

UBND THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN



**PHÂN TÍCH DỮ LIỆU**

**ĐỀ TÀI: PHÂN TÍCH VÀ DỰ BÁO GIÁ CỔ PHIẾU VINAMILK VÀ HỖ TRỢ QUYẾT ĐỊNH ĐẦU TƯ ĐỂ TỐI ƯU LỢI NHUẬN**

**Giảng viên hướng dẫn:** Phan Thành Huấn

**Nhóm sinh viên thực hiện:** Đặng Văn Bảo 3119410024

Huỳnh Lệ San 3121410415

TP. HCM, ngày 16 tháng 12 năm 2024

**BẢNG PHÂN CHIA CÔNG VIỆC**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **MSSV** | **HỌ TÊN** | **CÔNG VIỆC** | **MỨC ĐỘ HOÀN THÀNH** |
| 1 | 3119410024 | Đặng Văn Bảo | Phân rã và phân tích các thành phần trong chuỗi thời gian.  Xây dựng mô hình LSTM để dự đoán giá, Random Forest, SHAP để hoàn thành bộ quy tắc hỗ trợ quyết định đầu tư.  Viết báo cáo.  (50%) | 100% |
| 2 | 3121410415 | Huỳnh Lệ San | Tiền xử lý và phân tích dữ liệu.  Xây dựng mô hình hồi quy tuyến tính và XGBoost để dự đoán giá.  Viết báo cáo.  (50%) | 100% |

**NHẬN XÉT CỦA GIẢNG VIÊN**

**LỜI CẢM ƠN**

Chúng em xin gửi lời cảm ơn chân thành đến Thầy Phan Thành Huấn - giảng viên bộ môn Phân tích dữ liệu, khoa Công nghệ thông tin, trường Đại học Sài Gòn, đã trang bị cho chúng em những kiến thức và kỹ năng cơ bản cần thiết để có thể hoàn thành đồ án này.

Tuy nhiên, với vốn kiến thức cũng như kinh nghiệm còn khiêm tốn và mặc dù nhóm đã cố gắng hoàn thành hoàn thành đồ án với tất cả nỗ lực của từng thành viên trong nhóm, nhưng đồ án này chắc chắn không tránh khỏi những thiếu sót nhất định, chúng em rất mong nhận được sự quan tâm, cảm thông, chia sẻ và góp ý chỉ bảo của Thầy để đồ án của chúng em được đầy đủ và hoàn chỉnh hơn.

Xin kính chúc Thầy dồi dào sức khỏe và hạnh phúc để tiếp tục thực hiện sứ mệnh cao đẹp của mình là truyền đạt kiến thức cho thế hệ mai sau.

Chúng em xin chân thành cảm ơn.

TP. HCM, ngày 16 tháng 12 năm 2024.

**MỤC LỤC**

[CHƯƠNG 1: GIỚI THIỆU 1](#_Toc186555009)

[1.1. Bối cảnh và lý do chọn đề tài 1](#_Toc186555010)

[1.2. Mục tiêu nghiên cứu 2](#_Toc186555011)

[1.3. Phạm vi nghiên cứu 2](#_Toc186555012)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 4](#_Toc186555013)

[2.1. Các khái niệm cơ bản về thị trường chứng khoán 4](#_Toc186555014)

[2.2. Các yếu tố ảnh hưởng đến giá đóng cửa 5](#_Toc186555015)

[2.3. Tổng quan về các mô hình dự đoán giá cổ phiếu 7](#_Toc186555016)

[2.3.1. Mô hình hồi quy tuyến tính 7](#_Toc186555017)

[2.3.2. Mô hình XGBoost 8](#_Toc186555018)

[2.3.3. Mô hình LSTM (Long Short-Term Memory) 9](#_Toc186555019)

[2.4. Các chỉ số đánh giá mô hình 11](#_Toc186555020)

[2.4.1. Sai số trung bình tuyệt đối (MAE) 11](#_Toc186555021)

[2.4.2. Sai số trung bình bình phương (MSE) 11](#_Toc186555022)

[2.4.3. Căn bậc hai của sai số trung bình bình phương (RMSE) 12](#_Toc186555023)

[2.4.4. Hệ số xác định (R2) 12](#_Toc186555024)

[2.4.5. Sai số phần trăm tuyệt đối trung bình (MAPE) 12](#_Toc186555025)

[CHƯƠNG 3: THU THẬP VÀ PHÂN TÍCH DỮ LIỆU 14](#_Toc186555026)

[3.1. Thu thập dữ liệu 14](#_Toc186555027)

[3.2. Mô tả dữ liệu 14](#_Toc186555028)

[3.2.1. Tổng số dòng và cột trong tập dữ liệu 14](#_Toc186555029)

[3.2.2. Xác định kiểu dữ liệu của mỗi cột 15](#_Toc186555030)

[3.2.3. Xử lý kiểu dữ liệu 16](#_Toc186555031)

[3.3. Sắp xếp dữ liệu theo thời gian 16](#_Toc186555032)

[3.4. Thêm vào cột Đóng cửa (t+1) (Giá đóng cửa ngày mai) 17](#_Toc186555033)

[3.5. Kiểm tra tính toàn vẹn của dữ liệu 17](#_Toc186555034)

[3.5.1. Kiểm tra các dòng trùng lặp 18](#_Toc186555035)

[3.5.2. Kiểm tra và xử lý dữ liệu thiếu 18](#_Toc186555036)

[3.5.3. Xác định giá trị ngoại lệ trong các cột (outliers detection) 19](#_Toc186555037)

[3.6. Kiểm tra tính ổn định của dữ liệu 23](#_Toc186555038)

[3.7. Trực quan hóa dữ liệu 24](#_Toc186555039)

[3.7.1. Kiểm tra phân phối của dữ liệu 24](#_Toc186555040)

[3.7.2. Khám phá mối quan hệ giữa các biến thông qua ma trận tương quan 26](#_Toc186555041)

[3.7.3. Kiểm tra xu hướng dữ liệu 27](#_Toc186555042)

[CHƯƠNG 4: XÂY DỰNG MÔ HÌNH VÀ KẾT QUẢ 30](#_Toc186555043)

[4.1. Mô hình hồi quy tuyến tính đa biến 30](#_Toc186555044)

[4.1.1. Chuẩn bị xây dựng mô hình 30](#_Toc186555045)

[4.1.2. Xây dựng mô hình 34](#_Toc186555046)

[4.2. Mô hình XGBoost 36](#_Toc186555047)

[4.2.1. Chuẩn bị xây dựng mô hình 36](#_Toc186555048)

[4.2.2. Xây dựng mô hình 38](#_Toc186555049)

[4.3. Mô hình LSTM 39](#_Toc186555050)

[4.3.1. Chuẩn bị xây dựng mô hình 39](#_Toc186555051)

[4.3.2. Xây dựng mô hình 40](#_Toc186555052)

[4.4. Mô hình đưa ra quyết định và trọng số của các đặc trưng trong từng quyết định mua, bán và giữ. 42](#_Toc186555053)

[4.4.1. Random Forest Classifier 42](#_Toc186555054)

[4.4.2. SHAP (SHapley Additive exPlanations) 43](#_Toc186555055)

[CHƯƠNG 5: ĐÁNH GIÁ MÔ HÌNH 45](#_Toc186555056)

[5.1. Mô hình Hồi quy tuyến tính 45](#_Toc186555057)

[5.2. Mô hình XGBOOST 46](#_Toc186555058)

[5.3. Mô hình LSTM 47](#_Toc186555059)

[5.4. Đánh giá và so sánh 3 mô hình 48](#_Toc186555060)

[CHƯƠNG 6: TRÍCH XUẤT THÔNG TIN TỪ DỮ LIỆU THỜI GIAN 49](#_Toc186555061)

[6.1. Chỉ số MA 50 và MA 200 49](#_Toc186555062)

[6.2. Chỉ số Valatility 30 và Volatility 90 49](#_Toc186555063)

[6.3. Time series Decomposition (Trend và Seasonal): 50](#_Toc186555064)

[CHƯƠNG 7: BỘ HỖ TRỢ QUYẾT ĐỊNH 52](#_Toc186555065)

[7.1. Quyết định mua (Mở vị thế mua) 52](#_Toc186555066)

[7.2. Quyết định bán (Vị thế hoặc chốt lời) 52](#_Toc186555067)

[7.3. Quyết định giữ (không bán, cũng không mua thêm) 53](#_Toc186555068)

[7.4. Trọng số của các thuộc tính trong các quyết định mua, bán và giữ 54](#_Toc186555069)

[7.4.1. Quyết định bán 54](#_Toc186555070)

[7.4.2. Quyết định giữ 55](#_Toc186555071)

[7.4.3. Quyết định bán 56](#_Toc186555072)

[CHƯƠNG 8: KẾT LUẬN 57](#_Toc186555073)

**DANH MỤC HÌNH ẢNH VÀ BẢNG BIỂU**

*Hình ành*

[Hình 1: Khái niệm cơ bản về thị trường chứng khoán 5](#_Toc186555080)

[Hình 2: Các thành phần chính của LSTM 10](#_Toc186555081)

[Hình 3: Các chỉ số đánh giá mô hình 13](#_Toc186555082)

[Hình 4: Lệnh đọc dữ liệu từ dataset 14](#_Toc186555083)

[Hình 5: Hình ảnh dữ liệu của dataset 15](#_Toc186555084)

[Hình 6: Kiểu dữ liệu của từng cột 15](#_Toc186555085)

