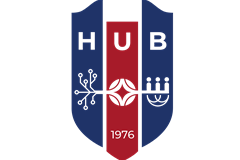
A blue design on a black background

Description automatically generated

TRƯỜNG ĐẠI HỌC NGÂN HÀNG TP. HỒ CHÍ MINH

**KHOA HỆ THỐNG THÔNG TIN QUẢN LÝ**

**🙤 🕮 🙦**





**CHỦ ĐỀ: Dự đoán giá cổ phiếu dựa trên dữ liệu giao dịch cổ phiếu trên các sàn giao dịch chứng khoán tại Việt Nam**

**Giảng viên hướng dẫn:**

TS. Mai Ngọc Thắng

**Lớp:**

D01

**Sinh viên thực hiện:**

Nguyễn Anh Tuấn

TPHCM, ngày 07 tháng 04 năm 2024

tT

030237210203

**MSSV:**

**Mục lục**

[**1.** **Giới thiệu về đề tài nghiên cứu** 3](#_Toc163915659)

[**1.1.** **Tổng quan** 3](#_Toc163915660)

[**1.2.** **Đặc điểm của bài toán** 3](#_Toc163915661)

[**1.3.** **Hướng giải quyết** 4](#_Toc163915662)

[**1.4.** **Kết quả dự kiến** 4](#_Toc163915663)

[**2.** **Bài toán thực tế và mô hình sử dụng** 5](#_Toc163915664)

[**2.1.** **Bài toán thực tế** 5](#_Toc163915665)

[**2.2.** **Thách thức của bài toán** 5](#_Toc163915666)

[**2.3.** **Cách thức triển khai và mô hình sử dụng** 6](#_Toc163915667)

[**3.** **Mô tả thuật toán và áp dụng mô hình vào bài toán** 6](#_Toc163915668)

[**3.1.** **Giới thiệu về thuật toán** 6](#_Toc163915669)

[**3.2.** **Cơ sở toán học** 7](#_Toc163915670)

[**3.3.** **Áp dụng mô hình vào bài toán** 9](#_Toc163915671)

[**4.** **Kết luận** 10](#_Toc163915672)

1. **Giới thiệu về đề tài nghiên cứu**
   1. **Tổng quan**

Trong bối cảnh tài chính, nơi mà sự biến động và không chắc chắn là điều bình thường, việc dự đoán giá cổ phiếu là một thách thức lớn và phức tạp. Trên thị trường chứng khoán Việt Nam, nơi mà hàng triệu nhà đầu tư tham gia mỗi ngày, khả năng dự đoán xu hướng giá cổ phiếu có thể mang lại lợi ích vô cùng lớn cho các nhà đầu tư và quản lý danh mục đầu tư. Bài toán "Dự đoán giá cổ phiếu dựa trên dữ liệu giao dịch cổ phiếu trên các sàn giao dịch chứng khoán tại Việt Nam" đã được xác định với một mục tiêu vô cùng quan trọng: tạo ra các mô hình dự đoán hiệu quả, từ đó giúp nhà đầu tư đưa ra các quyết định đầu tư thông minh và hiệu quả.

Nghiên cứu này tập trung vào việc áp dụng các phương pháp và công nghệ tiên tiến trong lĩnh vực khoa học dữ liệu và học máy để phân tích và dự đoán xu hướng giá cổ phiếu trên thị trường chứng khoán Việt Nam. Mục tiêu cuối cùng là cung cấp sự hỗ trợ cho các nhà đầu tư, giúp họ có thể đưa ra các quyết định đầu tư một cách thông minh và hiệu quả. Điều này không chỉ đòi hỏi sự hiểu biết sâu sắc về thị trường chứng khoán và các yếu tố ảnh hưởng đến giá cổ phiếu, mà còn đòi hỏi sự áp dụng linh hoạt và hiệu quả của các phương pháp và công nghệ trong lĩnh vực khoa học dữ liệu và học máy.

