACF ở độ trễ 1 rất cao, gần bằng 0.9 và giá trị ACF giảm dần qua các độ trễ nhưng rất chậm. Điều này cộng với việc giá đóng cửa của cả 3 ngân hàng đều thể hiện xu hướng tăng và có tính mùa vụ rõ ràng cho thấy chuỗi thời gian giá đóng cửa của cổ phiếu trong trường hợp này là một chuỗi thời gian không dừng (non-stationary). Vì vậy việc áp dụng các phương pháp truyền thống như ARIMA hay các thuật toán học máy truyền thống sẽ gặp phải một số khó khăn nhất định do hạn chế của các mô hình này trong việc xử lý chuỗi thời gian không dừng. Vì vậy, mô hình học sâu mà nhóm đã chọn sẽ hiệu quả hơn trong việc xử lý chuỗi thời gian của bộ dữ liệu này.

### Tính và phân tích các chỉ số kinh tế

Phân tích các chỉ số kinh tế là một phần không thể thiếu trong việc ra quyết định đầu tư và quản lý tài chính. Những chỉ số như Đường Trung Bình Đơn Giản (SMA), Đường Trung Bình Lũy Thừa (EMA) và Dải Bollinger (Bollinger Bands) giúp các nhà đầu tư hiểu rõ hơn về xu hướng và biến động của thị trường, từ đó các nhà đầu tư có thể tối ưu hóa chiến lược giao dịch, giảm thiểu rủi ro và nâng cao lợi nhuận đầu tư.

**a) SMA - Simple Moving Average (Đường trung bình động đơn giản)**

*Đường Simple Moving Average (SMA):* là đường trung bình động đơn giản được tính bằng trung bình cộng các mức giá đóng cửa trong một khoảng thời gian giao dịch nhất định. Đây là chỉ báo kỹ thuật cơ bản tuy nhiên rất hiệu quả trong giao dịch, đặc biệt là ngắn hạn. Đường SMA có ý nghĩa quan trọng trong những dự đoán về xu hướng của giá; từ đó giúp nhà đầu tư thực hiện những giao dịch ngắn hạn xác suất thành công cao. Nhà đầu tư có thể quan sát hướng chuyển động của SMA và đưa ra những quyết định giao dịch.

Việc sử dụng kết hợp các đường SMA ngắn và dài hạn sẽ là chỉ báo phù hợp để xác nhận xu hướng giá, tìm ra điểm thay đổi xu hướng. Khi đường SMA ngắn hạn tạo điểm cắt từ dưới lên với các đường SMA trung - dài hạn là có thể là tín hiệu xác nhận xu hướng tăng trong ngắn hạn. Ngược lại, khi các đường MA ngắn hạn tạo điểm cắt từ trên xuống với các đường trung - dài hạn, điều này có thể là tín hiệu cho thấy xu hướng giá sẽ giảm xuống trong ngắn hạn.

A graph of a stock market

Description automatically generated

*Hình 2.7. Đồ thị thể hiện đường SMA30 và SMA300 của giá đóng cửa của ngân hàng ACB*

Đồ thị hình 2.7 cho thấy rõ việc sử dụng kết hợp các đường SMA ngắn và dài hạn có hiệu quả khá cao khi muốn dự đoán xu hướng giá trong ngắn hạn phục vụ cho việc đầu tư. Vào tháng 7/2017, có thể thấy đường SMA30 (màu đen) đã cắt từ dưới lên đường SMA300 (màu đỏ) và sau đó là đợt tăng trưởng của giá đóng cửa trong ngắn hạn. Ngược lại, vào tháng 7/2019, đường SMA30 đã cắt từ trên xuống đường SMA300 kéo theo đó là giá đã giảm trong ngắn hạn.

**b) EMA - Exponential Moving Average (Đường trung bình động lũy thừa)**

Đây là một công cụ phân tích kỹ thuật phản ánh sự biến động giá được tính theo cấp số nhân, qua đó giúp nhà đầu tư xác định được sự biến động giá một cách chuẩn xác hơn và giảm thiểu được tình trạng nhiễu giá so với đường trung bình động thông thường (SMA). Công cụ này rất nhạy cảm với các các tín hiệu bất thường trong ngắn hạn, nên kết quả cho ra sẽ chuẩn xác hơn so với đường SMA. EMA được nhà đầu tư sử dụng để phản ứng nhanh, kịp thời trước các biến động phát sinh bất ngờ.

Về công dụng thì EMA cũng có thể được dùng để xác định xu hướng tăng giảm của thị trường bằng cách kết hợp các đường EMA ngắn và dài hạn tương tự như SMA. Ngoài ra, nhờ khả năng cập nhật dữ liệu mới, theo kịp xu hướng về giá nhanh vượt bậc so với nhiều công cụ chỉ báo khác nên đường EMA còn được dùng để tìm điểm đặt lệnh thích hợp.

Khi đường EMA dốc lên và đường giá nằm trên đường này đang có dấu hiệu cắt xuống (chuyển hướng): Ngay khi 2 đường chạm vào nhau chính là thời điểm thích hợp để đặt lệnh mua. Nếu đường EMA dốc xuống, đường giá đang ở dưới và bắt đầu di chuyển lên: Thời điểm 2 đường cắt nhau nhà đầu tư nên đặt lệnh bán.

A graph of a stock market

Description automatically generated

*Hình 2.8. Đồ thị thể hiện đường EMA20 và EMA70 của giá đóng cửa của ngân hàng VCB*

Dựa vào công dụng của đường EMA kết hợp cùng đồ thị hình 2.8 có thể thấy khoảng tháng 4/2023, đường EMA đang dốc lên trên, đường giá (màu xanh lá) chuyển hướng cắt xuống là lúc thích hợp để nhà đầu tư đặt lệnh mua, và ngay sau đó là đà tăng mạnh của giá đóng cửa. Tương tự, vào tháng 6/2020, khi đường EMA đang dốc xuống, đường giá đang ở dưới và bắt đầu di chuyển lên, thời điểm 2 đường cắt nhau nhà đầu tư đặt lệnh bán sẽ tránh được lỗ vì sau đó giá đóng cửa giảm mạnh chạm đáy.

**c) Bollinger Bands**

Bollinger bands là công cụ kết hợp giữa đường trung bình động MA (Moving Average) và độ lệch chuẩn. Cấu tạo chỉ báo này bao gồm một đường trung bình động ở giữa và hai đường biên trên, biên dưới. Khoảng cách giữa đường MA với các dải bollinger của nó được xác định bởi mức độ biến động giá. Khi giá chứng khoán biến động mạnh, dải bollinger sẽ mở rộng và ngược lại khi giá chứng khoán biến động ít hơn, dải bollinger sẽ dần thu hẹp lại.

Cách tính Bollinger bands như sau: Dải giữa là đường trung bình động chu kỳ 20 ngày (SMA20), được tính bằng giá trị trung bình của giá đóng cửa. Dải trên = SMA20 ngày + 2 x Độ lệch chuẩn 20 ngày. Dải dưới = SMA20 ngày – 2 x Độ lệch chuẩn 20 ngày.

A graph with green and red lines

Description automatically generated

*Hình 2.9. Đồ thị thể hiện đường EMA20 và EMA70 của giá đóng cửa của ngân hàng VCB*

Bollinger bands đặc biệt hữu ích để xác định xem một cổ phiếu đang bị mua quá mức hay bán quá mức. Khi giá cổ phiếu bằng hoặc cao hơn dải trên, cổ phiếu có thể bị mua quá mức. Khi giá cổ phiếu bằng hoặc thấp hơn biên độ, cổ phiếu có thể bị bán quá mức. Một trong những phương pháp giao dịch phổ biến khác khi sử dụng chỉ báo Bollinger Bands là Bollinger Squeeze (dải bollinger bands siết chặt) là khi khoảng cách giữa dải trên và dải dưới đường SMA thu hẹp lại hay còn gọi là hiện tượng “ thắt nút cổ chai”, giá cổ phiếu giai đoạn này biến động thấp. Đây là một dấu hiệu cho biết giá sẽ biến động mạnh trong tương lai và có thể xuất hiện cơ hội giao dịch cổ phiếu. Tuy nhiên trong trường hợp này dải Bollinger bands không cho biết giá sẽ biến động theo chiều hướng tăng hay giảm.

## Phân tích khai phá (EDA)

### Phân tích thống kê mô tả và phân phối của các biến

**a) Phân tích thống kê mô tả các biến**

Kết quả:

**Thống kê mô tả của VCB:**

      date                 open             high

 Min.   :2013-01-02   Min.   : 23.40   Min.   : 24.00

 1st Qu.:2015-10-08   1st Qu.: 37.25   1st Qu.: 37.60

 Median :2018-07-07   Median : 57.95   Median : 58.80

 Mean   :2018-05-21   Mean   : 60.25   Mean   : 61.01

 3rd Qu.:2020-12-09   3rd Qu.: 82.90   3rd Qu.: 83.50

 Max.   :2023-09-08   Max.   :116.70   Max.   :117.20

      low             close.Index           close.X[[i]]

 Min.   : 23.30   Min.   :2013-01-02    Min.   : 23.30000

 1st Qu.: 36.99   1st Qu.:2015-10-08    1st Qu.: 37.30000

 Median : 56.95   Median :2018-07-07    Median : 58.00000

 Mean   : 59.55   Mean   :2018-05-21    Mean   : 60.29951

 3rd Qu.: 82.00   3rd Qu.:2020-12-09    3rd Qu.: 82.80000

 Max.   :115.00   Max.   :2023-09-08    Max.   :116.40000

    nmVolume           change

 Min.   :  36380   Min.   :-6.80000

 1st Qu.: 551852   1st Qu.:-0.50000

 Median : 898480   Median : 0.00000

 Mean   :1099822   Mean   : 0.04564

 3rd Qu.:1407410   3rd Qu.: 0.50000

 Max.   :7265480   Max.   : 5.70000

- Biến động giá: từ mức thấp nhất là khoảng 23.30 - 23.40 đến mức cao nhất khoảng 115.00 - 117.20, cho thấy cổ phiếu VCB đã trải qua một giai đoạn tăng trưởng mạnh mẽ.

- Phân phối giá: phân phối giá cổ phiếu có độ lệch khá lớn giữa các quý phân vị, đặc biệt là từ Q1 đến Q3 cho thấy sự biến động giá lớn. Điều này được xác nhận bởi giá trị trung bình (Mean) và giá trị trung vị (Median) cũng có sự chênh lệch.

- Biến động khối lượng giao dịch: Khối lượng giao dịch biến động mạnh, từ mức thấp nhất là 36,380 đến mức cao nhất là 7,265,480. Điều này cho thấy có những giai đoạn mà cổ phiếu VCB được giao dịch rất sôi động.

- Phân phối khối lượng giao dịch: Giá trị trung vị (898,480) và giá trị trung bình (1,099,822) cũng cho thấy khối lượng giao dịch khá đồng đều nhưng có những ngày giao dịch rất lớn, gây ra sự chênh lệch.

**Thống kê mô tả của BID:**

      date                 open            high            low

 Min.   :2014-01-24   Min.   :12.60   Min.   :12.70   Min.   :12.50

 1st Qu.:2016-07-26   1st Qu.:18.46   1st Qu.:18.70   1st Qu.:18.20

 Median :2019-01-10   Median :32.70   Median :33.25   Median :32.33

 Mean   :2018-11-26   Mean   :30.59   Mean   :31.06   Mean   :30.17

 3rd Qu.:2021-03-17   3rd Qu.:40.59   3rd Qu.:41.10   3rd Qu.:40.04

 Max.   :2023-09-08   Max.   :54.90   Max.   :55.80   Max.   :53.80

     close.Index           close.X[[i]]        nmVolume

 Min.   :2014-01-24    Min.   :12.60000     Min.   :   51430

 1st Qu.:2016-07-26    1st Qu.:18.40000     1st Qu.:  855175

 Median :2019-01-10    Median :32.70000     Median : 1510920

 Mean   :2018-11-26    Mean   :30.60494     Mean   : 1855853

 3rd Qu.:2021-03-17    3rd Qu.:40.50000     3rd Qu.: 2423380

 Max.   :2023-09-08    Max.   :55.50000     Max.   :15701040

     change

 Min.   :-3.40000

 1st Qu.:-0.30000

 Median : 0.00000

 Mean   : 0.01545

 3rd Qu.: 0.35000

 Max.   : 3.05000

- Biến động giá: Giá cổ phiếu BID có sự biến động từ mức thấp nhất là khoảng 12.50 - 12.70 đến mức cao nhất khoảng 53.80 - 55.80.

- Phân phối giá: Giá trị trung vị (Median) và các phân vị (Qu.) cho thấy sự phân bố giá có  mức độ biến động rõ rệt, với khoảng cách lớn giữa các giá trị phân vị Q1 và Q3. Điều này cho thấy giá cổ phiếu BID có sự tăng trưởng đáng kể qua các năm.

- Biến động khối lượng giao dịch: Khối lượng giao dịch của BID cũng có sự biến động lớn, từ mức thấp nhất là 51,430 đến mức cao nhất là 15,701,040. Điều này cho thấy có những giai đoạn BID có khối lượng giao dịch rất cao, có thể do các sự kiện quan trọng hoặc tin tức ảnh hưởng đến cổ phiếu.

- Phân phối khối lượng giao dịch: Giá trị trung vị (1,510,920) và giá trị trung bình (1,855,853) cho thấy khối lượng giao dịch trung bình và phần lớn các ngày có khối lượng giao dịch lớn.