[Hình 7: Code xử lý kiểu dữ liệu 16](#_Toc186555086)

[Hình 8: Kiểu dữ liệu của từng cột sau khi được xử lý 16](#_Toc186555087)

[Hình 9: Code sắp xếp dữ liệu tăng dần theo thời gian 16](#_Toc186555088)

[Hình 10: Dữ liệu sau khi được sắp xếp tăng dần theo thời gian 17](#_Toc186555089)

[Hình 11: Code thêm cột Giá đóng cửa ngày mai vào dữ liệu 17](#_Toc186555090)

[Hình 12: Số dòng trùng lặp trong dữ liệu 18](#_Toc186555091)

[Hình 13: Kết quả kiểm tra dữ liệu thiếu 18](#_Toc186555092)

[Hình 14: Các dòng có giá trị thiếu 19](#_Toc186555093)

[Hình 15: Số lượng giá trị thiếu ở từng cột sau khi được xử lý giá trị thiếu 19](#_Toc186555094)

[Hình 16: Biểu đồ boxplot 19](#_Toc186555095)

[Hình 17: Giá trị ngoại lệ của cột giá đóng cửa 21](#_Toc186555096)

[Hình 18: Code kiểm tra độ lệch và độ nhọn của các biến 22](#_Toc186555097)

[Hình 19: Code kiểm tra tính ổn định của cột Đóng cửa 23](#_Toc186555098)

[Hình 20: Code kiểm tra tính ổn định của cột giá đóng cửa ngày mai 23](#_Toc186555099)

[Hình 21: Biểu đồ histogram 24](#_Toc186555100)

[Hình 22: Biểu đồ scatter 25](#_Toc186555101)

[Hình 23: Ma trận tương quan giữa các biến 26](#_Toc186555102)

[Hình 24: Biểu đồ đường của giá đóng cửa theo thời gian 27](#_Toc186555103)

[Hình 25: Biểu đồ đường của giá đóng cửa với MA50 và MA200 28](#_Toc186555104)

[Hình 26: Code kiểm tra đa cộng tuyến 31](#_Toc186555105)

[Hình 27: Code xử lý đa cộng tuyến 32](#_Toc186555106)

[Hình 28: Kiểm tra sự ảnh hưởng của biến đầu vào với biến đầu ra 32](#_Toc186555107)

[Hình 29: Code chuẩn hóa dữ liệu cho HQTT 33](#_Toc186555108)

[Hình 30: Code chia dữ liệu của HQTT 34](#_Toc186555109)

[Hình 31: Code xây dựng mô hình HQTT 35](#_Toc186555110)

[Hình 32: Kết quả dự đoán giá đóng cửa ngày mai 35](#_Toc186555111)

[Hình 33: Biểu đồ so sánh giá dự đoán và giá thực tế của HQTT 35](#_Toc186555112)

[Hình 34: Code chuẩn hóa dữ liệu của XGBoost 36](#_Toc186555113)

[Hình 35: Code chia dữ liệu của XGBoost 37](#_Toc186555114)

[Hình 36: Code chọn các chỉ số phù hợp với XGBoost 37](#_Toc186555115)

[Hình 37: Code xây dựng mô hình XGBoost 38](#_Toc186555116)

[Hình 38: Kết quả dự đoán của XGBoost 39](#_Toc186555117)

[Hình 39: Biểu đồ so sánh giá dự đoán và giá thực tế của XGBoost 39](#_Toc186555118)

[Hình 40: Code xây dựng mô hình 41](#_Toc186555119)

[Hình 41: Cấu trúc mô hình LSTM 41](#_Toc186555120)

[Hình 42: Biểu đồ so sánh giá dự đoán và giá thực tế của LSTM 42](#_Toc186555121)

[Hình 43: Các chỉ số đánh giá của mô hình HQTT 45](#_Toc186555122)

[Hình 44: Các chỉ số đánh giá của mô hình XGBoost 46](#_Toc186555123)

[Hình 45: Các chỉ số đánh giả của mô hình LSTM 47](#_Toc186555124)

[Hình 46: Biểu đồ đường của giá dự báo so với MA50 dự báo và MA200 dự báo 49](#_Toc186555125)

[Hình 47: Biểu đồ đường thể hiện độ biến động 30 ngày và 90 ngày của giá dự báo 50](#_Toc186555126)

[Hình 48: Biểu đồ thể hiện Trend của giá dự báo 51](#_Toc186555127)

[Hình 49: Biểu đồ thể hiện Seasonal của giá dự báo 51](#_Toc186555128)

[Hình 50: Trọng số của các thuộc tính cho quyết định bán 54](#_Toc186555129)

[Hình 51: Trọng số của các thuộc tính cho quyết định giữ 55](#_Toc186555130)

[Hình 52: Trọng số của các thuộc tính cho quyết định mua 56](#_Toc186555131)

*Bảng biểu*

[Bảng 1: So sánh các chỉ số đánh giá của 3 mô hình 46](#_Toc185363185)

[Bảng 2: Quy tắc mua, bán và giữ 52](#_Toc185363186)

# CHƯƠNG 1: GIỚI THIỆU

## 1.1. Bối cảnh và lý do chọn đề tài

Thị trường chứng khoán đóng vai trò quan trọng, là huyết mạch của nền kinh tế, không chỉ cung cấp nguồn vốn cho doanh nghiệp mà còn mở ra cơ hội đầu tư cho các cá nhân và tổ chức. Với vai trò là nơi giao thoa giữa lý trí và cảm xúc, thị trường chứng khoán đòi hỏi nhà đầu tư cần sự kiên nhẫn và kỷ luật để đạt được thành công. Đặc biệt, sự phát triển nhanh chóng của thị trường chứng khoán Việt Nam trong những năm gần đây đã cho thấy nhu cầu phân tích và dự đoán chính xác giá cổ phiếu để hỗ trợ các quyết định đầu tư quan trọng.

Vinamilk, với vị thế là một trong những doanh nghiệp lớn nhất trên thị trường chứng khoán Việt Nam, đại diện cho sự phát triển bền vững của nền kinh tế quốc gia. Là một doanh nghiệp hàng đầu trong ngành công nghiệp sữa, cổ phiếu Vinamilk thu hút sự quan tâm lớn từ các nhà đầu tư trong và ngoài nước. Việc nghiên cứu và phân tích các doanh nghiệp lớn như Vinamilk không chỉ cung cấp cơ hội đầu tư tiềm năng mà còn giúp hiểu rõ hơn về nền kinh tế và thị trường.

Dự đoán giá cổ phiếu giúp nhà đầu tư đưa ra quyết định mua bán hợp lý, tối ưu hóa lợi nhuận và giảm thiểu rủi ro, đặc biệt trong một môi trường đầy biến động như thị trường tài chính. Peter Lynch trong cuốn "One Up On Wall Street" đã nhấn mạnh: "Đầu tư mà không nghiên cứu giống như chơi bài mà không nhìn vào quân bài". Các mô hình dự đoán giá cổ phiếu sử dụng dữ liệu lịch sử và các yếu tố kinh tế để dự báo xu hướng tương lai, giúp nhà đầu tư có cái nhìn tổng quan và chiến lược đầu tư hiệu quả hơn.

Dự đoán giá cổ phiếu từ lâu đã là một bài toán phức tạp do chịu ảnh hưởng từ nhiều yếu tố như kinh tế vĩ mô, hiệu suất doanh nghiệp, tâm lý thị trường,... và đòi hỏi sự kết hợp của nhiều phương pháp phân tích kỹ thuật. Phân tích kỹ thuật là nghệ thuật và khoa học của việc dự đoán xu hướng giá dựa trên dữ liệu thị trường quá khứ. Sử dụng các công cụ và mô hình học máy (machine learning) có thể cải thiện độ chính xác của dự đoán. Việc nghiên cứu và phát triển các mô hình dự đoán không chỉ có ý nghĩa thực tiễn mà còn đóng góp vào sự phát triển của khoa học dữ liệu và kinh tế tài chính.

## 1.2. Mục tiêu nghiên cứu

* **Phân tích dữ liệu giá cổ phiếu Vinamilk (VNM):**
* Thu thập và xử lý dữ liệu giá cổ phiếu VNM từ các nguồn đáng tin cậy trong giai đoạn 2014-2024.
* Nghiên cứu các đặc điểm và xu hướng biến động giá cổ phiếu, bao gồm phân tích các yếu tố ảnh hưởng như xu hướng, mùa vụ và các sự kiện kinh tế.
* **Ứng dụng các phương pháp dự báo giá cổ phiếu:**
* Triển khai các mô hình dự báo truyền thống như hồi quy tuyến tính.
* Ứng dụng các mô hình hiện đại như XGBoost (phân tích phi tuyến) và LSTM (mô hình chuỗi thời gian dựa trên mạng nơ-ron).
* So sánh độ chính xác, hiệu quả và khả năng áp dụng của từng phương pháp trên dữ liệu thực tế.
* **Đánh giá và lựa chọn mô hình tối ưu:**
* Xây dựng các tiêu chí đánh giá hiệu quả của mô hình, như RMSE, MAE, và R-squared.
* Phân tích ưu nhược điểm của từng mô hình dựa trên kết quả dự báo.
* **Hỗ trợ ra quyết định đầu tư:**
* Ứng dụng mô hình dự báo để đề xuất các chiến lược đầu tư tối ưu cho cổ phiếu VNM.
* Đề xuất bộ quy tắc đầu tư để tối ưu hóa lợi nhuận.