* 1. **Đặc điểm của bài toán**
* Dữ liệu đa dạng và lớn: Dữ liệu giao dịch cổ phiếu thường bao gồm các thông tin như giá mở cửa, giá đóng cửa, khối lượng giao dịch, biến động giá, chỉ số kỹ thuật và các yếu tố tài chính khác. Dữ liệu này có thể rất lớn và phức tạp do số lượng cổ phiếu, thời gian giao dịch và các biến thay đổi liên tục.
* Biến động và không chắc chắn: Thị trường chứng khoán là một môi trường biến động và không chắc chắn, với sự ảnh hưởng của nhiều yếu tố như tin tức kinh tế, sự kiện chính trị, tâm lý đám đông và các yếu tố vĩ mô khác. Điều này tạo ra một môi trường dự đoán không chắc chắn và rủi ro.
* Phức tạp và đa chiều: Dự đoán giá cổ phiếu không chỉ phụ thuộc vào dữ liệu lịch sử mà còn phải tính đến các yếu tố kỹ thuật, cơ bản và tâm lý thị trường.
  1. **Hướng giải quyết**
* Xử lý và phân tích dữ liệu: Trong giai đoạn này, chúng ta tiến hành thu thập dữ liệu giao dịch cổ phiếu từ các sàn giao dịch chứng khoán tại Việt Nam và thực hiện quá trình xử lý và phân tích dữ liệu một cách kỹ lưỡng. Bằng cách này, chúng ta mong muốn đạt được một cái nhìn toàn diện và sâu sắc hơn về các yếu tố ảnh hưởng đến giá cổ phiếu, bao gồm cả yếu tố nội tại và ngoại tại của doanh nghiệp, thông tin thị trường và các biến động kinh tế xã hội.
* Xây dựng mô hình dự đoán: Sau khi hoàn thành giai đoạn xử lý và phân tích dữ liệu, chúng ta chuyển sang việc xây dựng các mô hình dự đoán. Đây là quá trình kỹ thuật mà chúng ta áp dụng các phương pháp và thuật toán trong lĩnh vực học máy và dự đoán thống kê để tạo ra các mô hình dự đoán giá cổ phiếu. Bằng cách này, chúng ta hy vọng có thể tạo ra những dự đoán chính xác và tin cậy về xu hướng giá cổ phiếu trong tương lai.
* Đánh giá và tinh chỉnh mô hình: Cuối cùng, sau khi xây dựng các mô hình dự đoán, chúng ta thực hiện việc đánh giá hiệu suất của chúng và tiến hành các biện pháp tinh chỉnh để nâng cao tính chính xác và độ tin cậy của dự đoán. Điều này bao gồm việc sử dụng các phương pháp đánh giá hiệu suất như Mean Squared Error (MSE), Root Mean Squared Error (RMSE) để đánh giá và điều chỉnh mô hình sao cho phù hợp nhất với dữ liệu và mục tiêu của chúng ta.
  1. **Kết quả dự kiến**

Kết quả của nghiên cứu không chỉ là sự dự đoán về xu hướng giá cổ phiếu mà còn là sự mở rộng về sự hiểu biết và sự thông minh trong quyết định đầu tư. Chúng tôi hy vọng rằng thông điệp từ nghiên cứu này sẽ thúc đẩy sự tự tin và sự hiểu biết sâu sắc hơn về thị trường chứng khoán, giúp nhà đầu tư và quản lý danh mục đầu tư điều chỉnh và tối ưu hóa chiến lược đầu tư của mình. Tuy nhiên, chúng tôi luôn nhắc nhở rằng thị trường chứng khoán là một môi trường phức tạp và dự đoán giá cổ phiếu vẫn mang tính chất dự đoán và rủi ro.