**Thống kê mô tả của ACB:**

      date                 open            high            low

 Min.   :2013-01-02   Min.   :14.80   Min.   :15.10   Min.   :14.20

 1st Qu.:2015-10-08   1st Qu.:18.20   1st Qu.:18.40   1st Qu.:17.90

 Median :2018-07-03   Median :23.30   Median :23.60   Median :23.00

 Mean   :2018-05-20   Mean   :24.55   Mean   :24.88   Mean   :24.25

 3rd Qu.:2020-12-10   3rd Qu.:29.80   3rd Qu.:30.20   3rd Qu.:29.40

 Max.   :2023-09-08   Max.   :51.20   Max.   :51.70   Max.   :50.60

     close.Index           close.X[[i]]        nmVolume

 Min.   :2013-01-02    Min.   :15.00000     Min.   :    3990

 1st Qu.:2015-10-08    1st Qu.:18.20000     1st Qu.:  259986

 Median :2018-07-03    Median :23.20000     Median : 1567378

 Mean   :2018-05-20    Mean   :24.56399     Mean   : 2885698

 3rd Qu.:2020-12-10    3rd Qu.:29.80000     3rd Qu.: 4115341

 Max.   :2023-09-08    Max.   :51.70000     Max.   :35156000

     change

 Min.   :-3.80000

 1st Qu.:-0.15000

 Median : 0.00000

 Mean   : 0.01946

 3rd Qu.: 0.20000

 Max.   : 3.10000

- Biến động giá: Giá cổ phiếu ACB có sự biến động từ mức thấp nhất là khoảng 14.20 - 15.10 đến mức cao nhất khoảng 50.60 - 51.70. Điều này cho thấy cổ phiếu ACB đã trải qua một giai đoạn tăng trưởng đáng kể.

- Phân phối giá: Phân phối giá cổ phiếu có sự phân tán rõ rệt giữa các quý phân vị, với sự tăng trưởng giá đáng kể từ Q1 đến Q3 và giá trị trung bình (Mean) cao hơn giá trị trung vị (Median), cho thấy xu hướng tăng giá tổng thể.

- Biến động khối lượng giao dịch: Khối lượng giao dịch biến động lớn, từ mức thấp nhất là 3,990 đến mức cao nhất là 35,156,000. Điều này cho thấy có những giai đoạn mà cổ phiếu ACB được giao dịch rất sôi động.

- Phân phối khối lượng giao dịch: Giá trị trung vị (1,567,378) và giá trị trung bình (2,885,698) cho thấy có một số ngày giao dịch với khối lượng rất cao, gây ra sự chênh lệch lớn trong phân phối.

**Tính phương sai và độ lệch chuẩn**

**Phương sai theo từng ngân hàng:**

**VCB**

                 open                  high                   low

"         616.709165" "         630.574751" "         601.586165"

                close              nmVolume                change

"         617.594033" "664046943891.170898" "           1.380969"

**BID**

                   open                    high                     low

"          126.8223156" "          130.6174920" "          123.1574477"

                  close                nmVolume                  change

"          126.5476568" "2123953974190.2370605" "            0.5539685"

**ACB**

                   open                    high                     low

"            55.988426" "            58.187146" "            53.719789"

                  close                nmVolume                  change

"            55.912143" "13812384086895.287109" "             0.246362"

Phương sai là phép đo mức chênh lệch giữa các số liệu trong một tập dữ liệu trong thống kê. Phương sai đo lường sự khác biệt đến điểm trung vị hoặc trung bình. Trong bối cảnh này, phương sai đóng vai trò phân tích mức độ biến động của dữ liệu cổ phiếu của từng ngân hàng.

VCB: Có sự biến động mạnh nhất về giá cổ phiếu và khối lượng giao dịch. Điều này có thể làm cho việc dự đoán giá khó khăn hơn, nhưng cũng có thể mang lại cơ hội lợi nhuận lớn hơn.

BID: Phương sai thấp hơn so với VCB, cho thấy sự ổn định hơn trong giá cổ phiếu BID, nhưng vẫn có sự biến động đáng kể về khối lượng giao dịch, phương sai cao cho thấy biến động mạnh trong khối lượng giao dịch.

ACB: Giá cổ phiếu có phương sai thấp nhất trong 3 ngân hàng cho thấy ổn định nhất về giá, nhưng có sự biến động rất lớn trong khối lượng giao dịch.

**Độ lệch chuẩn theo từng ngân hàng:**

**VCB**

           open            high             low           close

"    24.833630" "    25.111247" "    24.527254" "    24.851439"

       nmVolume          change

"814890.755802" "     1.175147"

**BID**

             open              high               low             close

"     11.2615414" "     11.4288010" "     11.0976325" "     11.2493403"

         nmVolume            change

"1457379.1456550" "      0.7442906"

**ACB**

             open              high               low             close

"      7.4825414" "      7.6280499" "      7.3293785" "      7.4774423"

         nmVolume            change

"3716501.5924785" "      0.4963487"

Độ lệch chuẩn là một chỉ số thống kê quan trọng giúp đo lường mức độ biến động của dữ liệu xung quanh giá trị trung bình. Trong bối cảnh phân tích dữ liệu chứng khoán của các ngân hàng, độ lệch chuẩn có thể cung cấp cái nhìn sâu sắc về mức độ biến động giá của cổ phiếu, khối lượng giao dịch, và sự thay đổi giá.

Xét về độ lệch chuẩn, nhìn chung các biến của 3 ngân hàng có thể thấy độ biến động của giá cổ phiếu theo thứ tự từ cao đến thấp là: VCB, BID, ACB.

- Cổ phiếu VCB có mức biến động khá cao, đặc biệt là ở các mức giá và khối lượng giao dịch. Tất cả các giá trị mở cửa, cao nhất, thấp nhất và đóng cửa đều có độ lệch chuẩn cao, cho thấy giá cổ phiếu VCB biến động mạnh. Độ lệch chuẩn của khối lượng giao dịch cũng khá cao, chỉ ra sự biến động lớn trong khối lượng giao dịch hàng ngày.

- BID có biến động giá cổ phiếu thấp hơn so với VCB. So với VCB, BID có độ lệch chuẩn của giá mở cửa, cao nhất, thấp nhất và đóng cửa thấp hơn, chỉ ra sự ổn định hơn trong giá cổ phiếu. Mô hình dự đoán có thể ít bị ảnh hưởng bởi các biến số ngẫu nhiên và có thể tập trung vào các yếu tố ổn định như xu hướng lịch sử. Khối lượng giao dịch có độ lệch chuẩn cao, tương tự VCB, cần được xem xét kỹ lưỡng trong mô hình dự đoán.

- ACB có mức biến động giá cổ phiếu thấp nhất, cho thấy tính ổn định cao. Tuy nhiên, khối lượng giao dịch có độ lệch chuẩn cao nhất trong ba ngân hàng, cho thấy sự quan tâm lớn từ nhà đầu tư.

Vẽ biểu đồ đường giá đóng cửa của 3 ngân hàng dựa trên giá đóng cửa:

A graph of a graph of a price

Description automatically generated with medium confidence

*Hình 2.10. Biểu đồ đường của giá đóng cửa của 3 ngân hàng VCB, BID, ACB*

Biểu đồ giá đóng cửa của 3 ngân hàng ở hình 2.1 cho thấy sự tăng trưởng theo thời gian của ngân hàng VCB tốt nhất, tiếp đó là ngân hàng BID cũng có sự phát triển tăng tiến theo xu hướng dài hạn, còn cổ phiếu ngân hàng ACB có nhiều biến động và giá đóng cửa cũng thấp nhất trong 3 ngân hàng. Nhóm sẽ tiến hành phân tích khám phá các biến của 3 ngân hàng chi tiết hơn ở các phần sau.

**b) Phân tích phân phối của các biến**

Vẽ biểu đồ boxplot kết hợp theo các biến open, low, high, close của 3 ngân hàng:

A graph of different colored squares

Description automatically generated

*Hình 2.11. Biểu đồ boxplot các biến giá low, high, open, close của 3 ngân hàng VCB, BID, ACB*

So sánh độ phân phối giá của 3 ngân hàng:

- Nhìn chung, VCB có giá cổ phiếu cao nhất trong ba ngân hàng.

- Độ biến động về giá: VCB có độ biến động giá cao nhất và phạm vi giá rộng nhất, thể hiện qua hộp mở rộng lớn (khoảng cách từ Q1 tới Q3 dài), cho thấy sự biến động lớn hơn trong phạm vi giá của ngân hàng này. BID có độ biến động giá thấp hơn, và cuối cùng là ACB với mức độ biến động giá thấp nhất. Kết hợp với thông tin từ biểu đồ đường giá đóng cửa hình 2.2, có thể thấy rằng nguyên nhân dẫn đến mức độ biến động giá cao nhất của VCB là do mức độ tăng trưởng tốt nhất theo thời gian.

- Hình dạng phân phối: hình dạng phân phối của VCB và ACB đều xấp xỉ đối xứng. Còn hình dạng phân phối của BID có sự lệch phải nhẹ.

- Giá trị ngoại lai: ACB là ngân hàng duy nhất có xuất hiện giá trị ngoại lai và tất cả các biến giá low, high, open, close đều có giá trị ngoại lai. Nhóm sẽ kiểm tra và phân tích chi tiết các giá trị này ở phần sau.

Vẽ biểu đồ boxplot của biến khối lượng giao dịch cho 3 ngân hàng:

A graph of different colored lines

Description automatically generated

*Hình 2.12. Biểu đồ boxplot biến Volume của 3 ngân hàng VCB, BID, ACB*

- Độ biến động khối lượng giao dịch: ACB có độ biến động về khối lượng lớn nhất thể hiện qua độ mở rộng của hộp (khoảng cách từ Q1 tới Q3 dài nhất). BID có độ biến động khối lượng giao dịch thấp hơn và độ biến động khối lượng thấp nhất là VCB.

- Hình dạng phân phối: cả 3 ngân hàng đều có hình dạng phân phối lệch phải nhẹ, điều này cho thấy có những phiên giao dịch có khối lượng lớn hơn so với trung bình, đặc biệt là ở phần cao nhất của phân phối. Điều này có thể phản ánh sự biến động mạnh mẽ hoặc các sự kiện đặc biệt trong thị trường giao dịch cổ phiếu của các ngân hàng.

- Giá trị ngoại lai: biến khối lượng giao dịch của cả ba ngân hàng đều có các giá trị ngoại lai. Nhóm sẽ tiến hành phân tích các giá trị ngoại lại này kĩ hơn ở phần sau.

### 2.3.2. Tìm và xử lý dữ liệu rỗng

Kiểm tra dữ liệu rỗng của 3 ngân hàng

Số dữ liệu rỗng của VCB:

    date     open     high      low          nmVolume   change

       0        0        0        0        0        0        0

Số dữ liệu rỗng của BID:

    date     open     high      low          nmVolume   change

       0        0        0        0        0        0        0

Số dữ liệu rỗng của ACB:

    date     open     high      low          nmVolume   change

       0        0        0        0        0        0        0

Với kết quả kiểm tra trên, dữ liệu của tất cả các biến của cả ba ngân hàng không có giá trị rỗng.

### 2.3.3. Tìm và xử lý ngoại lai

Dựa vào biểu đồ boxplot ở phần 2.1.2 có thể xác định một số biến có giá trị ngoại lai là:

- VCB: nmVolume

- BID: nmVolume

- ACB: high, low, open, close, nmVolume

**a) Tìm ngoại lai các biến Volume của VCB**

Tiến hành kiểm tra bằng biểu đồ phân phối Histogram

A graph of a number of numbers

Description automatically generated

*Hình 2.13. Biểu đồ phân phối histogram biến Volume của VCB*

Biểu đồ phân phối trên có dạng lệch phải, nhóm sử dụng phương pháp IQR (Interquartile Range) để kiểm tra ngoại lai của biến Volume. IQR là khoảng cách giữa phần tư thứ nhất (Q1) và phần tư thứ ba (Q3) của dữ liệu, IQR chỉ dựa trên các phần tư của dữ liệu và không bị ảnh hưởng bởi các giá trị cực đoan, nó cung cấp một thước đo tốt về sự phân tán mà không bị nhiễu bởi ngoại lai.

Tìm các ngoại lai bằng IQR. Với k = 1.5 sau đó tiến hành kiểm tra thông tin của các giá trị ngoại lai.