## 1.3. Phạm vi nghiên cứu

Phạm vi nghiên cứu tập trung vào việc ứng dụng các mô hình học máy vào bài toán dự đoán giá cổ phiếu. Nghiên cứu và đánh giá hiệu quả dự đoán của các mô hình.

Cụ thể, phạm vi nghiên cứu bao gồm:

* **Mô hình dự đoán:** Nghiên cứu và áp dụng hai mô hình chính là hồi quy tuyến tính, XGBoost và LSTM nhằm phân tích và dự đoán giá cổ phiếu của Vinamilk trong các khoảng thời gian nhất định (từ năm 2014 đến năm 2024).
* **Tập dữ liệu**: Sử dụng tập dữ liệu lịch sử giá cổ phiếu của Vinamilk, bao gồm các thông tin quan trọng như giá mở cửa, giá đóng cửa, giá cao nhất, giá thấp nhất, khối lượng giao dịch và tỉ lệ thay đổi giá. Dữ liệu sẽ được tiền xử lý để đảm bảo chất lượng và phù hợp với mục tiêu nghiên cứu.
* **Đánh giá mô hình**: Sử dụng các chỉ số đánh giá như MSE, MAE, RMSE và R² để đo lường hiệu suất của từng mô hình trong việc dự đoán giá cổ phiếu, từ đó so sánh và xác định mô hình tốt nhất cho bài toán này.

Phạm vi nghiên cứu được giới hạn ở việc dự đoán giá cổ phiếu trong một khoảng thời gian cụ thể, với mục tiêu phân tích và so sánh hiệu quả của các mô hình hồi quy tuyến tính và XGBoost trong bối cảnh dự đoán giá cổ phiếu của Vinamilk.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## 2.1. Các khái niệm cơ bản về thị trường chứng khoán

* Giá mở cửa (Open Price): Là mức giá giao dịch đầu tiên của một cổ phiếu trong phiên giao dịch. Giá này thường được xác định dựa trên lệnh mua/bán và cung cầu trước khi thị trường chính thức mở cửa.
* Giá đóng cửa (Close Price): Là mức giá giao dịch cuối cùng của một cổ phiếu trong phiên giao dịch. Đây là một chỉ số quan trọng, thường được sử dụng để phân tích xu hướng thị trường và đưa ra các quyết định đầu tư.
* Giá cao nhất (High Price): Là mức giá cao nhất mà một cổ phiếu đạt được trong phiên giao dịch. Giá này thể hiện mức độ kỳ vọng cao nhất của thị trường trong ngày.
* Giá thấp nhất (Low Price): Là mức giá thấp nhất mà một cổ phiếu đạt được trong phiên giao dịch. Giá này phản ánh mức độ lo ngại hoặc áp lực bán cao nhất trong ngày.
* % Thay đổi (Percent Change): Là tỷ lệ phần trăm thay đổi của giá cổ phiếu giữa hai phiên giao dịch liên tiếp hoặc giữa giá mở cửa và giá đóng cửa. Chỉ số này thường được sử dụng để đo lường sự biến động giá.
* Khối lượng giao dịch (Volume): Là tổng số lượng cổ phiếu được mua hoặc bán trong một phiên giao dịch. Đây là chỉ số quan trọng để đánh giá sức mạnh của một xu hướng giá, cho thấy mức độ quan tâm của nhà đầu tư đối với cổ phiếu.
* MA50: Trung bình giá của cổ phiếu trong 50 ngày giao dịch gần nhất.
* MA200: Trung bình giá của cổ phiếu trong 200 ngày giao dịch gần nhất.
* Volatility 30: Độ biến động của giá cổ phiếu trong 30 ngày gần nhất.
* Volatility 90: Độ biến động của giá cổ phiếu trong 90 ngày gần nhất.
* Time series decomposition (Phân rã chuỗi thời gian): Là phương pháp phân tích chuỗi thời gian bằng cách tách nó thành các thành phần cơ bản:
* Trend (Xu hướng):
  + Là hướng đi dài hạn của dữ liệu (tăng, giảm hoặc ổn định).
  + Ví dụ: Giá cổ phiếu tăng dần qua các năm.
* Seasonal (Mùa vụ):
  + Là các biến động có tính lặp lại theo chu kỳ, chẳng hạn như xu hướng giá tăng trong mùa lễ hội hoặc giảm vào những ngày thấp điểm.
* Residual (Ngẫu nhiên):
  + Là phần dữ liệu không thể giải thích bằng xu hướng hoặc mùa vụ, thường là các yếu tố ngẫu nhiên hoặc bất thường.

A diagram of a hexagon with colorful text

Description automatically generated

Hình 1: Khái niệm cơ bản về thị trường chứng khoán

## 2.2. Các yếu tố ảnh hưởng đến giá đóng cửa

Giá đóng cửa của cổ phiếu là kết quả tổng hợp của nhiều yếu tố từ trong và ngoài thị trường, và các yếu tố này tương tác chặt chẽ với dữ liệu giá lịch sử để định hình xu hướng giá. Các yếu tố có thể được chia thành các nhóm chính sau:

* Yếu tố nội tại doanh nghiệp:
* Hiệu quả kinh doanh: Các báo cáo tài chính như doanh thu, lợi nhuận, và thu nhập trên mỗi cổ phiếu là những chỉ số quan trọng từ dữ liệu giá lịch sử giúp xác định sức khỏe tài chính của doanh nghiệp và triển vọng tăng trưởng.
* Cơ cấu vốn và cổ tức: Thông qua dữ liệu giá cổ phiếu trong quá khứ, có thể thấy được các dấu hiệu về các quyết định tài chính của doanh nghiệp như chia tách cổ phiếu, phát hành thêm hoặc trả cổ tức, từ đó ảnh hưởng đến giá cổ phiếu hiện tại.
* Thay đổi trong ban lãnh đạo: Biến động của giá cổ phiếu trước và sau các thay đổi trong ban lãnh đạo thường phản ánh kỳ vọng thị trường và dữ liệu giá lịch sử là cơ sở để đánh giá sự tác động này.
* Yếu tố thị trường:
* Cung và cầu cổ phiếu: Biến động của giá đóng cửa chịu ảnh hưởng mạnh mẽ từ cung cầu trên thị trường, mà dữ liệu giá lịch sử giúp nhận diện rõ ràng các xu hướng tăng hoặc giảm, từ đó cho phép dự đoán những thay đổi tiếp theo.
* Thanh khoản thị trường: Mức độ thanh khoản, được đo lường qua dữ liệu khối lượng giao dịch cổ phiếu, giúp xác định khả năng chuyển đổi cổ phiếu thành tiền mặt và tác động đến giá đóng cửa.
* Chỉ số thị trường: Xu hướng của chỉ số thị trường như VN-Index có thể được dự đoán thông qua việc phân tích dữ liệu giá lịch sử, từ đó ảnh hưởng đến các cổ phiếu thành phần khác.
* Yếu tố kinh tế vĩ mô:
* Lãi suất: Mức lãi suất được phản ánh qua dữ liệu lãi suất trong quá khứ có thể ảnh hưởng đến chi phí tài chính của các doanh nghiệp và tác động trực tiếp đến giá cổ phiếu.
* Lạm phát: Dữ liệu lạm phát trước đây giúp dự đoán áp lực chi phí đối với các công ty và giá cả thị trường, từ đó ảnh hưởng đến giá cổ phiếu.
* Tăng trưởng kinh tế: Sự tăng trưởng của nền kinh tế được thể hiện qua dữ liệu lịch sử có thể dự đoán các xu hướng giá cổ phiếu, đặc biệt là khi nền kinh tế phát triển ổn định.
* Yếu tố chính trị và pháp luật:
* Chính sách nhà nước: Các chính sách kinh tế, thuế, hoặc hỗ trợ cho ngành nghề có thể được phản ánh qua dữ liệu giá lịch sử, cho phép dự đoán ảnh hưởng tiềm năng đến giá cổ phiếu.
* Tình hình chính trị: Những biến động chính trị lớn, được thể hiện qua dữ liệu giá lịch sử, thường gây ra các phản ứng tâm lý mạnh mẽ trên thị trường và ảnh hưởng đến giá cổ phiếu.
* Quy định pháp luật: Thay đổi quy định pháp lý, thể hiện qua dữ liệu lịch sử, có thể làm thay đổi tâm lý nhà đầu tư và tác động đến giá cổ phiếu.
* Yếu tố tâm lý và bất ngờ:
* Tâm lý nhà đầu tư: Sự kỳ vọng của nhà đầu tư, được phản ánh qua dữ liệu giá cổ phiếu trước đó, ảnh hưởng lớn đến quyết định mua bán cổ phiếu.
* Tin tức và sự kiện: Các sự kiện bất ngờ, như thiên tai, dịch bệnh, hoặc các vụ bê bối tài chính, có thể gây biến động lớn và được phản ánh qua dữ liệu giá lịch sử của cổ phiếu.

Giá đóng cửa của cổ phiếu không chỉ phụ thuộc vào dữ liệu giá lịch sử mà còn bị tác động bởi một loạt các yếu tố nội tại doanh nghiệp, thị trường, vĩ mô, chính trị, và tâm lý. Để dự đoán giá cổ phiếu một cách chính xác, cần kết hợp dữ liệu lịch sử với các yếu tố khác và đánh giá toàn diện tình hình hiện tại.