1. **Bài toán thực tế và mô hình sử dụng**
   1. **Bài toán thực tế**

Trong ngữ cảnh này, nghiên cứu đặt ra một bài toán cụ thể nhằm sử dụng dữ liệu giao dịch cổ phiếu từ các sàn giao dịch chứng khoán tại Việt Nam để tiên đoán giá cổ phiếu của Tập đoàn Hòa Phát (HPG) trong tương lai. Bài toán này đòi hỏi sự kết hợp hài hòa giữa các phương pháp trong lĩnh vực khoa học dữ liệu và kiến thức sâu sắc về cơ cấu và hoạt động của thị trường chứng khoán, nhằm xây dựng các mô hình dự đoán mạnh mẽ và hiệu quả. Điều này không chỉ là quá trình áp dụng các công cụ và kỹ thuật phức tạp trong khoa học dữ liệu, mà còn đòi hỏi hiểu biết sâu rộng về các yếu tố ảnh hưởng đến giá cổ phiếu, như tình hình kinh tế, chính trị, và các yếu tố vĩ mô khác, để đảm bảo tính chính xác và tin cậy của dự đoán.

* 1. **Thách thức của bài toán**

Trong việc áp dụng khoa học dữ liệu và phân tích dữ liệu giao dịch cổ phiếu để dự đoán giá cổ phiếu của Tập đoàn Hòa Phát (mã chứng khoán: HPG) trên thị trường chứng khoán Việt Nam, bài toán đặt ra nhiều thách thức khó khăn và phức tạp.

* Dữ liệu phức tạp và đa dạng: Dữ liệu giao dịch cổ phiếu trên sàn chứng khoán Việt Nam bao gồm hàng loạt thông tin như giá mở cửa, giá đóng cửa, khối lượng giao dịch, chỉ số kỹ thuật, và tin tức thị trường. Sự phức tạp và đa dạng của dữ liệu đòi hỏi sự chuyên sâu trong việc xử lý và phân tích.
* Biến động và sự không chắc chắn của thị trường: Thị trường chứng khoán luôn biến động và không chắc chắn, với sự ảnh hưởng của nhiều yếu tố bên ngoài như tin tức kinh tế, sự kiện quốc tế và tâm lý thị trường. Điều này tạo ra một môi trường dự đoán phức tạp và thách thức.
* Định hình mô hình phù hợp: Một thách thức lớn khác là lựa chọn và định hình mô hình phù hợp nhất cho bài toán. Việc này đòi hỏi sự hiểu biết sâu sắc về các phương pháp và thuật toán trong lĩnh vực học máy và dự đoán thống kê, cũng như khả năng áp dụng chúng một cách linh hoạt và hiệu quả.
* Yêu cầu về độ chính xác và tin cậy: Các nhà đầu tư và quản lý danh mục đầu tư đặt kỳ vọng vào các dự đoán chính xác và tin cậy về giá cổ phiếu. Điều này đòi hỏi các mô hình dự đoán phải được xây dựng và đánh giá một cách cẩn thận và kỹ lưỡng.
  1. **Cách thức triển khai và mô hình sử dụng**

Để tiếp cận và giải quyết bài toán dự đoán giá cổ phiếu dựa trên dữ liệu giao dịch trên các sàn chứng khoán tại Việt Nam một cách chính xác và hiệu quả, chúng tôi áp dụng một phương pháp tiên tiến trong lĩnh vực dự đoán chuỗi thời gian, được biết đến với tên gọi ARIMA Auto. Điều này đề cập đến việc sử dụng một biến thể tự động của mô hình ARIMA để tự động chọn ra các tham số phù hợp nhất cho mô hình.

Sau khi đã xác định được mô hình tối ưu từ ARIMA Auto, chúng tôi tiếp tục áp dụng mô hình ARIMA truyền thống để huấn luyện trên dữ liệu lịch sử và dự đoán giá cổ phiếu cho các khoảng thời gian trong tương lai.

Các bước thực hiện:

* Tải dữ liệu: Tải dữ liệu giao dịch cổ phiếu lên hệ thống
* Tiền xử lý dữ liệu: Loại bỏ các giá trị bị thiếu và chia dữ liệu ra thành tập dữ liệu huấn luyện và tập dữ liệu kiếm tra
* Lựa chọn và huấn luyện mô hình dự đoán: dùng mô hình ARIMA Auto để tìm ra mô hình tốt nhất và dùng ARIMA để huấn luyện và dự đoán mô hình.
* Đánh giá mô hình: Đánh giá hiệu suất của mô hình bằng Mean Squared Error (MSE), Root Mean Squared Error (RMSE).
* Dự đoán cổ phiếu cho 30 ngày tiếp theo: Cuối cùng, chúng ta sẽ sử dụng mô hình ARIMA đã huấn luyện để dự đoán giá cổ phiếu của Hòa Phát cho 30 ngày tiếp theo.