Kết quả cho thấy có 133 ngoại lai của biến Volume:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **nmVolume**  <dbl> | **date**  <date> |  |  | **nmVolume**  <dbl> | **date**  <date> |
| 243 | 2834510 | 2013-12-20 |  | 1317 | 4909820 | 2018-04-19 |
| 263 | 2886110 | 2014-01-20 |  | 1318 | 2996320 | 2018-04-20 |
| 300 | 3372760 | 2014-03-21 |  | 1319 | 2876760 | 2018-04-23 |
| 489 | 4879410 | 2014-12-19 |  | 1320 | 4451970 | 2018-04-24 |
| 503 | 3059730 | 2015-01-12 |  | 1321 | 4669710 | 2018-04-26 |
| 546 | 4058430 | 2015-03-20 |  | 1322 | 2940100 | 2018-04-27 |
| 663 | 2703220 | 2015-09-08 |  | 1323 | 2862030 | 2018-05-02 |
| 671 | 3095740 | 2015-09-18 |  | 1324 | 3451890 | 2018-05-03 |
| 736 | 3155650 | 2015-12-18 |  | 1325 | 3234250 | 2018-05-04 |
| 795 | 3546350 | 2016-03-18 |  | 1326 | 3511210 | 2018-05-07 |
| 911 | 3427730 | 2016-09-01 |  | 1327 | 2947160 | 2018-05-08 |
| 912 | 4191860 | 2016-09-05 |  | 1328 | 2787480 | 2018-05-09 |
| 916 | 3353710 | 2016-09-09 |  | 1329 | 4433980 | 2018-05-10 |
| 921 | 7265480 | 2016-09-16 |  | 1330 | 3954360 | 2018-05-11 |
| 1045 | 3101080 | 2017-03-17 |  | 1337 | 3235250 | 2018-05-22 |
| 1088 | 3432310 | 2017-05-22 |  | 1340 | 2718610 | 2018-05-25 |
| 1099 | 4750190 | 2017-06-06 |  | 1341 | 4579370 | 2018-05-28 |
| 1105 | 2818810 | 2017-06-14 |  | 1343 | 4432950 | 2018-05-30 |
| 1106 | 2996650 | 2017-06-15 |  | 1344 | 3736400 | 2018-05-31 |
| 1107 | 2708280 | 2017-06-16 |  | 1345 | 5930650 | 2018-06-01 |
| 1171 | 2799180 | 2017-09-15 |  | 1346 | 3837380 | 2018-06-04 |
| 1172 | 2916990 | 2017-09-18 |  | 1347 | 3532580 | 2018-06-05 |
| 1184 | 4836990 | 2017-10-04 |  | 1350 | 2693680 | 2018-06-08 |
| 1185 | 5672370 | 2017-10-05 |  | 1352 | 3024500 | 2018-06-12 |
| 1189 | 2835470 | 2017-10-11 |  | 1354 | 2814340 | 2018-06-14 |
| 1193 | 3062770 | 2017-10-17 |  | 1357 | 5149160 | 2018-06-19 |
| 1194 | 3053330 | 2017-10-18 |  | 1358 | 2767440 | 2018-06-20 |
| 1197 | 3568930 | 2017-10-23 |  | 1366 | 2846420 | 2018-07-02 |
| 1202 | 3808170 | 2017-10-30 |  | 1367 | 3355250 | 2018-07-03 |
| 1209 | 2695690 | 2017-11-08 |  | 1369 | 2865870 | 2018-07-05 |
| 1213 | 3008230 | 2017-11-14 |  | 1370 | 3443450 | 2018-07-06 |
| 1218 | 3799100 | 2017-11-21 |  | 1373 | 3312190 | 2018-07-11 |
| 1219 | 3116060 | 2017-11-22 |  | 1387 | 3438180 | 2018-07-31 |
| 1230 | 2737090 | 2017-12-07 |  | 1392 | 5488590 | 2018-08-07 |
| 1241 | 4459390 | 2017-12-22 |  | 1395 | 3849630 | 2018-08-10 |
| 1246 | 2753570 | 2017-12-29 |  | 1398 | 3238040 | 2018-08-15 |
| 1247 | 2801720 | 2018-01-02 |  | 1399 | 3426320 | 2018-08-16 |
| 1252 | 2853160 | 2018-01-09 |  | 1400 | 2711770 | 2018-08-17 |
| 1253 | 2842830 | 2018-01-10 |  | 1424 | 4889170 | 2018-09-21 |
| 1259 | 2845720 | 2018-01-18 |  | 1438 | 3432840 | 2018-10-11 |
| 1260 | 2998270 | 2018-01-19 |  | 1747 | 2794080 | 2019-09-20 |
| 1262 | 4788340 | 2018-01-25 |  | 1835 | 3541560 | 2020-01-30 |
| 1265 | 2818520 | 2018-01-30 |  | 1871 | 2763360 | 2020-03-20 |
| 1270 | 6724710 | 2018-02-06 |  | 1902 | 3865590 | 2020-05-08 |
| 1278 | 4825770 | 2018-02-23 |  | 2097 | 2902200 | 2021-02-08 |
| 1279 | 3453290 | 2018-02-26 |  | 2098 | 2978900 | 2021-02-09 |
| 1281 | 3475320 | 2018-02-28 |  | 2129 | 2697600 | 2021-04-01 |
| 1282 | 3615670 | 2018-03-01 |  | 2131 | 3592400 | 2021-04-05 |
| 1284 | 2914810 | 2018-03-05 |  | 2142 | 6274200 | 2021-04-20 |
| 1285 | 3103920 | 2018-03-06 |  | 2143 | 2981300 | 2021-04-22 |
| 1288 | 3236520 | 2018-03-09 |  | 2145 | 3595400 | 2021-04-26 |
| 1289 | 2899140 | 2018-03-12 |  | 2164 | 3816000 | 2021-05-25 |
| 1291 | 4053460 | 2018-03-14 |  | 2169 | 4533500 | 2021-06-01 |
| 1293 | 5327050 | 2018-03-16 |  | 2170 | 4274100 | 2021-06-02 |
| 1294 | 2803610 | 2018-03-19 |  | 2171 | 2802000 | 2021-06-03 |
| 1296 | 2952180 | 2018-03-21 |  | 2173 | 2969900 | 2021-06-07 |
| 1297 | 3871380 | 2018-03-22 |  | 2180 | 2898400 | 2021-06-16 |
| 1298 | 4563730 | 2018-03-23 |  | 2182 | 3234900 | 2021-06-18 |
| 1304 | 3055600 | 2018-04-02 |  | 2190 | 2768000 | 2021-06-30 |
| 1305 | 3500060 | 2018-04-03 |  | 2197 | 2936900 | 2021-07-09 |
| 1309 | 2693260 | 2018-04-09 |  | 2311 | 3475700 | 2021-12-17 |
| 1310 | 5026810 | 2018-04-10 |  | 2334 | 2794300 | 2022-01-20 |
| 1311 | 3911130 | 2018-04-11 |  | 2336 | 2728300 | 2022-01-24 |
| 1312 | 3262970 | 2018-04-12 |  | 2346 | 3629100 | 2022-02-14 |
| 1313 | 5148440 | 2018-04-13 |  | 2362 | 2936300 | 2022-03-08 |
| 1314 | 3992890 | 2018-04-16 |  | 2395 | 2776600 | 2022-04-22 |
|  |  |  |  | 2562 | 3158100 | 2022-12-16 |

Trực quan lại các điểm ngoại lai bằng biểu đồ phân tán:

A graph showing a number of red and blue dots

Description automatically generatedA graph with red and blue dots

Description automatically generated

*Hình 2.14. Biểu đồ phân tán các điểm ngoại lai của biến Volume của VCB*

Phân tích ngoại lai của biến Volume

- Phân bố trong khoảng giá: các giá trị ngoại lai của khối lượng giao dịch xuất hiện trong nhiều khoảng giá khác nhau.

- Phân bố ngoại lai: Các giá trị ngoại lai tập trung nhiều nhất trong khoảng thời gian từ cuối năm 2017 tới nửa đầu năm 2018.

- Liên hệ với xu hướng giá: Kết hợp với biểu đồ đường giá đóng cửa trong Hình 2. , ta thấy rằng từ đầu năm 2017, giá đóng cửa của VCB đang trong giai đoạn giảm (downtrend). Tuy nhiên, đến cuối năm 2017, giá bắt đầu dao động và chuyển sang xu hướng tăng (uptrend). Nguyên nhân có thể đến từ khả năng xuất hiện những sự kiện kinh tế và thông báo quan trọng từ VCB đã xảy ra trong giai đoạn này. Những thông tin này có thể đã tác động mạnh mẽ đến hành vi của nhà đầu tư, khiến họ nắm lấy cơ hội để mua/bán cổ phiếu. Kết quả là khối lượng giao dịch có nhiều biến động lớn, dẫn đến sự xuất hiện của các giá trị ngoại lai.

Khối lượng giao dịch của VCB trong giai đoạn từ cuối năm 2017 tới nửa đầu năm 2018 cho thấy sự tập trung của các giá trị ngoại lai. Điều này có thể phản ánh những thay đổi đáng kể trong thị trường và hành vi của nhà đầu tư do tác động của các sự kiện kinh tế hoặc thông báo quan trọng từ VCB, đặc biệt khi giá cổ phiếu chuyển từ xu hướng giảm sang tăng.

**b) Tìm ngoại lai các biến close, volume của BID**

Tiến hành kiểm tra bằng biểu đồ phân phối Histogram

A graph of a number of numbers

Description automatically generated with medium confidence

*Hình 2.15. Biểu đồ phân phối của biến Volume của BID*

Biểu đồ phân phối biến Volume của BID ở hình 2.4 cho thấy phân phối có dạng lệch phải, nhóm cũng tiến hành kiểm tra giá trị ngoại lai bằng phương pháp IQR sau đó kiểm tra thông tin của các giá trị đó.

Tìm các ngoại lai bằng IQR. Với k = 1.5

Kết quả cho thấy có 106 ngoại lai như bên dưới:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Date**  <date> | **nmVolume**  <dbl> |  |  | **date**  <date> | **nmVolume**  <dbl> |
| 1 | 2014-01-24 | 8417060 |  | 879 | 2017-08-09 | 9682480 |
| 32 | 2014-03-19 | 5624790 |  | 996 | 2018-01-25 | 7986250 |
| 245 | 2015-01-22 | 4915010 |  | 997 | 2018-01-26 | 5677170 |
| 246 | 2015-01-23 | 5685560 |  | 1080 | 2018-06-04 | 4878280 |
| 248 | 2015-01-27 | 6307270 |  | 1091 | 2018-06-19 | 5460120 |
| 249 | 2015-01-28 | 7664440 |  | 1101 | 2018-07-03 | 4944960 |
| 263 | 2015-02-25 | 6844330 |  | 1104 | 2018-07-06 | 5061910 |
| 297 | 2015-04-14 | 5097900 |  | 1105 | 2018-07-09 | 5423120 |
| 299 | 2015-04-16 | 4920810 |  | 1107 | 2018-07-11 | 5441810 |
| 355 | 2015-07-09 | 5260640 |  | 1121 | 2018-07-31 | 5307160 |
| 402 | 2015-09-15 | 9105340 |  | 1132 | 2018-08-15 | 4946070 |
| 405 | 2015-09-18 | 6916750 |  | 1133 | 2018-08-16 | 5017400 |
| 468 | 2015-12-16 | 6453910 |  | 1165 | 2018-10-02 | 5199330 |
| 494 | 2016-01-22 | 5376560 |  | 1172 | 2018-10-11 | 6573500 |
| 512 | 2016-02-24 | 5780650 |  | 1748 | 2020-10-14 | 6821050 |
| 678 | 2016-10-19 | 5957570 |  | 1853 | 2021-03-18 | 5666500 |
| 731 | 2017-01-03 | 4954000 |  | 1864 | 2021-04-02 | 5794100 |
| 734 | 2017-01-06 | 6338280 |  | 1865 | 2021-04-05 | 5969400 |
| 735 | 2017-01-09 | 6160700 |  | 1870 | 2021-04-12 | 5149300 |
| 737 | 2017-01-11 | 7292910 |  | 1888 | 2021-05-11 | 5349300 |
| 741 | 2017-01-17 | 6538420 |  | 1896 | 2021-05-21 | 13027500 |
| 742 | 2017-01-18 | 5523490 |  | 1897 | 2021-05-24 | 6994100 |
| 781 | 2017-03-21 | 10006770 |  | 1898 | 2021-05-25 | 6329600 |
| 782 | 2017-03-22 | 9321930 |  | 1899 | 2021-05-26 | 7369800 |
| 784 | 2017-03-24 | 8242990 |  | 1900 | 2021-05-27 | 5862300 |
| 785 | 2017-03-27 | 7212840 |  | 1901 | 2021-05-28 | 7857000 |
| 788 | 2017-03-30 | 5467270 |  | 1902 | 2021-05-31 | 5597900 |
| 814 | 2017-05-10 | 6204060 |  | 1904 | 2021-06-02 | 6043600 |
| 821 | 2017-05-19 | 5387170 |  | 1905 | 2021-06-03 | 6715400 |
| 822 | 2017-05-22 | 15701040 |  | 1906 | 2021-06-04 | 5939600 |
| 839 | 2017-06-14 | 7438470 |  | 1907 | 2021-06-07 | 5275100 |
| 840 | 2017-06-15 | 5827220 |  | 1908 | 2021-06-08 | 5685400 |
| 842 | 2017-06-19 | 4812540 |  | 1911 | 2021-06-11 | 6390300 |
| 843 | 2017-06-20 | 6648680 |  | 1922 | 2021-06-28 | 6685100 |
| 844 | 2017-06-21 | 7245070 |  | 1928 | 2021-07-06 | 6896000 |
| 850 | 2017-06-29 | 5414500 |  | 1932 | 2021-07-12 | 6021800 |
| 855 | 2017-07-06 | 5082620 |  | 2011 | 2021-11-03 | 5601700 |
| 856 | 2017-07-07 | 4883780 |  | 2024 | 2021-11-22 | 4942900 |
| 868 | 2017-07-25 | 4854620 |  | 2053 | 2021-12-31 | 4857000 |
| 869 | 2017-07-26 | 4938950 |  | 2056 | 2022-01-06 | 5998800 |
| 871 | 2017-07-28 | 5408060 |  | 2060 | 2022-01-12 | 8028300 |
| 872 | 2017-07-31 | 6728350 |  | 2061 | 2022-01-13 | 9664800 |
| 874 | 2017-08-02 | 6829100 |  | 2063 | 2022-01-17 | 7741100 |
| 885 | 2017-08-17 | 5639850 |  | 2064 | 2022-01-18 | 4849200 |
| 895 | 2017-08-31 | 5421350 |  | 2066 | 2022-01-20 | 5637900 |
| 927 | 2017-10-17 | 5851150 |  | 2078 | 2022-02-14 | 7129000 |
| 936 | 2017-10-30 | 5516030 |  | 2100 | 2022-03-17 | 5819900 |
| 995 | 2018-01-22 | 5040120 |  | 2175 | 2022-07-05 | 4933500 |

Trực quan lại các điểm ngoại lai của biến Volume bằng biểu đồ scatter.