## 2.3. Tổng quan về các mô hình dự đoán giá cổ phiếu

Dự đoán giá cổ phiếu là một bài toán phức tạp và đa chiều, đòi hỏi sự kết hợp của nhiều phương pháp và kỹ thuật khác nhau để phân tích và dự đoán giá trong tương lai. Dưới đây là tổng quan về các mô hình phổ biến được sử dụng trong việc dự đoán giá cổ phiếu:

### 2.3.1. Mô hình hồi quy tuyến tính

Hồi quy tuyến tính là một trong những mô hình cơ bản nhất trong dự đoán giá cổ phiếu. Mô hình này sử dụng một phương trình tuyến tính để thiết lập mối quan hệ giữa biến phụ thuộc (giá cổ phiếu) và các biến độc lập (như giá mở cửa, giá cao nhất, giá thấp nhất và khối lượng giao dịch).

Hồi quy tuyến tính đơn biến: Một biến phụ thuộc được dự đoán trên một biến độc lập (phương pháp đơn giản nhất trong hồi quy dữ liệu).

Trong đó, là biến phụ thuộc, là biến độc lập, là hệ số chặn, là hệ số hồi quy và là sai số.

Hồi quy tuyến tính bội: Một phần mở rộng của hồi quy tuyến tính đơn biến, trong đó biến phụ thuộc Y được dự đoán dựa trên nhiều biến độc lập

Trong đó, là biến phụ thuộc, là các biến độc lập, là hệ số chặn, là các hệ số hồi quy tương ứng với từng biến độc lập và là sai số.

Mô hình LR có ưu điểm là đơn giản, dễ hiểu và dễ áp dụng với các dữ liệu lịch sử. Nó cung cấp cái nhìn trực quan về mối quan hệ tuyến tính giữa các biến và giá cổ phiếu.

Nhược điểm của mô hình này là nó phụ thuộc mạnh vào giả định tuyến tính. Không phù hợp với các dữ liệu có mối quan hệ phi tuyến hoặc có xu hướng biến động cao, đặc trưng của thị trường chứng khoán. Đồng thời, dễ bị ảnh hưởng bởi các điểm dữ liệu ngoại lai (outliers).

### 2.3.2. Mô hình XGBoost

XGBoost (Extreme Gradient Boost) là một thuật toán học máy nâng cao dựa trên cây quyết định, sử dụng kỹ thuật boosting để kết hợp nhiều cây con thành một mô hình mạnh mẽ hơn. Trong dự đoán giá cổ phiếu, XGBoost khai thác dữ liệu lịch sử và các yếu tố tác động (như khối lượng giao dịch, biến động giá) để xây dựng các dự báo.

Ưu điểm của XGBoost là khả năng xử lý dữ liệu lớn và phức tạp, khả năng phát hiện mối quan hệ phi tuyến giữa các yếu tố. Mô hình cung cấp độ chính xác cao nhờ tối ưu hoá quá trình huấn luyện và giảm thiểu hiện tượng overfitting.

Tuy nhiên, mô hình này đòi hỏi thời gian huấn luyện lâu hơn so với các mô hình đơn giản như hồi quy tuyến tính. Quá trình điều chỉnh tham số phức tạp, yêu cầu kinh nghiệm và thời gian thử nghiệm. Ngoài ra, kết quả của mô hình XGBoost có thể khó giải thích so với các mô hình tuyến tính.

### 2.3.3. Mô hình LSTM (Long Short-Term Memory)

Mô hình LSTM (Long Short-Term Memory) là một biến thể đặc biệt của RNN (Recurrent Neural Network), được thiết kế để giải quyết vấn đề phụ thuộc dài hạn (long-term dependency) trong dữ liệu chuỗi thời gian. Khác với RNN truyền thống, LSTM sử dụng một cơ chế đặc biệt gồm các cổng (forget gate, input gate, output gate) và ô nhớ (memory cell) để chọn lọc thông tin cần lưu giữ hoặc loại bỏ qua các bước thời gian. Điều này giúp LSTM khắc phục các vấn đề thường gặp ở RNN như vanishing gradient, cho phép mô hình học được các mối quan hệ dài hạn trong dữ liệu. Một trong những ưu điểm nổi bật của LSTM là khả năng ghi nhớ dài hạn, giúp nó xử lý tốt các bài toán như dự báo chuỗi thời gian, dịch máy, nhận diện giọng nói và phân tích cảm xúc. Tuy nhiên, LSTM cũng có những nhược điểm như yêu cầu tài nguyên tính toán lớn do cấu trúc phức tạp, cần lượng dữ liệu lớn để tránh overfitting, và đòi hỏi nhiều thời gian tối ưu hóa siêu tham số. Dù vậy, với khả năng mạnh mẽ trong việc xử lý dữ liệu chuỗi, LSTM giúp xử lý hiệu quả hơn các bài toán liên quan đến chuỗi thời gian phức tạp như dự đoán giá cổ phiếu.

**Kiến trúc cơ bản của LSTM:**

Ở mỗi thời điểm (**state**) *t*, LSTM xử lý:

* **Input**:
  + : Dữ liệu đầu vào tại thời điểm *t* (ví dụ: giá cổ phiếu ngày thứ t).
  + : Hidden state từ thời điểm *t-1*, đại diện cho thông tin ngắn hạn.
  + : Cell state từ thời điểm *t−1*, đại diện cho thông tin dài hạn.
* **Output**:
  + : Hidden state ở thời điểm *t*, đóng vai trò lưu trữ thông tin để chuyển sang bước tiếp theo.
  + : Cell state ở thời điểm *t*, lưu trữ thông tin dài hạn.

**Các thành phần chính:**

A diagram of a flowchart

Description automatically generated

Hình 2: Các thành phần chính của LSTM

LSTM được cấu tạo từ nhiều "cổng" (**gates**) để điều chỉnh dòng chảy thông tin:

* **Forget Gate (ft):**
  + Quyết định thông tin nào cần "quên" từ trạng thái cell trước đó (ct-1)
  + **.**
* **Input Gate ():**
  + Quyết định thông tin mới nào sẽ được lưu trữ vào cell state hiện tại.
  + **.**
* **Candidate Cell State ():**
  + Tạo ra thông tin tiềm năng để cập nhật trạng thái cell.
  + .
* **Output Gate ():**
  + Quyết định phần nào của cell state *ct*  sẽ được đưa ra làm hidden state *ht*
* **Cập nhật Cell State và Hidden State**:
  + Cell state:
  + Hidden state:

Việc dự đoán giá cổ phiếu đòi hỏi một sự kết hợp của các mô hình khác nhau để có cái nhìn tổng thể về thị trường. Mỗi mô hình đều có ưu và nhược điểm riêng, và sự lựa chọn mô hình dựa vào loại dữ liệu và mục tiêu phân tích cụ thể. Kết hợp nhiều mô hình hoặc sử dụng phương pháp học máy nâng cao có thể cung cấp các dự báo chính xác và đáng tin cậy hơn.

## 2.4. Các chỉ số đánh giá mô hình

### 2.4.1. Sai số trung bình tuyệt đối (MAE)

Sai số tuyệt đối trung bình (Mean Absolute Error – MAE) đại diện cho giá trị trung bình của sự chênh lệch tuyệt đối giữa các giá trị thực tế và giá trị dự đoán trong tập dữ liệu. Nó đo lường mức trung bình của các phần dư trong tập dữ liệu.

Trong đó:

* : Tổng số dữ liệu (số lượng mẫu trong tập dữ liệu).
* : Giá trị thực.
* : Giá trị dự đoán.

### 2.4.2. Sai số trung bình bình phương (MSE)

Sai số bình phương trung bình (Mean Squared Error - MSE) đại diện cho giá trị trung bình của bình phương sự chênh lệch giữa các giá trị thực tế và giá trị dự đoán trong tập dữ liệu. Nó đo lường phương sai của các phần dư.

Sai số này nhạy cảm với các giá trị ngoại lai (outliers) do sử dụng bình phương.

### 2.4.3. Căn bậc hai của sai số trung bình bình phương (RMSE)

Căn bậc hai của sai số bình phương trung bình (Root Mean Squared Error - RMSE) là căn bậc hai của Sai số bình phương trung bình (MSE). Nó đo lường độ lệch chuẩn của các phần dư

### 2.4.4. Hệ số xác định (R2)

Hệ số xác định (Coefficient of Determination hay R-squared - R2) biểu thị tỷ lệ phương sai trong biến phụ thuộc được giải thích bởi mô hình hồi quy tuyến tính. Đây là một chỉ số không phụ thuộc vào thang đo (scale-free), nghĩa là bất kể giá trị lớn hay nhỏ, giá trị của R-squared luôn nhỏ hơn hoặc bằng 1.

Với là giá trị trung bình của tất cả giá trị thực .

### 2.4.5. Sai số phần trăm tuyệt đối trung bình (MAPE)

Sai số phần trăm tuyệt đối trung bình (Mean Absolute Percentage Error – MAPE) biểu diễn sai số tuyệt đối trung bình dưới dạng phần trăm so với giá trị thực. Nó giúp đánh giá độ chính xác của mô hình mà không phụ thuộc vào đơn vị dữ liệu.

Trong đó:

* : để tránh trường hợp chia cho 0.
* MAPE thường được sử dụng để đánh giá độ chính xác trong các bài toán dự báo, đặc biệt hữu ích khi so sánh các mô hình khác nhau.