1. **Mô tả thuật toán và áp dụng mô hình vào bài toán**
   1. **Giới thiệu về thuật toán**

ARIMA được phát triển vào những năm 1970 bởi George Box và Gwilym Jenkins, và phát triển tiếp bởi người khác như James Durbin và David Brillinger. Ý tưởng ban đầu của ARIMA xuất phát từ ARMA (AutoRegressive Moving Average), một mô hình được sử dụng trong phân tích dự báo chuỗi thời gian. Box và Jenkins đã mở rộng ARMA bằng cách kết hợp nó với phương pháp tích phân, tạo ra mô hình ARIMA.

Mục đích chính của ARIMA là mô hình hóa mối quan hệ phức tạp giữa các quan sát liên tiếp trong chuỗi thời gian và sử dụng mô hình này để dự đoán các giá trị tương lai dựa trên quan sát hiện tại và quá khứ. ARIMA đã trở thành một trong những phương pháp phổ biến và mạnh mẽ nhất cho việc dự đoán chuỗi thời gian trong nhiều lĩnh vực, bao gồm tài chính, kinh tế, và dự báo thời tiết..

Mô hình trên là sự kết hợp của ba thành phần:

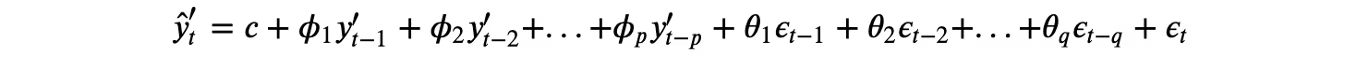
* AR (Auto Regressive): mô hình cố gắng dự đoán các giá trị tương lai dựa trên các giá trị trong quá khứ. Mô hình AR yêu cầu chuỗi thời gian phải là ổn định.
* I (Integrated): phần này biểu diễn việc làm cho chuỗi dữ liệu trở nên ổn định bằng cách chuyển đổi dữ liệu ban đầu thành dữ liệu dễ dự đoán hơn. Điều này thường đạt được bằng cách lấy sai phạm giữa các giá trị liên tiếp.
* MA (Moving Average): mô hình cố gắng dự đoán các giá trị tương lai dựa trên các sai số dự đoán trong quá khứ. Mô hình MA giả định rằng một mô hình tự hồi quy có thể xấp xỉ chuỗi thời gian đã cho. Điều này không phải là mô hình trung bình di động, mà là một quy trình làm mịn thay vì một mô hình dự đoán.
  1. **Cơ sở toán học**

Mỗi thành phần trong ARIMA hoạt động như một tham số có ký hiệu chuẩn. Đối với các mô hình ARIMA, ký hiệu tiêu chuẩn sẽ là ARIMA với p, d và q, trong đó các giá trị nguyên thay thế cho các tham số để chỉ ra loại mô hình ARIMA được sử dụng. Các tham số có thể được định nghĩa là:

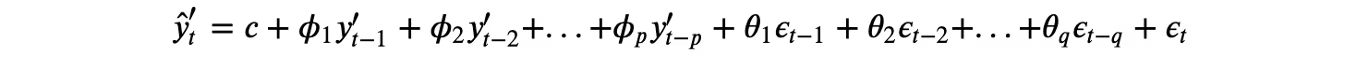
* Mô hình Auto Regressive AR(p) là một mô hình hồi quy với các giá trị trễ của y cho đến thời điểm thứ p trong quá khứ được sử dụng làm dữ liệu dự báo. Ở đây, p là số lượng quan sát trễ trong mô hình, ε là nhiễu trắng tại thời điểm t, c là một hằng số và φs là các tham số.