A chart with red and blue dots

Description automatically generatedA graph with red and blue dots

Description automatically generated

*Hình 2.16. Biểu đồ phân tán các điểm ngoại lai của biến Volume của BID*

Phân tích các giá trị ngoại lai của biến Volume

- Phân bố trong khoảng giá: có hầu hết ở các khoảng giá, tập trung nhiều nhất ở khoảng giá 15 - 12 và 40 - 50.

- Phân bố trong khoảng thời gian: các giá trị ngoại lai tập trung nhiều ở khoảng thời gian 2017 - 2018 và 2021 - 2022. Một số nguyên nhân có thể đến từ tin tức, sự kiện kinh tế, thay đổi chính sách từ ngân hàng BIDV,... gây ảnh hưởng đến quyết định của các nhà đầu tư.

**c) Tìm và xử các ngoại lại trong biến high, low, open, volume của ACB**

Vẽ biểu đồ phân phối histogram đối với các biến nghi ngờ có ngoại lai

A graph of a graph

Description automatically generatedA graph of a graph

Description automatically generatedA graph of a graph

Description automatically generated

A graph of a graph

Description automatically generatedA graph of a number of numbers

Description automatically generated

*Hình 2.17. Biểu đồ phân phối các biến low, high, open, close của ACB*

Biểu đồ phân phối của các biến low, high, open, close đều có hình dạng lệch phải, điều này cho thấy dữ liệu có phân bố không đồng đều với một số giá trị rất cao nằm ở phía bên phải của biểu đồ. Điều này có thể gợi ý về sự hiện diện của các giá trị ngoại lai. Biểu đồ phân phối của biến Volume cũng có hình dạng lệch phải và lệch hơn nhiều so với các biến giá, điều này cho thấy dữ liệu biến khối lượng giao dịch có sự chênh lệch lớn giữa các giá trị trong tập dữ liệu.

Tìm các ngoại lai bằng IQR. Với k = 1.5. Sau đó, kiểm tra các thông tin của các ngoại lai.

**Outliers in Low (including date):**

|  |  |  |
| --- | --- | --- |
|  | **date**  <date> | **low**  <dbl> |
| 1291 | 2018-03-12 | 47.0 |
| 1292 | 2018-03-13 | 48.0 |
| 1293 | 2018-03-14 | 49.0 |
| 1294 | 2018-03-15 | 48.6 |
| 1295 | 2018-03-16 | 49.5 |
| 1296 | 2018-03-19 | 50.6 |
| 1297 | 2018-03-20 | 47.0 |
| 1298 | 2018-03-21 | 47.5 |
| 1299 | 2018-03-22 | 46.8 |
| 1302 | 2018-03-27 | 47.2 |
| 1303 | 2018-03-28 | 46.7 |
| 1306 | 2018-04-02 | 46.8 |
| 1307 | 2018-04-03 | 48.5 |
| 1308 | 2018-04-04 | 48.9 |
| 1309 | 2018-04-05 | 49.0 |
| 1310 | 2018-04-06 | 49.8 |
| 1311 | 2018-04-09 | 50.4 |
| 1312 | 2018-04-10 | 50.1 |
| 1313 | 2018-04-11 | 48.8 |
| 1314 | 2018-04-12 | 48.1 |
| 1315 | 2018-04-13 | 48.4 |
| 1316 | 2018-04-16 | 47.7 |
| 1317 | 2018-04-17 | 48.3 |
| 1318 | 2018-04-18 | 48.7 |
| 1319 | 2018-04-19 | 47.2 |
| 1320 | 2018-04-20 | 47.4 |

**Outliers in High (including date):**

|  |  |  |
| --- | --- | --- |
|  | **date**  <date> | **high**  <dbl> |
| 1291 | 2018-03-12 | 48.6 |
| 1292 | 2018-03-13 | 49.2 |
| 1293 | 2018-03-14 | 51.0 |
| 1294 | 2018-03-15 | 49.6 |
| 1295 | 2018-03-16 | 50.7 |
| 1296 | 2018-03-19 | 51.7 |
| 1297 | 2018-03-20 | 48.5 |
| 1298 | 2018-03-21 | 48.8 |
| 1299 | 2018-03-22 | 48.1 |
| 1302 | 2018-03-27 | 48.3 |
| 1306 | 2018-04-02 | 49.2 |
| 1307 | 2018-04-03 | 49.5 |
| 1308 | 2018-04-04 | 49.8 |
| 1309 | 2018-04-05 | 49.8 |
| 1310 | 2018-04-06 | 51.4 |
| 1311 | 2018-04-09 | 51.5 |
| 1312 | 2018-04-10 | 51.6 |
| 1313 | 2018-04-11 | 50.9 |
| 1314 | 2018-04-12 | 49.9 |
| 1315 | 2018-04-13 | 50.4 |
| 1316 | 2018-04-16 | 49.1 |
| 1317 | 2018-04-17 | 49.5 |
| 1318 | 2018-04-18 | 50.3 |
| 1319 | 2018-04-19 | 49.0 |
| 1320 | 2018-04-20 | 48.3 |
| 1321 | 2018-04-23 | 48.8 |
| 1335 | 2018-05-16 | 48.9 |

**Outliers in Open (including date):**

|  |  |  |
| --- | --- | --- |
| **date**  <date> | **open**  <dbl> |  |
| 1292 | 2018-03-13 | 48.7 |
| 1293 | 2018-03-14 | 50.5 |
| 1294 | 2018-03-15 | 49.3 |
| 1295 | 2018-03-16 | 49.8 |
| 1296 | 2018-03-19 | 51.0 |
| 1297 | 2018-03-20 | 47.5 |
| 1298 | 2018-03-21 | 48.4 |
| 1299 | 2018-03-22 | 47.9 |
| 1302 | 2018-03-27 | 47.6 |
| 1303 | 2018-03-28 | 47.4 |
| 1307 | 2018-04-03 | 49.2 |
| 1308 | 2018-04-04 | 49.0 |
| 1309 | 2018-04-05 | 49.2 |
| 1310 | 2018-04-06 | 49.9 |
| 1311 | 2018-04-09 | 51.0 |
| 1312 | 2018-04-10 | 51.2 |
| 1313 | 2018-04-11 | 50.5 |
| 1314 | 2018-04-12 | 48.1 |
| 1315 | 2018-04-13 | 50.0 |
| 1316 | 2018-04-16 | 48.4 |
| 1317 | 2018-04-17 | 48.6 |
| 1318 | 2018-04-18 | 49.5 |
| 1319 | 2018-04-19 | 48.0 |
| 1320 | 2018-04-20 | 48.0 |
| 1321 | 2018-04-23 | 48.4 |

Trực quan lại các điểm ngoại lai của biến Close bằng biểu đồ phân tán:

A graph with blue lines and white text

Description automatically generated

*Hình 2.18. Biểu đồ phân tán các giá trị ngoại lai của giá đóng cửa của ACB*

***Phân tích các giá trị ngoại lai của các biến giá low, high, open, close***

Có thể thấy rằng các biến giá low, high, open, và close đều có cùng số lượng khoảng 25 - 27 giá trị ngoại lai và cùng nằm trong khoảng thời gian từ 12/03/2018 đến 23/04/2018. Điều này dễ hiểu được vì các biến giá thường có sự tương quan với nhau. Thời điểm này nằm trong giai đoạn giá cổ phiếu tăng (uptrend), và những giá trị ngoại lai xuất hiện ở đỉnh điểm của giai đoạn đó. Nguyên nhân xuất hiện các giá trị ngoại lai có thể là do vào thời điểm này đã có một sự kiện kinh tế hoặc một thông báo quan trọng từ ngân hàng, ảnh hưởng mạnh mẽ đến giá cổ phiếu.

**Outliers in Volume (including date):**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **date**  <date> | **nmVolume**  <dbl> |  |  | **date**  <date> | **nmVolume**  <dbl> |
| 1047 | 2017-03-21 | 10618307 |  | 2162 | 2021-05-26 | 9976000 |
| 1281 | 2018-02-26 | 10048941 |  | 2165 | 2021-05-31 | 16030300 |
| 1846 | 2020-02-12 | 10516207 |  | 2166 | 2021-06-01 | 11159800 |
| 1854 | 2020-02-24 | 10321149 |  | 2167 | 2021-06-02 | 16009600 |
| 1864 | 2020-03-09 | 16776046 |  | 2168 | 2021-06-03 | 12265700 |
| 1867 | 2020-03-12 | 11082125 |  | 2170 | 2021-06-07 | 18505000 |
| 1874 | 2020-03-23 | 10024177 |  | 2171 | 2021-06-08 | 14322800 |
| 1920 | 2020-06-01 | 13859688 |  | 2172 | 2021-06-09 | 12565100 |
| 1921 | 2020-06-02 | 10935462 |  | 2173 | 2021-06-10 | 14000300 |
| 1972 | 2020-08-11 | 15333430 |  | 2175 | 2021-06-14 | 10946100 |
| 1979 | 2020-08-20 | 19475136 |  | 2176 | 2021-06-15 | 18178200 |
| 1991 | 2020-09-08 | 12273484 |  | 2181 | 2021-06-22 | 14872000 |
| 1998 | 2020-09-17 | 12323866 |  | 2182 | 2021-06-23 | 10244600 |
| 2009 | 2020-10-02 | 23518516 |  | 2183 | 2021-06-24 | 11494200 |
| 2010 | 2020-10-05 | 12891936 |  | 2188 | 2021-07-01 | 10393200 |
| 2011 | 2020-10-06 | 10977492 |  | 2190 | 2021-07-05 | 16668600 |
| 2012 | 2020-10-07 | 17695173 |  | 2191 | 2021-07-06 | 10070000 |
| 2013 | 2020-10-08 | 13901830 |  | 2192 | 2021-07-07 | 10121600 |
| 2017 | 2020-10-14 | 20294474 |  | 2194 | 2021-07-09 | 11315400 |
| 2018 | 2020-10-15 | 16349453 |  | 2195 | 2021-07-12 | 18353700 |
| 2020 | 2020-10-19 | 14160428 |  | 2197 | 2021-07-14 | 10810500 |
| 2021 | 2020-10-20 | 11362976 |  | 2200 | 2021-07-19 | 10019200 |
| 2025 | 2020-10-26 | 11665717 |  | 2209 | 2021-07-30 | 35156000 |
| 2026 | 2020-10-27 | 11726997 |  | 2213 | 2021-08-05 | 13787200 |
| 2027 | 2020-10-28 | 14916975 |  | 2215 | 2021-08-09 | 19240200 |
| 2032 | 2020-11-04 | 12607096 |  | 2216 | 2021-08-10 | 10457700 |
| 2038 | 2020-11-12 | 18955350 |  | 2217 | 2021-08-11 | 16492300 |
| 2041 | 2020-11-17 | 16779929 |  | 2218 | 2021-08-12 | 12147100 |
| 2045 | 2020-11-23 | 10897510 |  | 2219 | 2021-08-13 | 15125200 |
| 2046 | 2020-11-24 | 10360554 |  | 2220 | 2021-08-16 | 12056700 |
| 2048 | 2020-11-26 | 10037160 |  | 2224 | 2021-08-20 | 20952900 |
| 2052 | 2020-12-09 | 24200930 |  | 2225 | 2021-08-23 | 15602700 |
| 2057 | 2020-12-16 | 10461450 |  | 2273 | 2021-10-29 | 13597100 |
| 2058 | 2020-12-17 | 13567290 |  | 2276 | 2021-11-03 | 13060100 |
| 2059 | 2020-12-18 | 11693630 |  | 2291 | 2021-11-24 | 14295800 |
| 2060 | 2020-12-21 | 22481640 |  | 2333 | 2022-01-24 | 13185100 |
| 2062 | 2020-12-23 | 11216860 |  | 2337 | 2022-01-28 | 10258100 |
| 2063 | 2020-12-24 | 15831250 |  | 2339 | 2022-02-08 | 32461200 |
| 2069 | 2021-01-04 | 10776300 |  | 2340 | 2022-02-09 | 10149700 |
| 2070 | 2021-01-05 | 18693900 |  | 2343 | 2022-02-14 | 13153200 |
| 2071 | 2021-01-06 | 15883900 |  | 2355 | 2022-03-02 | 13881600 |
| 2072 | 2021-01-07 | 11636100 |  | 2528 | 2022-11-04 | 12616400 |
| 2073 | 2021-01-08 | 15519300 |  | 2625 | 2023-03-29 | 15179800 |
| 2076 | 2021-01-13 | 9915300 |  | 2628 | 2023-04-03 | 9966000 |
| 2080 | 2021-01-19 | 13077400 |  | 2661 | 2023-05-23 | 21722400 |
| 2081 | 2021-01-20 | 12323000 |  | 2662 | 2023-05-24 | 11048500 |
| 2087 | 2021-01-28 | 15154800 |  | 2665 | 2023-05-29 | 16572400 |
| 2098 | 2021-02-19 | 31184200 |  | 2666 | 2023-05-30 | 22595600 |
| 2099 | 2021-02-22 | 12587300 |  | 2667 | 2023-05-31 | 29153500 |
| 2100 | 2021-02-23 | 13951000 |  | 2668 | 2023-06-01 | 12761200 |
| 2101 | 2021-02-24 | 18685200 |  | 2669 | 2023-06-02 | 15030000 |
| 2102 | 2021-02-25 | 15014200 |  | 2670 | 2023-06-05 | 13842800 |
| 2103 | 2021-02-26 | 13071400 |  | 2671 | 2023-06-06 | 9941200 |
| 2108 | 2021-03-05 | 10157000 |  | 2679 | 2023-06-16 | 14410500 |
| 2109 | 2021-03-08 | 16047000 |  | 2684 | 2023-06-23 | 11433200 |
| 2111 | 2021-03-10 | 12152700 |  | 2685 | 2023-06-26 | 11784600 |
| 2114 | 2021-03-15 | 18052500 |  | 2706 | 2023-07-25 | 13703900 |
| 2117 | 2021-03-18 | 10281900 |  | 2710 | 2023-07-31 | 19206000 |
| 2120 | 2021-03-23 | 10347000 |  | 2711 | 2023-08-01 | 13714200 |
| 2121 | 2021-03-24 | 9941600 |  | 2712 | 2023-08-02 | 12406100 |
| 2126 | 2021-04-01 | 10439300 |  | 2713 | 2023-08-03 | 17006600 |
| 2127 | 2021-04-02 | 13979400 |  | 2714 | 2023-08-04 | 18672300 |
| 2133 | 2021-04-12 | 12513500 |  | 2715 | 2023-08-07 | 28085700 |
| 2145 | 2021-04-29 | 16646300 |  | 2716 | 2023-08-08 | 15266300 |
| 2147 | 2021-05-05 | 17558400 |  | 2717 | 2023-08-09 | 15555900 |
| 2149 | 2021-05-07 | 12059100 |  | 2718 | 2023-08-10 | 12692000 |
| 2150 | 2021-05-10 | 13491300 |  | 2720 | 2023-08-14 | 10439500 |
| 2153 | 2021-05-13 | 11724300 |  | 2723 | 2023-08-17 | 10120500 |
| 2158 | 2021-05-20 | 12263200 |  | 2724 | 2023-08-18 | 20451400 |