A screenshot of a diagram

Description automatically generated

Hình 3: Các chỉ số đánh giá mô hình

# CHƯƠNG 3: THU THẬP VÀ PHÂN TÍCH DỮ LIỆU

Trong chương này, chúng tôi trình bày quy trình thu thập dữ liệu và các bước phân tích nhằm phục vụ cho bài toán dự báo giá cổ phiếu mã VNM. Quá trình thu thập dữ liệu được thực hiện từ các nguồn đáng tin cậy, đảm bảo tính chính xác và đầy đủ. Dữ liệu sau khi thu thập được làm sạch, chuẩn hóa và phân tích để hiểu rõ các đặc trưng cũng như mối quan hệ giữa các biến, làm nền tảng cho việc xây dựng và đánh giá các mô hình dự báo.

Thông qua chương này, bạn đọc sẽ có cái nhìn tổng quan về dữ liệu được sử dụng, đồng thời hiểu rõ cách dữ liệu hỗ trợ việc xây dựng mô hình dự báo giá cổ phiếu hiệu quả.

## 3.1. Thu thập dữ liệu

Dataset được thu thập từ trang web [Simplize.vn](https://simplize.vn/co-phieu/VNM), cung cấp thông tin liên quan đến giá cổ phiếu của mã chứng khoán **VNM** (Vinamilk).

## 3.2. Mô tả dữ liệu

### 3.2.1. Tổng số dòng và cột trong tập dữ liệu

Dữ liệu giá cổ phiếu được thu thập từ ngày 26/11/2014 đến ngày 2/1/2024, tổng cộng có 2719 dòng. Bộ dữ liệu này bao gồm 7 cột chính: Ngày, Đóng cửa, Mở, Cao, Thấp, Khối lượng (KL) và % Thay đổi. Cột Ngày cung cấp thông tin về các ngày giao dịch, giúp xác định rõ mốc thời gian của dữ liệu. Cột Đóng cửa ghi lại giá cổ phiếu khi kết thúc phiên giao dịch mỗi ngày, trong khi Mở thể hiện giá cổ phiếu khi bắt đầu phiên giao dịch. Cột Cao và Thấp cung cấp thông tin về mức giá cao nhất và thấp nhất trong phiên giao dịch mỗi ngày. Cột Khối lượng (KL) cho biết số lượng cổ phiếu được giao dịch trong ngày, và % Thay đổi thể hiện tỷ lệ thay đổi giá giữa ngày hôm trước và ngày hiện tại.

**A screenshot of a computer

Description automatically generated**

Hình 4: Lệnh đọc dữ liệu từ dataset

**A screenshot of a computer

Description automatically generated**

Hình 5: Hình ảnh dữ liệu của dataset

### 3.2.2. Xác định kiểu dữ liệu của mỗi cột

**A screenshot of a computer

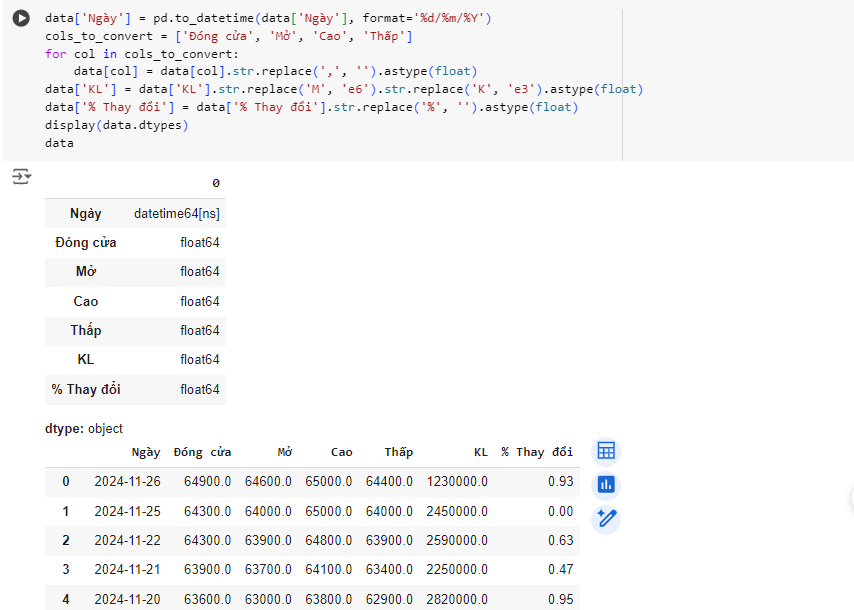
Description automatically generated**

Hình 6: Kiểu dữ liệu của từng cột

Sau khi phân tích, chúng em nhận thấy rằng các cột trong bộ dữ liệu giá cổ phiếu có kiểu dữ liệu object, phản ánh sự kết hợp của chuỗi và giá trị số. Để chuẩn bị dữ liệu một cách phù hợp cho các phân tích tiếp theo, chúng em sẽ thực hiện các chuyển đổi cần thiết.

### 3.2.3. Xử lý kiểu dữ liệu

Đầu tiên, cột Ngày sẽ được chuyển đổi thành kiểu datetime để thuận tiện cho việc theo dõi và phân tích xu hướng theo thời gian. Các cột Đóng cửa, Mở, Cao, Thấp, và % Thay đổi, KL sẽ được chuyển đổi sang kiểu float để dễ dàng thực hiện các phép toán thống kê và so sánh.

****

Hình 7: Code xử lý kiểu dữ liệu

**A screenshot of a computer

Description automatically generated**

Hình 8: Kiểu dữ liệu của từng cột sau khi được xử lý

## 3.3. Sắp xếp dữ liệu theo thời gian

Sau khi xử lý kiểu dữ liệu, chúng em tiến hành sắp xếp dữ theo thời gian bằng cách sắp xếp theo cột Ngày. Điều này giúp tổ chức dữ liệu theo trình tự thời gian từ sớm đến muộn, hỗ trợ phân tích xu hướng và các mối tương quan trong thời gian một cách chính xác hơn.

**A screenshot of a computer

Description automatically generated**

Hình 9: Code sắp xếp dữ liệu tăng dần theo thời gian

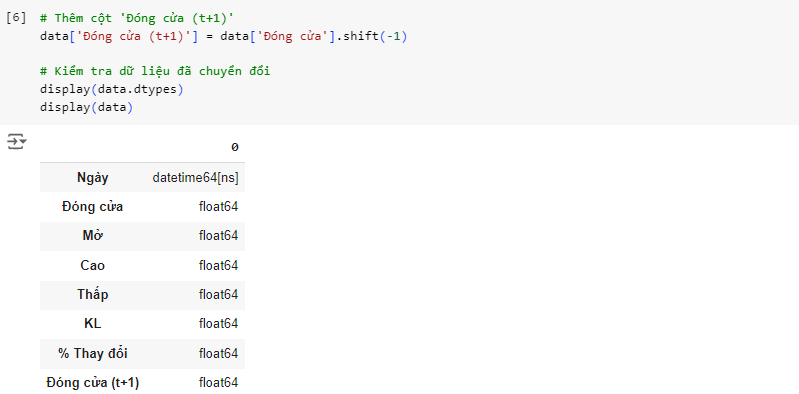
**A screenshot of a computer

Description automatically generated**

Hình 10: Dữ liệu sau khi được sắp xếp tăng dần theo thời gian

## 3.4. Thêm vào cột Đóng cửa (t+1) (Giá đóng cửa ngày mai)

Tiến hành thêm vào cột Đóng cửa (t+1) để lưu giá trị giá đóng cửa ngày mai để phục vụ cho mục đích dự đoán giá đóng cửa ngày mai.



Hình 11: Code thêm cột Giá đóng cửa ngày mai vào dữ liệu

## 3.5. Kiểm tra tính toàn vẹn của dữ liệu

Trong quá trình phân tích dữ liệu, việc kiểm tra các dòng trùng lặp, dữ liệu thiếu và giá trị ngoại lệ là rất quan trọng. Đầu tiên, chúng tôi xác định các dòng trùng lặp để loại bỏ hoặc sửa chữa, đảm bảo dữ liệu duy nhất và chính xác. Sau đó, chúng tôi kiểm tra dữ liệu thiếu để phát hiện và xử lý các giá trị NaN. Cuối cùng, sử dụng các thống kê mô tả, biểu đồ boxplot và hệ số IQR,… để xác định giá trị ngoại lệ, giúp nhận diện và xử lý những giá trị không hợp lý trong dữ liệu. Việc làm sạch dữ liệu giúp nâng cao chất lượng và tính tin cậy của mô hình dự đoán giá cổ phiếu.

### 3.5.1. Kiểm tra các dòng trùng lặp

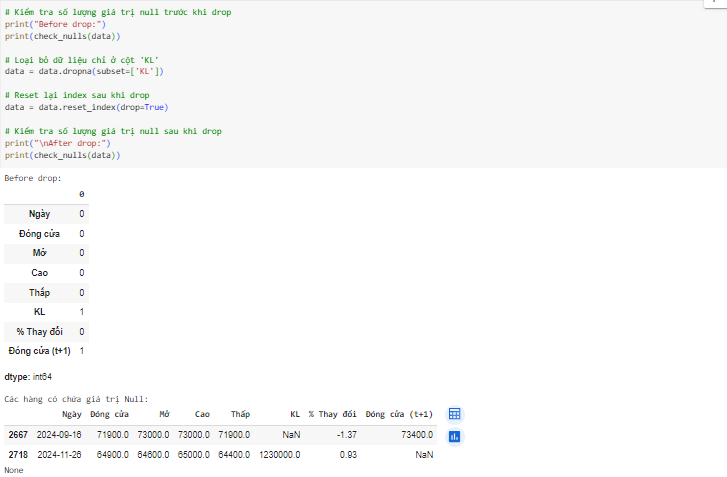
Sau khi kiểm tra, không thấy dòng dữ liệu nào bị trùng lặp trong dataset.