* Mô hình Integrated I(d) là sự khác biệt được thực hiện d lần cho đến khi chuỗi gốc trở thành ổn định. Một chuỗi thời gian ổn định là chuỗi mà các tính chất của nó không phụ thuộc vào thời điểm mà chuỗi được quan sát.
* Mô hình Moving Average MA(q) - Một mô hình trung bình động sử dụng một mô hình tương tự như mô hình hồi quy trên các sai số dự đoán trong quá khứ. Ở đây, ε là nhiễu trắng tại thời điểm t, c là một hằng số, và θs là các tham số.



Kết hợp tất cả ba loại mô hình trên cho ra mô hình ARIMA(p,d,q) hoàn chỉnh:



Phương pháp ARIMA là một phương pháp thống kê để phân tích và xây dựng một mô hình dự báo mà đại diện tốt nhất cho một chuỗi thời gian bằng cách mô hình hóa các tương quan trong dữ liệu. Nhờ vào các phương pháp thống kê một cách tuyệt đối, các mô hình ARIMA chỉ cần có dữ liệu lịch sử của một chuỗi thời gian để tổng quát hóa dự báo và đảm bảo tăng độ chính xác của dự báo trong khi vẫn giữ cho mô hình đơn giản.

Tuy nhiên, mặc dù có tính đơn giản, có nhiều điểm hạn chế tiềm ẩn khi sử dụng các mô hình ARIMA. Một điểm quan trọng nhất là sự chủ quan liên quan đến việc xác định các tham số p và q. Mặc dù các tự tương quan và tự tương quan một phần được sử dụng, việc chọn lựa các tham số p và q phụ thuộc vào kỹ năng và kinh nghiệm của người phát triển mô hình. Ngoài ra, so với phương pháp smoothing mũi tên đơn giản và phương pháp Holt Winters, các mô hình ARIMA phức tạp hơn và do đó, có ít sức mạnh giải thích hơn.

Tương tự như tất cả các phương pháp dự báo khác, bằng cách nhìn lại quá khứ, các mô hình ARIMA không tốt cho dự báo dài hạn và khá tệ trong việc dự đoán các điểm chuyển động. Chúng cũng có thể tốn nhiều chi phí tính toán.

Do đó, các mô hình ARIMA có thể dễ dàng và chính xác được sử dụng cho dự báo ngắn hạn chỉ bằng dữ liệu chuỗi thời gian, nhưng có thể mất một số kinh nghiệm và thử nghiệm để tìm ra một tập hợp tham số tối ưu cho mỗi trường hợp sử dụng.