Trực quan lại các điểm ngoại lai của biến Volume bằng biểu đồ phân tán:

A graph of red and blue dots

Description automatically generated

*Hình 2.19. Biểu đồ phân tán các điểm ngoại lai của biến Volume của ACB*

Phân tích các giá trị ngoại lai của biến volume của ACB

Các giá trị ngoại lai của biến Volume của ACB tập trung nhiều vào các khoảng thời gian năm 2020, 2021, 2023.

***Kết luận chung***

Các biến giá giao dịch, khối lượng giao dịch của 3 ngân hàng cùng đạt mức cao vào thời điểm năm 2018, cùng với đó các giá trị ngoại lai cũng tập trung nhiều vào thời điểm đó. Sau khi nhóm tìm hiểu và đối chiếu với thực tế thì nguyên do có thể đến từ việc tăng trưởng nóng của thị trường chứng khoán (Tạp chí Tài chính điện tử, 2018). Do thị trường chung đang có những dấu hiệu tích cực trong đó có những thông tin từ sự phát triển của nền kinh tế và sự tăng trưởng vĩ mô ổn định, bên cạnh đó qua các chính sách thúc đẩy kinh tế như cổ phần hóa các doanh nghiệp Nhà nước, đẩy mạnh tái cơ cấu, tăng năng lực cạnh tranh hay chính sách nới lỏng tiền tệ tạo tiền đề hỗ trợ doanh nghiệp đã và đang phát huy hiệu quả. Vì vậy, đây là động lực tăng trưởng trực tiếp cho nhóm ngành Ngân hàng năm 2018. Trong xu thế đi lên chung của thị trường, cổ phiếu ngân hàng tiếp tục là nhóm dẫn dắt thị trường, một vài phiên điều chỉnh sau đó tích lũy có thể diễn ra nhưng theo xu hướng dài hạn thì thị trường nói chung và nhóm ngân hàng nói riêng vẫn đang trong xu thế tăng (Tạp chí Tài chính điện tử, 2018), điều này cũng được thể hiện ở xu hướng dài hạn chung trong một vài năm tiếp theo (2018 - 2022) của hai ngân hàng VCB và BID, thấy rõ ở đồ thị đường thẳng. Riêng ngân hàng ACB lại có xu hướng đi xuống so với thị trường lý do có thể đến từ một vài lý do nội bộ như chiến lược phát triển, chính sách quản lý có thể chưa thực sự phù hợp.

### 2.3.4. Phân tích tương quan giữa các biến

Nhóm sử dụng hệ số tương quan Pearson để kiểm tra cường độ và hướng của mối quan hệ tuyến tính giữa hai biến liên tục.

Hệ số tương quan có thể có giá trị dao động từ −1 đến +1. Giá trị tuyệt đối của hệ số càng lớn thì mối quan hệ giữa các biến càng chặt chẽ. Đối với tương quan Pearson, giá trị tuyệt đối bằng 1 biểu thị mối quan hệ tuyến tính hoàn hảo. Mối tương quan gần bằng 0 cho thấy không có mối quan hệ tuyến tính giữa các biến.

Dấu của hệ số cho biết hướng của mối quan hệ. Nếu cả hai biến có xu hướng tăng hoặc giảm cùng nhau thì hệ số này dương và đường biểu thị mối tương quan dốc lên. Nếu một biến có xu hướng tăng trong khi biến kia giảm, thì hệ số âm và đường biểu thị mối tương quan dốc xuống.

Nhóm tiến hành tính toán và trực quan kết quả tương quan giữa các biến với nhau bằng biểu đồ nhiệt (heatmap).

***a) Tính toán độ tương quan giữa các biến trong VCB***

**Ma trận tương quan:**

               low      high      open     close  nmVolume

low      1.0000000 0.9992794 0.9994996 0.9994029 0.1887476

high     0.9992794 1.0000000 0.9994116 0.9995713 0.2048113

open     0.9994996 0.9994116 1.0000000 0.9989656 0.1948435

close    0.9994029 0.9995713 0.9989656 1.0000000 0.1982759

nmVolume 0.1887476 0.2048113 0.1948435 0.1982759 1.0000000

**P-value Matrix:**

         low high open close nmVolume

low       NA    0    0     0        0

high       0   NA    0     0        0

open       0    0   NA     0        0

close      0    0    0    NA        0

nmVolume   0    0    0     0       NA

A blue and white chart with white text

Description automatically generated

*Hình 2.20. Biểu đồ nhiệt thể hiện độ tương quan giữa các biến trong VCB*

Kết quả trên cho thấy các biến giá low, high, open, close của VCB có mối quan hệ đồng biến/mối tương quan tuyến tính tuyệt đối (giá trị tương quan bằng 1) với nhau. Điều này có nghĩa là khi một trong các giá này tăng hoặc giảm, các giá khác cũng có xu hướng tăng hoặc giảm tương tự.

Ngược lại, biến Volume với giá trị tương quan gần như bằng 0 với các biến còn lại cho thấy một mối tương quan yếu, là mối quan hệ độc lập với nhau, nhưng có ý nghĩa thống kê. Điều này cho thấy rằng khối lượng giao dịch không liên quan nhiều đến biến động giá.

P-values: Tất cả các p-value đều bằng 0 (ngoại trừ các giá trị trên đường chéo chính là NA). Điều này cho thấy rằng các hệ số tương quan đều có ý nghĩa thống kê ở mức ý nghĩa 0.05.

***b) Tính toán tương quan giữa các biến trong BID***

**Ma trận tương quan:**

               low       high       open      close   nmVolume

low      1.0000000 0.99872572 0.99916015 0.99892279 0.03806350

high     0.9987257 1.00000000 0.99897611 0.99913169 0.05826488

open     0.9991602 0.99897611 1.00000000 0.99793504 0.04359884

close    0.9989228 0.99913169 0.99793504 1.00000000 0.05161511

nmVolume 0.0380635 0.05826488 0.04359884 0.05161511 1.00000000

**P-value Matrix:**

                low        high       open      close    nmVolume

low              NA 0.000000000 0.00000000 0.00000000 0.058564416

high     0.00000000          NA 0.00000000 0.00000000 0.003771017

open     0.00000000 0.000000000         NA 0.00000000 0.030253443

close    0.00000000 0.000000000 0.00000000         NA 0.010298682

nmVolume 0.05856442 0.003771017 0.03025344 0.01029868          NA

A blue squares with white text

Description automatically generated

*Hình 2.21. Biểu đồ nhiệt thể hiện độ tương quan giữa các biến trong BID*

Các biến giá (low, high, open, close) có mối tương quan đồng biến rất mạnh với nhau. Điều này có nghĩa là khi một trong các giá này tăng hoặc giảm, các giá khác cũng có xu hướng tăng hoặc giảm tương tự.

Mối tương quan giữa khối lượng giao dịch (nmVolume) và các biến giá là khá yếu, nhưng hầu hết đều có ý nghĩa thống kê. Điều này cho thấy rằng khối lượng giao dịch có một mức độ tương quan nhỏ với biến động giá, nhưng không mạnh.

Các hệ số tương quan giữa các biến giá đều có ý nghĩa thống kê, do p-value của chúng đều bằng 0. Mối tương quan giữa hai biến low và volume không có ý nghĩa thống kê, do p-value là 0.058564416 > 0.05.

***c) Tính toán tương quan giữa các biến trong ACB***

**Ma trận tương quan:**

              low      high      open     close  nmVolume

low      1.0000000 0.9976629 0.9985813 0.9983389 0.4665238

high     0.9976629 1.0000000 0.9982651 0.9986966 0.4777618

open     0.9985813 0.9982651 1.0000000 0.9969580 0.4705393

close    0.9983389 0.9986966 0.9969580 1.0000000 0.4741330

nmVolume 0.4665238 0.4777618 0.4705393 0.4741330 1.0000000

**P-value Matrix:**

         low high open close nmVolume

low       NA    0    0     0        0

high       0   NA    0     0        0

open       0    0   NA     0        0

close      0    0    0    NA        0

nmVolume   0    0    0     0       NA

A blue and white chart

Description automatically generated with medium confidence

*Hình 2.22. Biểu đồ nhiệt thể hiện độ tương quan giữa các biến trong ACB*

Các biến giá (low, high, open, close) có mối tương quan đồng biến rất mạnh với nhau (xấp xỉ tuyệt đối bằng 1), điều này cho thấy khi một trong các giá này tăng hoặc giảm, các giá khác cũng có xu hướng tăng hoặc giảm tương tự.

Mối tương quan giữa khối lượng giao dịch (nmVolume) và các biến giá là trung bình (gần bằng 0.5) và có ý nghĩa thống kê, cho thấy có một mức độ tương quan nhất định giữa khối lượng giao dịch và giá.

Tất cả các hệ số tương quan đều có ý nghĩa thống kê, cho thấy rằng các mối tương quan này là đáng kể.

## 2.4. Tiền xử lí dữ liệu

Việc chuẩn hóa dữ liệu trong một mô hình học máy nói chung hay học sâu nói riêng có thể cải thiện thuật toán tối ưu hóa bằng cách làm cho quá trình hoạt động của thuật toán gradient descent trở nên mượt mà hơn và tìm cực tiểu của hàm chi phí nhanh chóng hơn. Dữ liệu đầu vào không được chuẩn hóa có thể khiến mô hình bị sai lệch (biased) về phía các giá trị lớn. (Enjoy Computer Science, n.d., para. 20).

Với lý do trên, nhóm đã thực hiện tiền xử lý dữ liệu bằng cách chuẩn hóa dữ liệu bằng phương pháp min-max để đưa giá đóng cửa về khoảng [0,1]. Phương pháp này đảm bảo rằng tất cả các giá trị đầu vào đều nằm trong một khoảng nhất định, giúp cân bằng các giá trị và cải thiện hiệu suất của cả hai mô hình mà nhóm đang phát triển.

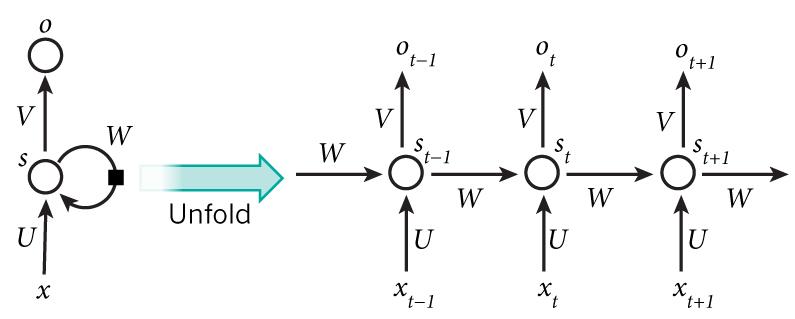
# Xây dựng mô hình dự báo bằng Deep learning

## Mô hình RNN

### Giới thiệu mô hình

Mạng nơ-ron hồi tiếp (RNN) là một loại mô hình mạng nơ-ron học sâu được huấn luyện để xử lý và chuyển đổi dữ liệu tuần tự đầu vào để trả về dữ liệu tuần tự đầu ra cụ thể (AWS, n.d.). Trong mô hình mạng nơ-ron truyền thống, dữ liệu đầu vào 𝑥 sẽ đi qua lớp ẩn s và trả về đầu ra *o* với kết nối đầy đủ giữa các nút, nhưng không có sự kết nối giữa các lớp. Vì vậy, tất cả các đầu vào và đầu ra từ mô hình mạng nơ-ron truyền thống đều độc lập với nhau (Denny’s Blog, 2015). Điều đó khiến mô hình này không phù hợp với những bài toán mà kết quả dự đoán phụ thuộc vào các dữ liệu trước đó. Như vậy, mô hình này không phù hợp với những bài toán mà kết quả dự đoán phụ thuộc vào các dữ liệu trước đó như dịch tự động, nhận diện giọng nói, phân tích chuỗi thời gian,...