**A computer screen shot of a code

Description automatically generated**

Hình 12: Số dòng trùng lặp trong dữ liệu

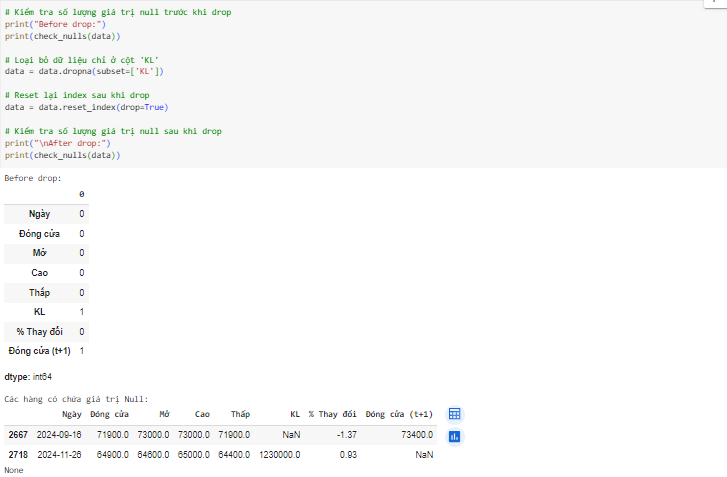
### 3.5.2. Kiểm tra và xử lý dữ liệu thiếu



A screenshot of a computer

Description automatically generated

Hình 13: Kết quả kiểm tra dữ liệu thiếu



Hình 14: Các dòng có giá trị thiếu

Sau khi kiểm tra thì phát hiện có 2 dòng tồn tại giá trị thiếu, tiến hành loại bỏ dòng có giá trị NaN.

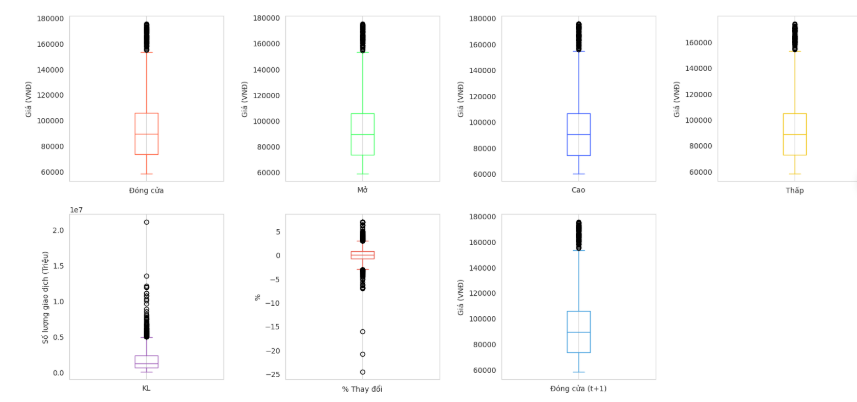
A screenshot of a computer

Description automatically generated

Hình 15: Số lượng giá trị thiếu ở từng cột sau khi được xử lý giá trị thiếu

### 3.5.3. Xác định giá trị ngoại lệ trong các cột (outliers detection)

#### 3.5.3.1. Biểu đồ boxplot:



Hình 16: Biểu đồ boxplot

Nhận xét biểu đồ trên:

1. Đóng cửa và Đóng cửa (t+1):

* Mức giá bình thường dao động chủ yếu trong khoảng từ 80,000 đến 140,000 VNĐ.
* Một số ít ngày có giá đóng cửa vượt quá 160,000 VNĐ, đây là giá trị ngoại lệ rõ ràng.

1. Mở:

* Giá mở chủ yếu nằm trong khoảng từ 80,000 đến 120,000 VNĐ.
* Một số ít điểm nằm ngoài phạm vi này với giá mở cao hơn 140,000 VNĐ.

1. Cao:

* Mức giá cao chủ yếu dao động từ 80,000 đến 150,000 VNĐ.
* Một số ít ngày có mức giá cao trên 160,000 VNĐ.

1. Thấp:

* Mức giá thấp chủ yếu nằm trong khoảng từ 60,000 đến 90,000 VNĐ.
* Một vài điểm nằm ngoài phạm vi này với giá thấp dưới 50,000 VNĐ.

1. KL (Khối lượng giao dịch):

* Khối lượng giao dịch chủ yếu dao động từ 1 đến 5 triệu (M).
* Một số ít ngày có khối lượng lớn hơn 10 triệu, cho thấy các giao dịch đặc biệt hoặc biến động lớn.

1. % Thay đổi:

* Phần lớn thay đổi giá nằm trong khoảng từ -5% đến +5%.
* Một số ít ngày có thay đổi lớn hơn 10%, cho thấy các biến động đột ngột trong giá cổ phiếu.

#### 3.5.3.2. Xác định ngoại lệ bằng cách tính IQR

IQR (Interquartile Range) hay khoảng tứ phân vị là thước đo thống kê biểu thị độ phân tán của dữ liệu. Nó được tính bằng hiệu số giữa Q3 (tứ phân vị thứ ba) và Q1 (tứ phân vị thứ nhất):

* Q1 (tứ phân vị thứ nhất): Là giá trị mà 25% dữ liệu nằm dưới nó.
* Q3 (tứ phân vị thứ ba): Là giá trị mà 75% dữ liệu nằm dưới nó.
* IQR: Thể hiện khoảng giá trị trung bình của 50% dữ liệu ở giữa.

Các giá trị ngoại lệ nằm ngoài khoảng:

* Lower Bound: Giá trị nhỏ nhất dữ liệu không ngoại lệ.
* Upper Bound: Giá trị lớn nhất dữ liệu không ngoại lệ.

Dữ liệu nằm ngoài khoảng lower bound và upper bound được coi là ngoại lệ.

A screenshot of a computer

Description automatically generated

Hình 17: Giá trị ngoại lệ của cột giá đóng cửa

#### 3.5.3.3. Độ nhọn và Độ lệch

Skewness phản ánh độ lệch của dữ liệu khỏi dạng chuẩn. Nếu giá trị Skewness lớn hơn 0, phân phối lệch phải; nếu nhỏ hơn 0, phân phối lệch trái.

Kurtosis đo độ nhọn của phân phối, cho biết có nhiều hay ít điểm cực trị (điểm nhọn hoặc đuôi dài) trong dữ liệu. Giá trị Kurtosis lớn hơn 3 cho thấy phân phối có đuôi dài và nhọn hơn phân phối chuẩn, trong khi giá trị nhỏ hơn 3 cho thấy đuôi ngắn và phân phối ít nhọn hơn.

A screenshot of a computer program

Description automatically generated

Hình 18: Code kiểm tra độ lệch và độ nhọn của các biến

Từ dữ liệu trên, có được nhận xét như sau:

* Skewness:
* Đóng cửa, Mở, Cao, Thấp có skewness dương (~1), cho thấy các giá trị thường cao hơn giá trị trung bình.
* KL (Khối lượng) có skewness rất cao (~2.64), chỉ ra rằng dữ liệu này có sự biến động mạnh.
* % Thay đổi có skewness âm (-2.18), cho thấy xu hướng giảm nhiều hơn tăng.
* Kurtosis:
* Đóng cửa, Mở, Cao, Thấp có kurtosis gần bằng 1, cho thấy phân bố gần với phân phối chuẩn.
* KL có kurtosis rất cao (~15.28), phản ánh nhiều điểm cực trị hơn.
* % Thay đổi có kurtosis cực kỳ cao (~34.34), cho thấy sự biến động lớn với nhiều giá trị cực đoan.

## 3.6. Kiểm tra tính ổn định của dữ liệu

Kiểm tra tính ổn định của dữ liệu (stationarity) là một bước quan trọng trong phân tích chuỗi thời gian vì nó ảnh hưởng trực tiếp đến hiệu quả và độ chính xác của các mô hình dự đoán. Một chuỗi thời gian được coi là ổn định (stationary) nếu các đặc trưng thống kê của nó (như trung bình, phương sai, và tự tương quan) không thay đổi theo thời gian.

A screenshot of a computer

Description automatically generated

Hình 19: Code kiểm tra tính ổn định của cột Đóng cửa

A screenshot of a computer code

Description automatically generated

Hình 20: Code kiểm tra tính ổn định của cột giá đóng cửa ngày mai

Nhận xét:

* ADF Statistic là -1.435 và -1.419 cho thấy rằng chuỗi không đạt được giá trị đủ thấp để từ chối giả thuyết rằng chuỗi là không ổn định. Thường thì giá trị của ADF Statistic cần nằm dưới -2.58 (với mức ý nghĩa 1%) hoặc -1.96 (với mức ý nghĩa 5%) để có thể kết luận chuỗi là ổn định (stationary).
* p-value là 0.565 và 0.57 lớn hơn 0.05. Điều này cho thấy không có đủ bằng chứng để bác bỏ giả thuyết rằng chuỗi là không ổn định (non-stationary).

Do đó, từ những kết quả này, chuỗi "Đóng cửa" và “ Đóng cửa (t+1) không được coi là ổn định theo kiểm định ADF.