* 1. **Áp dụng mô hình vào bài toán**

Các bước thực hiện chi tiết

* Tải dữ liệu: Trong quá trình này, chúng tôi tiến hành tải dữ liệu giao dịch cổ phiếu của Tập đoàn Hòa Phát từ nguồn dữ liệu đáng tin cậy bằng việc sử dụng thư viện vnstock. Dữ liệu này bao gồm một loạt các thông tin quan trọng như giá mở cửa, giá đóng cửa, khối lượng giao dịch và các chỉ số kỹ thuật khác, được sử dụng để nghiên cứu và phân tích động thái của thị trường chứng khoán.
* Thông tin chi tiết và thống kê dữ liệu: Sau khi dữ liệu đã được tải xuống, quá trình này sẽ tiếp tục với việc thực hiện phân tích thống kê cơ bản nhằm khám phá và hiểu rõ hơn về đặc điểm và tính chất của chuỗi thời gian. Phân tích này sẽ bao gồm việc đánh giá trung bình, phương sai, phân phối và các đặc điểm khác của dữ liệu, nhằm cung cấp cái nhìn toàn diện về động thái của thị trường chứng khoán.
* Trực quan sự biến động của cổ phiếu HPG: Bằng cách sử dụng công cụ trực quan như biểu đồ và đồ thị, quá trình này sẽ trực quan hóa sự biến động của giá cổ phiếu Hòa Phát (HPG) theo thời gian, nhằm phân tích và hiểu rõ hơn về xu hướng và cấu trúc của dữ liệu giao dịch
* Tiền xử lý dữ liệu: Trước khi tiến hành triển khai mô hình dự đoán, quá trình tiền xử lý dữ liệu là bước không thể thiếu, bao gồm loại bỏ các giá trị thiếu và phân chia tập dữ liệu thành hai phần: tập dữ liệu huấn luyện và tập dữ liệu kiểm tra. Điều này giúp đảm bảo tính toàn vẹn và độ tin cậy của dữ liệu được sử dụng trong quá trình huấn luyện và kiểm tra mô hình.
* Lựa chọn và huấn luyện mô hình dự đoán: Trong giai đoạn này, quá trình lựa chọn và huấn luyện mô hình dự đoán được thực hiện bằng cách sử dụng phương pháp ARIMA Auto để xác định mô hình tối ưu nhất, sau đó áp dụng mô hình ARIMA để thực hiện quá trình huấn luyện và dự đoán.
* So sánh giá trị dự đoán và giá trị thật của dữ liệu: Sau khi huấn luyện mô hình, chúng ta sẽ đánh giá hiệu suất của mô hình bằng cách so sánh giá trị dự đoán với giá trị thực tế của dữ liệu. Điều này giúp chúng ta đánh giá được độ chính xác và tin cậy của mô hình.
* Đánh giá mô hình: Chúng ta sẽ đánh giá hiệu suất của mô hình bằng các phương pháp đánh giá như Mean Squared Error (MSE), Root Mean Squared Error (RMSE).
* Dự đoán cổ phiếu cho 30 ngày tiếp theo: Cuối cùng, chúng ta sẽ sử dụng mô hình ARIMA đã huấn luyện để dự đoán giá cổ phiếu của Hòa Phát cho 30 ngày tiếp theo. Dự đoán này cung cấp thông tin hữu ích cho các nhà đầu tư và quản lý danh mục đầu tư để họ có thể đưa ra quyết định đầu tư thông minh.

1. **Kết luận**

Trong nghiên cứu này, chúng tôi đã tiến hành dự đoán giá cổ phiếu của tập đoàn Hòa Phát (mã chứng khoán: HPG) dựa trên dữ liệu giao dịch từ các sàn giao dịch chứng khoán tại Việt Nam, sử dụng mô hình Auto ARIMA để chọn mô hình tốt nhất và sau đó huấn luyện và đánh giá mô hình bằng mô hình ARIMA. Cuối cùng, chúng tôi đã sử dụng mô hình ARIMA để dự đoán giá cổ phiếu cho 30 ngày tiếp theo.

Kết quả của nghiên cứu này đã cung cấp cái nhìn sâu sắc vào khả năng dự đoán giá cổ phiếu trên thị trường chứng khoán Việt Nam bằng cách sử dụng mô hình ARIMA. Việc sử dụng mô hình ARIMA Auto đã giúp chúng tôi chọn ra một mô hình tối ưu dựa trên các tiêu chí như độ chính xác và hiệu suất. Sau đó, việc huấn luyện và đánh giá mô hình ARIMA đã cho thấy khả năng dự đoán tương đối tốt của mô hình trên dữ liệu giao dịch thực tế.

Dự đoán cho 30 ngày tiếp theo từ mô hình ARIMA cung cấp thông tin quan trọng cho các nhà đầu tư và quản lý danh mục đầu tư để họ có thể đưa ra quyết định đầu tư thông minh và hiệu quả trong tương lai. Tuy nhiên, cần lưu ý rằng thị trường chứng khoán là một môi trường phức tạp và dự đoán giá cổ phiếu vẫn mang tính chất dự đoán và rủi ro.

Tổng quan, nghiên cứu này đã đề xuất một phương pháp hữu ích và tiên tiến để dự đoán giá cổ phiếu trên thị trường chứng khoán Việt Nam, đồng thời cung cấp cái nhìn sâu sắc về khả năng và hạn chế của mô hình ARIMA trong việc này. Đây có thể là một công cụ hữu ích cho các nhà đầu tư và quản lý danh mục đầu tư để họ có thể đưa ra các quyết định đầu tư có kiến thức và đào tạo.