RNN là mô hình phù hợp hơn để giải quyết các bài toán như thế. RNN sử dụng một “bộ nhớ” để lưu lại thông tin từ các bước tính toán trước, dựa vào đó có thể đưa ra dự đoán chính xác nhất cho bước dự đoán hiện tại. RNN được gọi là hồi quy vì chúng thực hiện cùng một nhiệm vụ cho mọi phần tử của một chuỗi, với đầu ra phụ thuộc vào các tính toán trước đó (Zhu, 2020). Hình 3.1 minh họa cho một mô hình RNN cơ bản.



*Hình 3.1. Mô hình RNN cơ bản (Nguồn: Denny’s Blog)*

Trong mô hình RNN, các input 𝑥𝑡 sẽ được kết hợp với hidden layer s𝑡-1 bằng hàm 𝑓𝑊 với W là ma trận các trọng số để tính toán ra hidden layer s𝑡. Sau đó, output o𝑡 sẽ được tính ra từ s𝑡. Như vậy kết quả từ các quá trình tính toán trước t đã được "nhớ" bằng cách kết hợp thêm s𝑡-1 tính ra s𝑡, giúp tăng độ chính xác cho những dự đoán ở hiện tại. Hàm 𝑓 thường là một hàm phi tuyến như tanh hoặc ReLU.

Huấn luyện mô hình RNN cũng sử dụng thuật toán lan truyền ngược của sai số (back propagation of errors), nhưng có một chút khác biệt. Trong quá trình huấn luyện, các tham số W, U, V được chia sẻ, trong khi ở mô hình mạng nơ-ron truyền thống không thì không. Và đầu ra của thuật toán gradient descent ở mỗi bước của mô hình RNN không chỉ phụ thuộc vào bước hiện tại của mạng, mà còn phụ thuộc vào trạng thái của các bước trước đó của mạng.

Hình 3.1 trên có đầu ra tại mỗi bước thời gian, nhưng điều này có thể không cần thiết tùy thuộc vào bài toán. Ví dụ, khi dự đoán cảm xúc trong một câu, đầu ra cuối cùng là kết quả cần thiết chứ không phải cảm xúc trong mỗi từ. Tương tự, đầu vào tại mỗi bước thời gian cũng không cần thiết.

### Chỉ số đánh giá

Kết quả dự đoán của mô hình được đánh giá dựa trên 4 chỉ số sau:

***- MSE (Mean Squared Error)***

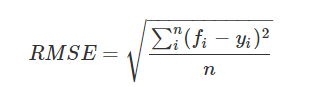
MSE cho thấy sai số bình phương trung bình giữa các giá trị được dự đoán và thực tế. MSE không âm và các giá trị càng gần 0 thì mô hình càng chính xác. Công thức tính MSE như sau:



Trong đó, n là số điểm dữ liệu, fᵢ là giá trị quan sát và yᵢ là giá trị dự đoán.

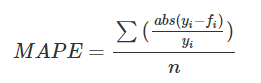
***- RMSE***

Giá trị RMSE là căn bậc hai của mức trung bình của các sai số bình phương, cho biết mức độ tập trung của dữ liệu xung quanh đường phù hợp nhất. RMSE càng thấp thì mô hình càng chính xác. Công thức tính RMSE như sau:



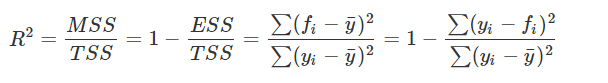
***- MAPE***

Giá trị MAPE cho thấy sai số trung bình tuyệt đối giữa giá trị dự đoán và giá trị thực tế. MAPE càng thấp thì mô hình càng chính xác. Công thức tính MAPE như sau:



***- R-Squared***

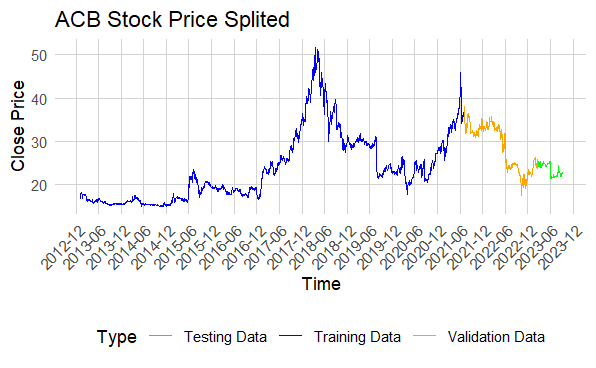
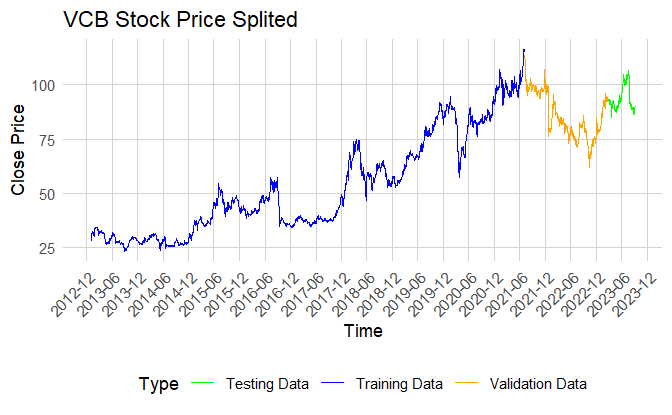
Giá trị R-Squared cho thấy mức độ mà biến phụ thuộc được giải thích bởi mô hình. R-Squared càng gần 1 thì mô hình càng phù hợp với dữ liệu.

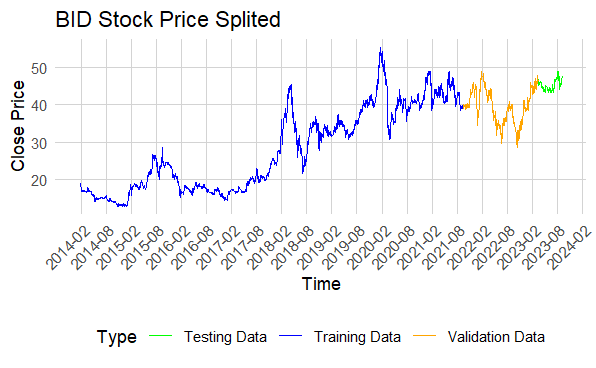


### Chuẩn bị dữ liệu

Sau khi được chuẩn hóa, dữ liệu của 3 ngân hàng được chia thành 3 tập con gồm train\_data (80%), validation\_data (15%) và test\_data (5%) được biểu diễn trong Hình 3.2.

Trong đó, tập train\_data là tập dữ liệu được sử dụng để huấn luyện mô hình, giúp xác định các trọng số và độ lệch của mô hình. Mô hình sẽ “nhìn thấy” và học từ dữ liệu này. Tập validation\_data được sử dụng để kiểm thử trong quá trình huấn luyện nhằm tinh chỉnh các siêu tham số của mô hình, và tập test\_data được dùng làm tiêu chuẩn để đánh giá mô hình sau quá trình huấn luyện.



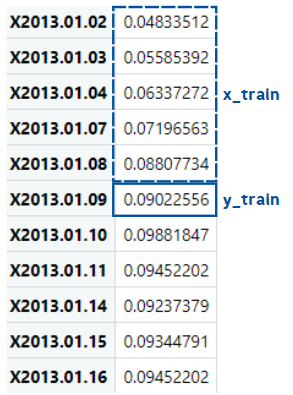


*Hình 3.2. Biểu đồ phân chia tập con của dữ liệu ngân hàng VCB, ACB, BID*

### Huấn luyện mô hình và đánh giá

Với dữ liệu của mỗi ngân hàng, bài phân tích này chỉ tập trung đưa ra mô hình dự đoán giá đóng cửa của cổ phiếu nên nhóm chỉ thực hiện huấn luyện mô hình với dữ liệu đầu vào là giá trị của giá đóng cửa. *Window size (kích thước cửa sổ)* trong lĩnh vực xử lý dữ liệu và học máy thường đề cập đến một khoảng thời gian hoặc số lượng mẫu dữ liệu liên tiếp được xem xét đồng thời để phân tích hoặc dự đoán. Trong bài phân tích này, window size là số lượng bước thời gian (timestep) trước đó được sử dụng để dự đoán giá trị tiếp theo. Ví dụ, nếu window size = 10, mô hình sẽ sử dụng giá đóng cửa của 10 bước thời gian trước đó để đưa ra dự đoán giá trị giá đóng cửa tiếp theo. Nhằm tìm ra window size tối ưu nhất cho mô hình, dữ liệu của mỗi ngân hàng sẽ được huấn luyện và đánh giá hiệu suất lần lượt qua các window size từ 1 đến 15.

Với mỗi window size, các tập dữ liệu được array thành các array phù hợp để đưa vào mô hình. Ví dụ như với window size = 5, tập dữ liệu train\_data sẽ được chia thành x\_train gồm các giá đóng cửa 5 ngày và y\_train là giá đóng cửa ngày thứ 6 tương ứng. Các tập dữ liệu validation\_data và test\_data cũng được xử lý tương tự. Hình 3.3 minh họa x\_train và y\_train.



*Hình 3.3. x\_train và y\_train với window size = 5*

Sau khi xử lý dữ liệu đầu vào, nhóm tạo và đưa dữ liệu huấn huyện vào mô hình RNN gồm lớp hidden layer với các siêu tham số như trong Bảng 1.

*Bảng 1. Các siêu tham số của mô hình RNN*

|  |  |
| --- | --- |
| layer\_simple\_rnn | 4 |
| units | 64; 128; 256; 512 |
| activation | tanh |
| dropout rate | 0.2 |
| optimizer | adam |
| loss | mean\_squared\_error |
| metrics | accuracy |
| epochs | 40 |
| batch\_size | 32 |

Sau khi lần lượt huấn luyện và đánh giá hiệu suất của mô hình dựa trên 4 chỉ số đã chọn, nhóm đã tổng hợp được các đánh giá về thông số như trong các bảng sau:

*Bảng 2. Kết quả đánh giá mô hình RNN trên dữ liệu ngân hàng VCB*

| Window | MSE | RMSE | MAPE | R-Squared |
| --- | --- | --- | --- | --- |
| **1** | **0.000399** | **0.019969** | **1.770766** | **0.882238** |
| 2 | 0.000770 | 0.027754 | 2.768323 | 0.772523 |
| 3 | 0.001223 | 0.034978 | 3.985002 | 0.638690 |
| 4 | 0.000689 | 0.026248 | 2.419054 | 0.796533 |
| 5 | 0.000679 | 0.026054 | 2.330023 | 0.799537 |
| 6 | 0.003778 | 0.061466 | 7.365655 | -0.115748 |
| 7 | 0.001523 | 0.039026 | 4.084504 | 0.550220 |
| 8 | 0.000794 | 0.028182 | 2.585864 | 0.765451 |
| 9 | 0.000765 | 0.027661 | 2.646640 | 0.774030 |
| 10 | 0.000721 | 0.026844 | 2.486457 | 0.787197 |
| 11 | 0.001234 | 0.035129 | 3.663881 | 0.635550 |
| 12 | 0.000880 | 0.029663 | 2.862439 | 0.740137 |
| 13 | 0.000894 | 0.029895 | 3.021387 | 0.736062 |
| 14 | 0.000731 | 0.027042 | 2.612803 | 0.784041 |
| 15 | 0.000969 | 0.031122 | 3.157508 | 0.713956 |

*Bảng 3. Kết quả đánh giá mô hình RNN trên dữ liệu ngân hàng ACB*

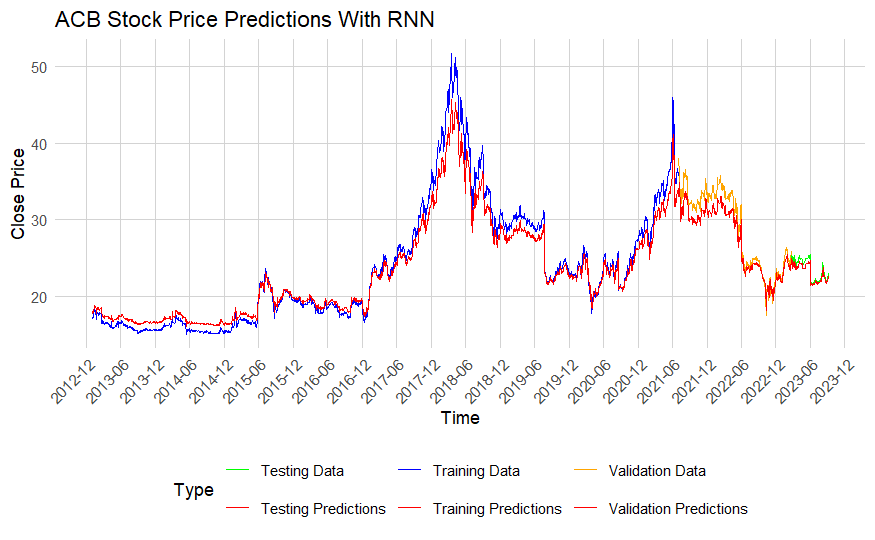
| Window | MSE | RMSE | MAPE | R-Squared |
| --- | --- | --- | --- | --- |
| **1** | **0.00014** | **0.011828** | **2.648492** | **0.898165** |
| 2 | 0.000206 | 0.014349 | 3.972940 | 0.850131 |
| 3 | 0.000251 | 0.015856 | 4.419946 | 0.816989 |
| 4 | 0.000402 | 0.020054 | 6.881329 | 0.707263 |
| 5 | 0.000355 | 0.018832 | 6.200716 | 0.741844 |
| 6 | 0.000362 | 0.019032 | 5.954971 | 0.736353 |
| 7 | 0.000401 | 0.020029 | 6.005281 | 0.707992 |
| 8 | 0.000298 | 0.017275 | 5.225281 | 0.782785 |
| 9 | 0.000253 | 0.015906 | 4.521315 | 0.815839 |
| 10 | 0.000266 | 0.016299 | 5.072792 | 0.806617 |
| 11 | 0.001144 | 0.033816 | 12.800730 | 0.167617 |
| 12 | 0.000843 | 0.029033 | 11.992100 | 0.386403 |
| 13 | 0.000236 | 0.01535 | 5.012636 | 0.828485 |
| 14 | 0.000538 | 0.023195 | 8.248867 | 0.608386 |
| 15 | 0.000682 | 0.026118 | 10.228599 | 0.503474 |