## 3.7. Trực quan hóa dữ liệu

### 3.7.1. Kiểm tra phân phối của dữ liệu

#### 3.7.1.1. Biểu đồ histogram

A group of blue and white graphs

Description automatically generated

Hình 21: Biểu đồ histogram

Dựa trên biểu đồ trên, nhận xét như sau:

* Các biểu đồ phân bố cho thấy dữ liệu của bạn có sự phân tán không đều. Giá Đóng cửa và Giá Mở có xu hướng lệch trái, với phần lớn giá trị thấp tập trung ở một vùng và giá trị cao hơn rải rác. Giá Cao và Giá Thấp có hình dạng lệch phải, cho thấy sự biến động lớn hơn trong các giá trị này. Khối lượng giao dịch (KL) có thể có một vài giá trị cực lớn hoặc đỉnh, biểu thị sự biến động mạnh.
* Biến % Thay đổi có phân phối gần như đồng đều quanh giá trị trung bình, cho thấy sự thay đổi không lớn trong giá trị % thay đổi.
* Những đặc điểm này cho thấy rằng dữ liệu có tính không tuyến tính và có sự biến động lớn.

#### 3.7.1.2. Biểu đồ scatter

A group of graphs showing different types of data

Description automatically generated with medium confidence

Hình 22: Biểu đồ scatter

Dựa trên biểu đồ trên, nhận xét như sau:

* Các biểu đồ scatter cho thấy mối quan hệ giữa các biến khác nhau với giá Đóng cửa. Giữa Giá Mở, Giá Cao và Giá Thấp với Giá Đóng cửa có mối tương quan tuyến tính rõ ràng, cho thấy sự phản ánh gần đúng của giá Đóng cửa với các giá trị này. Tuy nhiên, biểu đồ giữa % Thay đổi và Giá Đóng cửa không tuyến tính, cho thấy sự thay đổi lớn trong giá trị % có thể không tương ứng tuyến tính với sự thay đổi của giá Đóng cửa.
* Biểu đồ mối quan hệ giữa Khối lượng (KL) và Giá Đóng cửa cũng thể hiện sự tương quan nghịch tuyến tính, với sự giảm khối lượng không nhất thiết phải dẫn đến sự giảm tương ứng của giá Đóng cửa.

### 3.7.2. Khám phá mối quan hệ giữa các biến thông qua ma trận tương quan

A red and blue squares with numbers

Description automatically generated

Hình 23: Ma trận tương quan giữa các biến

Dựa trên ma trận tương quan của các biến trong dữ liệu, nhận xét như sau:

* + Giá đóng cửa (Đóng cửa):
* Giá mở (Mở), Giá cao (Cao), và Giá thấp (Thấp) đều có tương quan rất cao với giá đóng cửa, với giá trị gần 1. Điều này cho thấy rằng giá đóng cửa có sự liên quan mạnh mẽ với các biến này, phản ánh mối quan hệ chặt chẽ giữa các giá trị khác nhau của giá.
* Khối lượng (KL) và % thay đổi (% Thay đổi) có tương quan thấp hơn với giá đóng cửa, với giá trị âm hoặc gần 0. Điều này có thể chỉ ra rằng giá đóng cửa không bị ảnh hưởng mạnh mẽ bởi các biến này.
  + Các biến khác:
* % Thay đổi và Khối lượng cũng có tương quan thấp với nhau, điều này có thể cho thấy rằng sự thay đổi giá ít ảnh hưởng đến khối lượng giao dịch hoặc ngược lại.
* Tương quan giữa các biến như Cao và Mở, Thấp và Mở, và tương tự, đều rất cao, cho thấy rằng giá trong các phiên giao dịch có mối liên hệ mật thiết với nhau.

### 3.7.3. Kiểm tra xu hướng dữ liệu

#### 3.7.3.1. Biểu đồ đường của giá đóng cửa

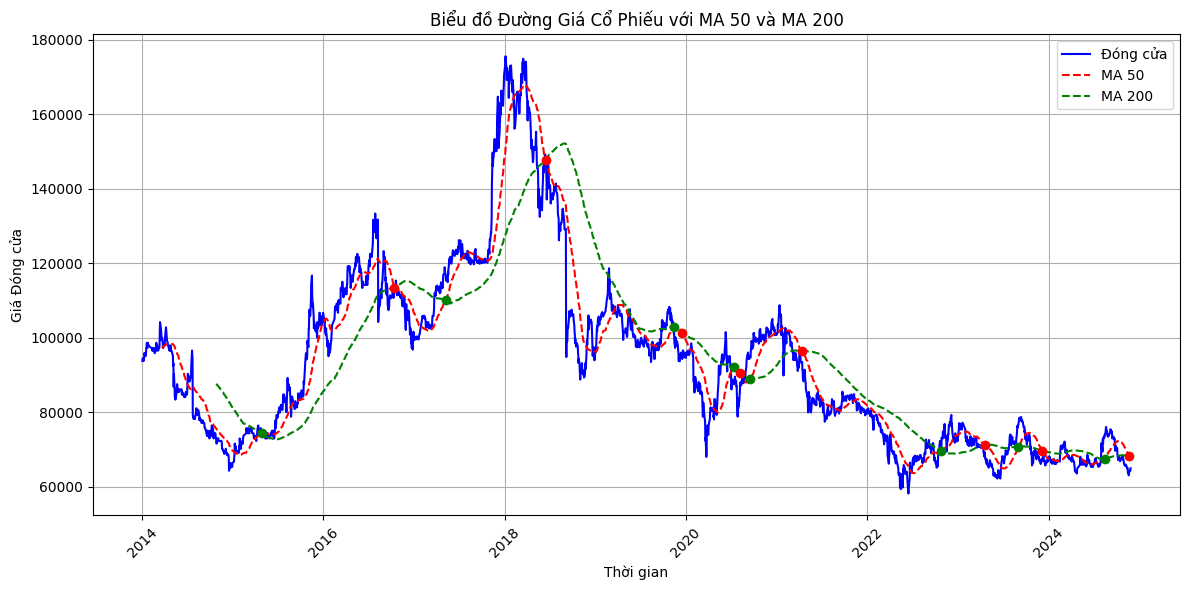
****

Hình 24: Biểu đồ đường của giá đóng cửa theo thời gian

Dựa trên biểu đồ trên, nhận xét như sau:

* Xu hướng chính:
* Giai đoạn từ 2014 đến 2018, giá đóng cửa có xu hướng tăng mạnh, đạt đỉnh vào năm 2018.
* Sau năm 2018, giá giảm mạnh và biến động liên tục cho đến 2024, nhưng xu hướng tổng thể là giảm dần.
* Biến động:
* Giai đoạn từ 2016 đến 2018 thể hiện sự tăng trưởng đột phá, nhưng cũng xuất hiện mức độ biến động lớn.
* Từ 2020 đến 2024, giá ít biến động hơn nhưng duy trì xu hướng giảm.
* Ý nghĩa:
* Thị trường có dấu hiệu không ổn định và chịu ảnh hưởng lớn bởi các yếu tố kinh tế hoặc sự kiện cụ thể trong giai đoạn 2018-2020.
* Việc giảm giá sau 2018 có thể phản ánh xu hướng suy thoái hoặc sự điều chỉnh sau giai đoạn tăng trưởng nóng.

#### 3.7.3.2. Biểu đồ đường của giá đóng cửa với MA50 và MA200



Hình 25: Biểu đồ đường của giá đóng cửa với MA50 và MA200

Nhận xét biểu đồ đường giá cổ phiếu với MA 50 và MA 200:

* Xu hướng dài hạn (MA 200):
  + Đường MA 200 thể hiện xu hướng tổng thể giảm dần từ 2018 đến 2024, phản ánh sự suy giảm dài hạn của giá cổ phiếu.
* Xu hướng ngắn hạn (MA 50):
  + Đường MA 50 có xu hướng dao động mạnh hơn so với MA 200, phản ánh sự biến động ngắn hạn trong giá cổ phiếu.
  + Các giao điểm giữa MA 50 và MA 200 (gọi là "Golden Cross" hoặc "Death Cross") có thể gợi ý về các tín hiệu mua/bán.
* Mối quan hệ giữa giá đóng cửa và MA:
  + Giai đoạn giá tăng (2016-2018): Đường giá đóng cửa nằm trên cả MA 50 và MA 200.
  + Giai đoạn giá giảm (sau 2018): Đường giá thường nằm dưới MA 200, thể hiện áp lực giảm giá.
* Ý nghĩa:
  + Biểu đồ cho thấy sự chuyển đổi giữa xu hướng tăng trưởng (trước 2018) và suy giảm (sau 2018).

# CHƯƠNG 4: XÂY DỰNG MÔ HÌNH VÀ KẾT QUẢ

## 4.1. Mô hình hồi quy tuyến tính đa biến

Mô hình hồi quy tuyến tính đa biến là một mở rộng của mô hình hồi quy tuyến tính đơn biến, trong đó chúng ta cố gắng dự đoán giá trị của một biến phụ thuộc (Y) thông qua một hoặc nhiều biến độc lập (X\_1, X\_2, ..., X\_n). Đây là một phương pháp thống kê thường được sử dụng để xác định và đánh giá mối quan hệ tuyến tính giữa các biến trong dữ liệu.