*Bảng 4. Kết quả đánh giá mô hình RNN trên dữ liệu ngân hàng BIDV*

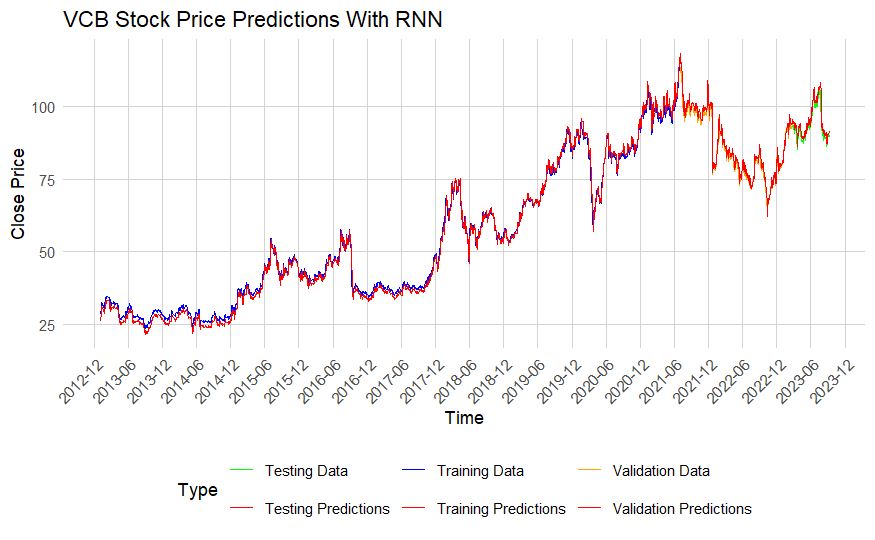
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Window | MSE | RMSE | MAPE | R-Squared |
| 1 | 0.000417 | 0.020423 | 2.205335 | 0.620752 |
| **2** | **0.000261** | **0.016142** | **1.416364** | **0.763077** |
| 3 | 0.000452 | 0.021271 | 2.224658 | 0.588607 |
| 4 | 0.001934 | 0.043972 | 5.387990 | -0.758125 |
| 5 | 0.001006 | 0.031715 | 3.610515 | 0.085399 |
| 6 | 0.002143 | 0.046287 | 5.591866 | -0.948132 |
| 7 | 0.000702 | 0.026500 | 2.852659 | 0.361454 |
| 8 | 0.001188 | 0.034468 | 3.954910 | -0.080228 |
| 9 | 0.002546 | 0.050455 | 6.236541 | -1.314738 |
| 10 | 0.001248 | 0.035327 | 4.220076 | -0.134773 |
| 11 | 0.000855 | 0.029239 | 3.365563 | 0.222632 |
| 12 | 0.000546 | 0.023360 | 2.530416 | 0.503809 |
| 13 | 0.000493 | 0.022203 | 2.398759 | 0.551773 |
| 14 | 0.000549 | 0.023426 | 2.591681 | 0.500996 |
| 15 | 0.000733 | 0.027080 | 3.115700 | 0.333209 |

Kết quả trên cho thấy không có mức biến động lớn trong chỉ số MSE và RMSE của mô hình trên dữ liệu của cả 3 ngân hàng, trong khi chỉ số MAPE và R-Squared với window size = 1 ở ngân hàng VCB, ACB và window size = 2 ở ngân hàng BIDV lại tốt hơn đáng kể.

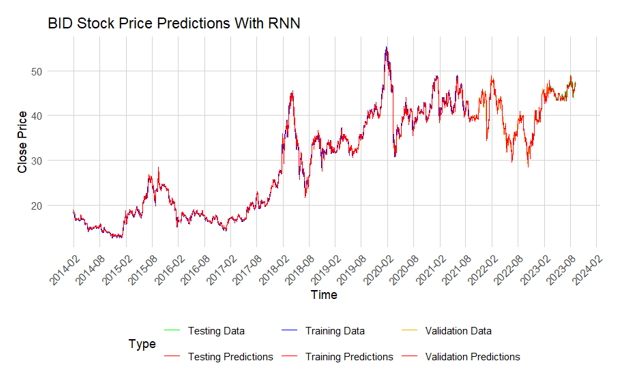
Với kết quả của từng ngân hàng như trên trên, nhóm sẽ tiến hành chạy các dự đoán về giá của từng ngân hàng sử dụng window size tốt nhất của từng ngân hàng và trực quan hóa kết quả qua các biểu đồ:



*Hình 3.4. Biểu đồ kết quả dự đoán của mô hình RNN trên dữ liệu ngân hàng ACB*



*Hình 3.5. Biểu đồ kết quả dự đoán của mô hình RNN trên dữ liệu ngân hàng VCB*



*Hình 3.6. Biểu đồ kết quả dự đoán của mô hình RNN trên dữ liệu ngân hàng BIDV*

## Mô hình LSTM

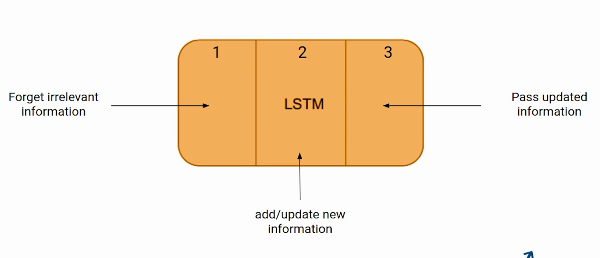
### Giới thiệu mô hình

LSTM là một loại mô hình RNN sử dụng để dự đoán về các loại dữ liệu Time Series, trong đó bao gồm giá về chứng khoán và lợi nhuận đầu tư.

LSTM là một loại mạng nơ-ron nhân tạo được thiết kế để xử lý dữ liệu tuần tự, chẳng hạn như chuỗi thời gian, âm thanh hoặc văn bản. Nó đặc biệt hữu ích để xử lý dữ liệu có sự phụ thuộc lâu dài, trong đó đầu ra tại một bước thời gian nhất định phụ thuộc vào thông tin từ các bước thời gian trước đó. LSTM cũng có thể xử lý dữ liệu nhiễu và phi tuyến tính, thường xảy ra với giá cổ phiếu, do các yếu tố bên ngoài khác nhau như sự kiện tin tức, chỉ số kinh tế và tâm lý thị trường. Hơn nữa, mô hình có thể xử lý lượng lớn dữ liệu một cách hiệu quả, điều này rất quan trọng để phân tích và dự đoán giá cổ phiếu được cập nhật thường xuyên, chẳng hạn như dữ liệu hàng ngày hoặc trong ngày.

LSTM phù hợp với dữ liệu chuỗi thời gian về giá cổ phiếu vì giá cổ phiếu thể hiện các mô hình hoặc sự phụ thuộc nhất định theo thời gian, chẳng hạn như xu hướng và tính thời vụ. Những mẫu này thường khó nắm bắt bằng các mô hình thống kê truyền thống, nhưng LSTM có thể nắm bắt chúng một cách hiệu quả và đưa ra dự đoán chính xác.

**Cách thức hoạt động**

****

*Hình 3.7. Cấu trúc cơ bản về mô hình LSTM*

Mạng LSTM có thể ghi nhớ thông tin này trong một khoảng thời gian dài hơn bằng cách sử dụng các ô nhớ, cổng đầu vào, cổng đầu ra và cổng quên. Các cổng này kiểm soát luồng thông tin vào và ra khỏi các ô nhớ, cho phép mạng lưu trữ và truy xuất thông tin có chọn lọc khi cần. LSTM thường được sử dụng cho các nhiệm vụ như dịch ngôn ngữ, nhận dạng giọng nói và dự đoán giá cổ phiếu**.** Chúng hoạt động bằng cách xử lý một chuỗi dữ liệu đầu vào, mỗi lần một phần tử, trong khi vẫn duy trì “bộ nhớ” của các dữ liệu đầu vào trước đó. Bộ nhớ này cho phép LSTM nắm bắt các yếu tố phụ thuộc lâu dài trong trình tự, chẳng hạn như các xu hướng hoặc mô hình tồn tại theo thời gian.

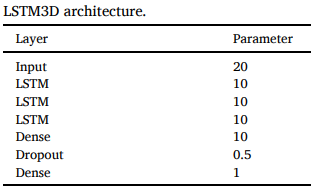
*Bảng 5. So sánh sự khác biệt giữa mô hình RNN và mô hình LSTM*

|  |  |  |
| --- | --- | --- |
| **Yếu tố** | **LSTM** | **RNN** |
| **Cấu trúc** | Một loại RNN có thêm ô nhớ | RNN cơ bản |
| **Duy trì bộ nhớ** | Xử lý các phần phụ thuộc dài hạn và ngăn chặn vấn đề biến mất độ dốc | Gặp khó khăn với sự phụ thuộc lâu dài và vấn đề độ dốc biến mất |
| **Cấu trúc cell** | Cấu trúc ô phức tạp với các cổng đầu vào, đầu ra và cổng quên | Cấu trúc ô đơn giản chỉ có một hidden state |
| **Trình tự xử lý** | Thích hợp để xử lý dữ liệu tuần tự | Cũng được thiết kế cho dữ liệu tuần tự nhưng bộ nhớ hạn chế |
| **Hiệu quả đào tạo** | Quá trình đào tạo chậm hơn do độ phức tạp tăng lên | Quá trình đào tạo nhanh hơn do kiến trúc đơn giản hơn |
| **Hiệu suất trên chuỗi dài** | Hoạt động tốt hơn trên các chuỗi dài | Gặp khó khăn khi lưu giữ thông tin về chuỗi dài |
| **Công dụng** | Phù hợp nhất cho các công việc đòi hỏi trí nhớ dài hạn, chẳng hạn như dịch ngôn ngữ và phân tích tình cảm | Thích hợp cho các nhiệm vụ tuần tự đơn giản, chẳng hạn như dự báo chuỗi thời gian |
| **Vấn đề về độ dốc biến mất** | Giải quyết vấn đề độ dốc biến mất | Dễ gặp vấn đề về độ dốc biến mất |

### Huấn luyện mô hình

Tương tự với RNN, nhóm sẽ tiến hành chia dữ liệu thành 3 phần training data (80%), validation data (15%) và testing data (5%) và để tìm ra window size tối ưu nhất cho mô hình, mô hình LSTM sẽ được huấn luyện và đánh giá hiệu suất lần lượt từng ngân hàng qua các window size từ 1 đến 10.

Sau khi dữ liệu đầu vào đã được xử lý, mô hình LSTM sẽ được huấn luyện dựa vào mô hình LSTM3D (Burak Gülmez, 2023). Mô hình LSTM3D gồm 7 lớp trong đó có 3 lớp LSTM với tham số mỗi lớp là 10 units và các tham số như sau:



*Hình 3.8. Kiến trúc LSTM*

*Bảng 6. Các siêu tham số của mô hình LSTM*

|  |  |
| --- | --- |
| learning rate | 0.01 |
| batch size | 4 |
| epochs | 40 |
| optimizer | adam |
| loss | mean squared error |

Sau khi lần lượt huấn luyện và đánh giá kết quả của các windows size, nhóm đã tổng hợp được các đánh giá về thông số của VCB, ACB, và BID lần lượt như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Window Size** | **MSE** | **RMSE** | **MAPE** | **R-Squared** |
| **1** | **0.000808** | **0.028424** | **1.936273** | **0.857689** |
| 2 | 0.033802 | 0.183854 | 23.13228 | -10.0845 |
| 3 | 0.079966 | 0.282782 | 30.16411 | -15.8566 |
| 4 | 0.016899 | 0.129996 | 6.370067 | 0.018116 |
| 5 | 0.017572 | 0.132558 | 19.12816 | -6.36131 |
| 6 | 0.018222 | 0.134989 | 11.44513 | -2.34362 |
| 7 | 0.260372 | 0.510267 | 55.15096 | -53.0163 |
| 8 | 0.078833 | 0.280773 | 30.25631 | -15.9002 |
| 9 | 0.050123 | 0.223881 | 20.02565 | -7.11811 |
| 10 | 0.100959 | 0.317741 | 20.89402 | -7.96989 |

*Bảng 7. Kết quả đánh giá mô hình LSTM trên dữ liệu ngân hàng VCB*

*Bảng 8. Kết quả đánh giá mô hình LSTM trên dữ liệu ngân hàng ACB*

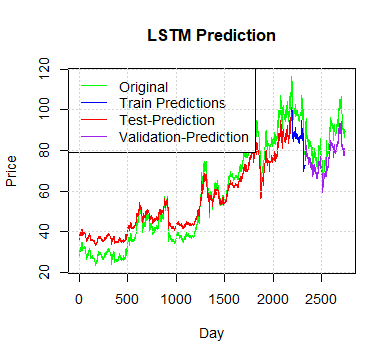
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Window Size** | **MSE** | **RMSE** | **MAPE** | **R-Squared** |
| 1 | 0.000175 | 0.013229 | 3.355911 | 0.872608 |
| 2 | 0.000251 | 0.015847 | 4.866712 | 0.817195 |
| 3 | 0.000278 | 0.016683 | 5.690928 | 0.797405 |
| 4 | 0.000197 | 0.014026 | 4.105943 | 0.856797 |
| 5 | 0.000171 | 0.013084 | 3.387743 | 0.875388 |
| **6** | **0.000158** | **0.012572** | **3.270681** | **0.884946** |
| 7 | 0.00018 | 0.013434 | 3.485627 | 0.868638 |
| 8 | 0.000617 | 0.024843 | 9.831519 | 0.550768 |
| 9 | 0.000274 | 0.01655 | 5.965983 | 0.800622 |
| 10 | 0.001331 | 0.036478 | 15.47321 | 0.031392 |

*Bảng 9. Kết quả đánh giá mô hình LSTM trên dữ liệu ngân hàng BIDV*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Window Size** | **MSE** | **RMSE** | **MAPE** | **R-Squared** |
| 1 | 0.000426 | 0.02064 | 2.243192 | 0.612655 |
| 2 | 0.000271 | 0.016468 | 1.467959 | 0.753421 |
| 3 | 0.001408 | 0.037529 | 4.530668 | -0.28061 |
| 4 | 0.001425 | 0.03775 | 4.586349 | -0.29575 |
| 5 | 0.000321 | 0.017914 | 1.766632 | 0.708206 |
| **6** | **0.000236** | **0.015352** | **1.458489** | **0.785687** |
| 7 | 0.004356 | 0.066002 | 8.366603 | -2.96107 |
| 8 | 0.000703 | 0.026514 | 3.039877 | 0.360803 |
| 9 | 0.001149 | 0.033899 | 4.089639 | -0.04487 |
| 10 | 0.000375 | 0.019353 | 1.951372 | 0.659443 |

Kết quả cho thấy rằng đối với từng ngân hàng hay từng dữ liệu sẽ cho ra những kết quả khác nhau. Đối với ngân hàng VCB, window size hiệu quả nhất là 1, nghĩa là sử dụng giá đóng cửa của ngày hôm trước để đoán giá đóng cửa của ngày hôm sau, hiệu quả nhất so với các window size còn lại với các chỉ số MSE, RMSE, MAPE, và R-Square đều tốt hơn đáng kể so với các window size khác, đặc biệt là chỉ số MAPE có sự khác biệt lớn nhất.

Với kết quả của từng ngân hàng như trên trên, nhóm sẽ tiến hành chạy các dự đoán về giá của từng ngân hàng sử dụng window size tốt nhất của từng ngân hàng và trực quan hóa kết quả qua các biểu đồ:

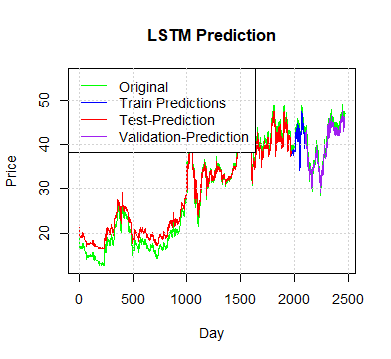
******

*Hình 3.9. Biểu đồ kết quả dự đoán của mô hình LSTM trên dữ liệu ngân hàng VCB*

***A graph with red green and blue lines

Description automatically generated***

*Hình 3.10 Biểu đồ kết quả dự đoán của mô hình LSTM trên dữ liệu ngân hàng ACB*

******

*Hình 3.11. Biểu đồ kết quả dự đoán của mô hình LSTM trên dữ liệu ngân hàng BIDV*

Nhìn chung, các kết quả dự đoán đều khá sát với dữ liệu gốc, với các đường dự đoán bám sát dữ liệu được huấn luyện và các dữ luyện được kiểm tra. Dự đoán của ngân hàng BIDV có kết quả chính xác nhất với đường dự đoán gần như trùng với đường dữ liệu gốc.

## Kết luận mô hình

*Bảng 10. Chỉ số đánh giá hoạt động tốt nhất của mô hình RNN trên dữ liệu của ba ngân hàng*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ngân hàng** | **Window** | **MSE** | **RMSE** | **MAPE** | **R-Squared** |
| **VCB** | 1 | 0.000399 | 0.019969 | 1.770766 | 0.882238 |
| **ACB** | 1 | 0.00014 | 0.011828 | 2.648492 | 0.898165 |
| **BIDV** | 2 | 0.000261 | 0.016142 | 1.416364 | 0.763077 |

*Bảng 11. Chỉ số đánh giá hoạt động tốt nhất của mô hình LSTM trên dữ liệu của ba ngân hàng*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ngân hàng** | **Window** | **MSE** | **RMSE** | **MAPE** | **R-Squared** |
| **VCB** | 1 | 0.000808 | 0.028424 | 1.936273 | 0.857689 |
| **ACB** | 6 | 0.000158 | 0.012572 | 3.270681 | 0.884946 |
| **BIDV** | 6 | 0.000236 | 0.015352 | 1.458489 | 0.785687 |

Hai bảng trên liệt kê các chỉ số tốt nhất của mô hình RNN và LSTM trên dữ liệu của ba ngân hàng VCB, ACB, và BIDV. Khi so sánh trên cùng dữ liệu ngân hàng, chỉ số của mô hình RNN và LSTM không có sự chênh lệch đáng kể và cả hai đều hoạt động tốt.

Tuy nhiên, mô hình RNN hoạt động tốt hơn một chút trên cả ba dữ liệu ngân hàng, mặc dù mô hình LSTM được đánh giá cao hơn về mặt lý thuyết. Điều này có thể là do dữ liệu của cả ba ngân hàng còn khá ít (trung bình 2.650 dòng/ngân hàng), đủ để mô hình RNN hoạt động tốt nhưng chưa đủ để mô hình LSTM phát huy hiệu quả.

# Kết luận và khuyến nghị

## Tóm tắt kết quả và những phát hiện quan trọng

Sau khi lần lượt huấn luyện và đánh giá kết quả bằng các chỉ số như MSE, RMSE, MAPE, R-squared của 2 mô hình, nhóm nhận thấy đối với mô hình RNN các chỉ số của tập dữ liệu cả 3 ngân hàng đều cho kết quả tốt nhất với windows size là 1 (ngân hàng BIDV cho kết quả tốt hơn ở windows size = 2 nhưng không đáng kể so với 1). Về LSTM, mô hình cho kết quả tốt nhất ở windows size là 1 với tập dữ liệu của ngân hàng VCB, tuy nhiên các chỉ số của tập dữ liệu của 2 ngân hàng còn lại ACB, BIDV lại cho kết quả tối ưu nhất ở windows size = 6.

Nhìn chung, có thể thấy các mô hình được huấn luyện tốt vì kết quả dự đoán khá gần với dữ liệu thực tế của ba ngân hàng, với các đường dự đoán theo sát cả dữ liệu huấn luyện lẫn kiểm tra. Đặc biệt, dự đoán cho ngân hàng BIDV là chính xác nhất ở cả 2 mô hình, với đường dự đoán gần như trùng khớp hoàn toàn với dữ liệu gốc.

## Đánh giá hạn chế

Mặc dù hai mô hình RNN, LSTM đã có nhiều cải tiến so với các mô hình truyền thống nhưng hai mô hình này vẫn còn một số hạn chế:

+ ***Cấu trúc khá phức tạp và thời gian huấn luyện dài***: Cả RNN và LSTM đều có cấu trúc khá phức tạp và yêu cầu thời gian huấn luyện dài, đặc biệt là với các chuỗi thời gian dài và dữ liệu lớn.

+ ***Overfitting***: Khi mô hình quá phức tạp hoặc có quá nhiều tham số, có thể dẫn đến việc học quá kỹ các chi tiết của dữ liệu huấn luyện và kém hiệu quả trong việc dự đoán dữ liệu mới. Điều này đặc biệt quan trọng khi dự đoán giá chứng khoán, nơi dữ liệu thường nhiễu.

**+ *Nhạy cảm với thông số và cấu trúc mô hình***: Hiệu suất của RNN và LSTM phụ thuộc rất nhiều vào các tham số và cấu trúc của mô hình, như số lượng lớp ẩn, số lượng đơn vị trong mỗi lớp, và các tham số điều chỉnh khác. Việc tìm ra cấu trúc tối ưu và các tham số phù hợp có thể tốn nhiều thời gian.

Với RNN, ô hình này hoạt động tốt nhất khi sử dụng dữ liệu từ ngày hôm trước để dự đoán giá của ngày hôm sau, đặc biệt với những dữ liệu có hệ số tự tương quan (ACF) cao. Tuy nhiên, điều này khiến cho RNN gặp khó khăn trong việc bao quát các tính chất thời vụ và xu hướng dài hạn của dữ liệu. RNN cũng thường gặp vấn đề về quên ngược (vanishing gradient), làm giảm khả năng ghi nhớ và học từ các chuỗi dữ liệu dài. Những hạn chế này làm cho RNN không phải là lựa chọn lý tưởng cho các bài toán cần nắm bắt các xu hướng phức tạp và biến động theo mùa.

Với LSTM, mô hình được huấn luyện theo trình tự thời gian, điều này khiến cho quá trình huấn luyện trở nên tốn nhiều thời gian và phức tạp vì LSTM tiêu tốn rất nhiều tài nguyên xử lý, bao gồm cả bộ nhớ và sức mạnh tính toán, làm cho việc áp dụng nó vào các phân tích thời gian thực trở nên khó khăn. Ngoài ra, việc xử lý các chuỗi dữ liệu dài có thể dẫn đến tình trạng quên ngược (vanishing gradient), làm giảm hiệu suất của mô hình. Những hạn chế này khiến cho LSTM không phải lúc nào cũng là lựa chọn tối ưu cho các bài toán yêu cầu xử lý nhanh và hiệu quả trong thời gian thực.

Ngoài ra, dữ liệu giá chứng khoán thường xuyên biến động mạnh và có nhiều yếu tố bên ngoài tác động, như tin tức thị trường, tâm lý nhà đầu tư, và các sự kiện kinh tế vĩ mô. Vì vậy, việc chỉ dùng giá đóng cửa làm đầu vào cho hai mô hình RNN và LSTM có thể gặp khó khăn trong việc dự đoán các biến động này một cách chính xác.

## Hướng phát triển và khuyến nghị

Để cải thiện hiệu quả của mô hình RNN và LSTM trong dự đoán giá chứng khoán, có thể thực hiện thêm các nghiên cứu theo một số hướng phát triển sau:

Đầu tiên, có thể kết hợp thêm các giá trị có sẵn trong dữ liệu như giá mở cửa (open), giá thấp nhất (low), giá cao nhất (high), và các chỉ báo kỹ thuật khác để đưa ra dự đoán toàn diện hơn, không chỉ dựa trên giá đóng cửa.

Thứ hai, thay vì dự đoán giá chính xác, mô hình có thể được áp dụng để dự đoán xu hướng tăng/giảm dựa trên cột thay đổi (change), giúp nâng cao tính chính xác của kết quả dự đoán cũng như giúp nhà đầu tư có cái nhìn rõ ràng hơn về xu hướng thị trường.

Ngoài ra, sử dụng thêm kỹ thuật tối ưu hóa tham số như Adaptive Random Optimization (ARO) để tìm ra các hyperparameter tốt nhất cho mô hình như tối ưu hóa số lượng lớp ẩn, số lượng đơn vị trong mỗi lớp, và các tham số khác để đạt được hiệu suất cao nhất.

Tài liệu tham khảo

AWS. (n.d.). *What Is RNN?* Retrieved May 10, 2024 from

[https://aws.amazon.com/what-is/recurrent-neural-network/#:~:text=A%20 recurrent%20 neural%20network%20](https://aws.amazon.com/what-is/recurrent-neural-network/#:~:text=A%20recurrent%20neural%20network%20)

Bytepeaker. (2021, August 31). *LSTM | Introducción a LSTM*. Datapeaker. <https://datapeaker.com/en/big--data/lstm-introduction-to-lstm/>

Denny’s Blog. (September 17, 2015). Recurrent Neural Networks Tutorial, Part 1 – Introduction to RNNs [Blog post]. Retrieved from <https://dennybritz.com/posts/wildml/recurrent-neural-networks-tutorial-part-1/>

Enjoy Computer Science. (n.d.). Introduction to Feature Scaling: Normalization and

Standardization. Retrieved May 10, 2024 from <https://www.enjoyalgorithms.com/blog/need-of-feature-scaling-in-machine-learning>

Gülmez, B. (2023). Stock price prediction with optimized deep LSTM network with artificial rabbits optimization algorithm. *Expert Systems With Applications*, *227*, 120346. <https://doi.org/10.1016/j.eswa.2023.120346>

Tạp chí Tài chính điện tử. (2018). *Cổ phiếu ngân hàng có đang tăng trưởng quá nóng?*. Retrieved from

<https://tapchitaichinh.vn/co-phieu-ngan-hang-co-dang-tang-truong-qua-nong.html>

Zhu, Y. (2020). Stock price prediction using the RNN model. Journal of Physics: Conference Series, 1650. doi:10.1088/1742-6596/1650/3/032103