### 4.1.1. Chuẩn bị xây dựng mô hình

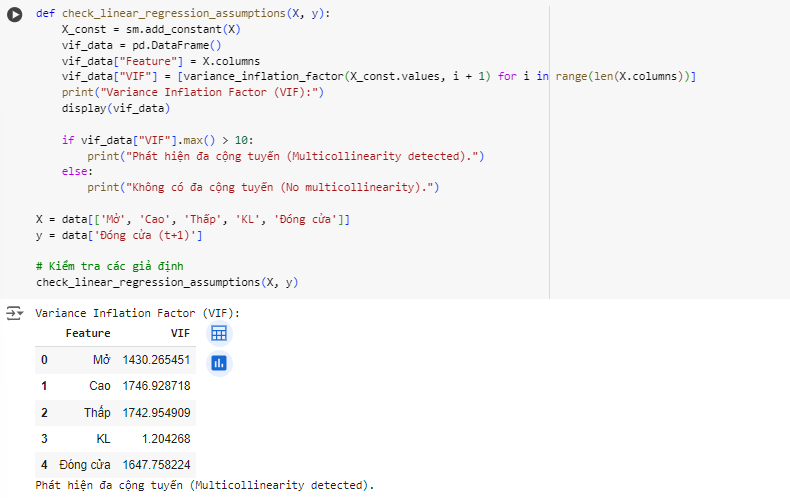
Từ các bước trực quan hóa cũng như kiểm tra ma trận tương quan giữa các biến thì chúng em quyết định chọn các biến như Đóng cửa, Mở, Thấp, Cao, KL để làm biến đầu vào cho mô hình hồi quy tuyến tính.

#### 4.1.1.1. Kiểm tra đa cộng tuyến (VIF)

VIF đo lường mức độ mà một biến độc lập trong tập dữ liệu bị "giải thích" bởi các biến độc lập khác.

Vì chúng em muốn dự đoán giá cổ phiếu với các biến còn lại trừ biến Đóng cửa (t+1) nên chúng em sẽ tiến hành kiểm tra đa cộng tuyến giữa 5 biến Cao, Thấp, Mở, Đóng cửa, KL để xem giữa các biến có mối quan hệ chặt chẽ với nhau hay không.

Tiến hành kiểm tra VIF cho các biến độc lập để đánh giá mức độ tương quan giữa chúng. Sử dụng hàm variance\_inflation\_factor từ thư viện statsmodels để tính VIF cho từng biến.

****

Hình 26: Code kiểm tra đa cộng tuyến

Theo kết quả trên, ta có nhận xét như sau:

* Các biến Cao, Mở, Thấp, Đóng cửa có chỉ số VIF lần lượt là 1781.954090, 1771.592315, 1747.769727 và 2306.403804, vượt xa ngưỡng chấp nhận được (VIF > 10). Điều này cho thấy các biến này có mối tương quan cực kỳ cao với nhau, gây ảnh hưởng lớn đến độ ổn định của hệ số hồi quy và tính giải thích của mô hình.
* KL có VIF thấp (1.21), không có vấn đề về đa cộng tuyến.

#### 4.1.1.2. Xử lý đa cộng tuyến bằng cách tạo biến Trung bình

Để xử lý đa cộng tuyến, một cách hiệu quả là tổng hợp các biến có tương quan cao thành một biến mới. Trong trường hợp này, các biến Cao, Mở, Thấp và Đóng cửa có thể được tổng hợp thành một biến đại diện bằng cách tính trung bình cộng hoặc trung vị của chúng.

A screenshot of a computer

Description automatically generated

Hình 27: Code xử lý đa cộng tuyến

Sau khi có được biến Trung bình ta tiến hành kiểm tra lại đa cộng tuyến thì thấy đã xử lý được tình trạng đa cộng tuyến.

#### 4.1.1.3. Kiểm tra lại sự ảnh hưởng của các biến đầu vào mục tiêu đến giá trị cần dự đoán (Đóng cửa (t+1))

A screenshot of a computer

Description automatically generated

Hình 28: Kiểm tra sự ảnh hưởng của biến đầu vào với biến đầu ra

Từ thông tin trên, có nhận xét như sau:

* Trung bình:
* Hệ số là 0.9991, rất gần với 1, cho thấy biến Trung bình có ảnh hưởng lớn và tích cực đến biến Đóng cửa (t+1).
* p-value = 0.000 (rất nhỏ), ảnh hưởng này có ý nghĩa thống kê.
* KL (Khối lượng):
* Hệ số là 3.577e-05 (0.00003577), cho thấy ảnh hưởng của KL là rất nhỏ đến giá đóng cửa.
* p-value = 0.074:
  + Ở mức ý nghĩa 5% (0.05), biến KL không có ý nghĩa thống kê (tức là nó có thể không ảnh hưởng đáng kể đến giá đóng cửa).
  + Tuy nhiên, ở mức ý nghĩa 10% (0.10), nó có thể được coi là có ý nghĩa, nhưng ảnh hưởng vẫn nhỏ.

#### 4.1.1.4. Chuẩn hóa dữ liệu

Chuẩn hóa dữ liệu bằng cách sử dụng hàm MinMaxScaler() của thư viện sklearn.preprocessing. Chuẩn hóa các biến đầu vào cần sử dụng cho mô hình.

**A screenshot of a computer

Description automatically generated**

Hình 29: Code chuẩn hóa dữ liệu cho HQTT

#### 4.1.1.5. Chia dữ liệu

Tiến hành chia dữ liệu thành 2 tập train và test theo tỉ lệ 80:20. Tập train sẽ bao gồm dữ liệu giá từ đầu cho đến 80% dữ liệu. Tập test sẽ là phần còn lại của dữ liệu. Các cột đầu vào của mô hình bao gồm: KL và Trung bình. Đâu ra sẽ là Đóng cửa (t+1) (Giá đóng cửa ngày mai).

**A screenshot of a computer

Description automatically generated**

Hình 30: Code chia dữ liệu của HQTT

### 4.1.2. Xây dựng mô hình

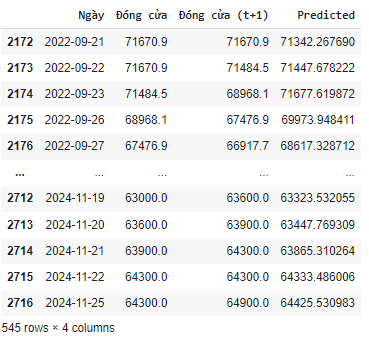
Viết hàm xây dựng và dự đoán mô hình sử dụng thư viện LinearRegression.Vẽ biểu đồ giữa giá thực tế và giá dự đoán để trực quan mức độ chính xác của dự đoán.





Hình 31: Code xây dựng mô hình HQTT

Hàm này sẽ in ra được giá đóng cửa dự đoán và vẽ đồ thị trực quan kết quả dự đoán so với giá trị giá đóng cửa thực tế.



Hình 32: Kết quả dự đoán giá đóng cửa ngày mai

A graph of a graph

Description automatically generated with medium confidence

Hình 33: Biểu đồ so sánh giá dự đoán và giá thực tế của HQTT

## 4.2. Mô hình XGBoost

XGBoost (eXtreme Gradient Boosting) là một thuật toán tăng cường học máy, được thiết kế đặc biệt để cải thiện hiệu quả của các mô hình hồi quy và phân loại. Nó kết hợp kỹ thuật Boosting với các lý thuyết học máy tiên tiến như Gradient Boosting và Regularization, giúp cải thiện độ chính xác và tốc độ của mô hình. XGBoost sử dụng thuật toán chia nhỏ nhánh (tree-based model) để giảm thiểu lỗi dự đoán bằng cách học một loạt các cây quyết định (decision trees) theo thứ tự từ đơn giản đến phức tạp.

Mặc dù XGBoost không phải là mô hình chuyên biệt cho dữ liệu chuỗi thời gian như LSTM hay ARIMA, nó vẫn có thể được áp dụng để dự đoán dữ liệu chuỗi thời gian như giá cổ phiếu. Tuy nhiên, cần lưu ý rằng hiệu quả của nó có thể không đạt được mức tốt nhất như các mô hình chuyên biệt hơn, nhưng nó vẫn có thể là một lựa chọn hợp lý.

### 4.2.1. Chuẩn bị xây dựng mô hình

Đối với XGBOOST thì chúng em cũng cho nó đầu vào giống với Hồi quy tuyến tính. Input cũng bao gồm các cột như Trung bình và KL.

#### 4.2.1.1. Chuẩn hóa dữ liệu

Chuẩn hóa dữ liệu bằng cách sử dụng hàm MinMaxScaler() của thư viện sklearn.preprocessing. Chuẩn hóa các biến đầu vào cần sử dụng cho mô hình.

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

Hình 34: Code chuẩn hóa dữ liệu của XGBoost

#### 4.2.1.2. Chia dữ liệu

Tiến hành chia dữ liệu thành 2 tập train và test theo tỉ lệ 80:20. Tập train sẽ bao gồm dữ liệu giá từ đầu cho đến 80% dữ liệu. Tập test sẽ là phần còn lại của dữ liệu. Các cột đầu vào của mô hình bao gồm: KL và Trung bình. Đâu ra sẽ là Đóng cửa (t+1) (Giá đóng cửa ngày mai).

**A screenshot of a computer

Description automatically generated**

Hình 35: Code chia dữ liệu của XGBoost

#### 4.2.1.3. Chọn các chỉ số phù hợp với mô hình XGBOOST bằng GridSearchCV

Chọn các chỉ số phù hợp bằng GridSearchCV để mô hình đạt được độ chính xác cũng như tối ưu nhất.

**A screenshot of a computer

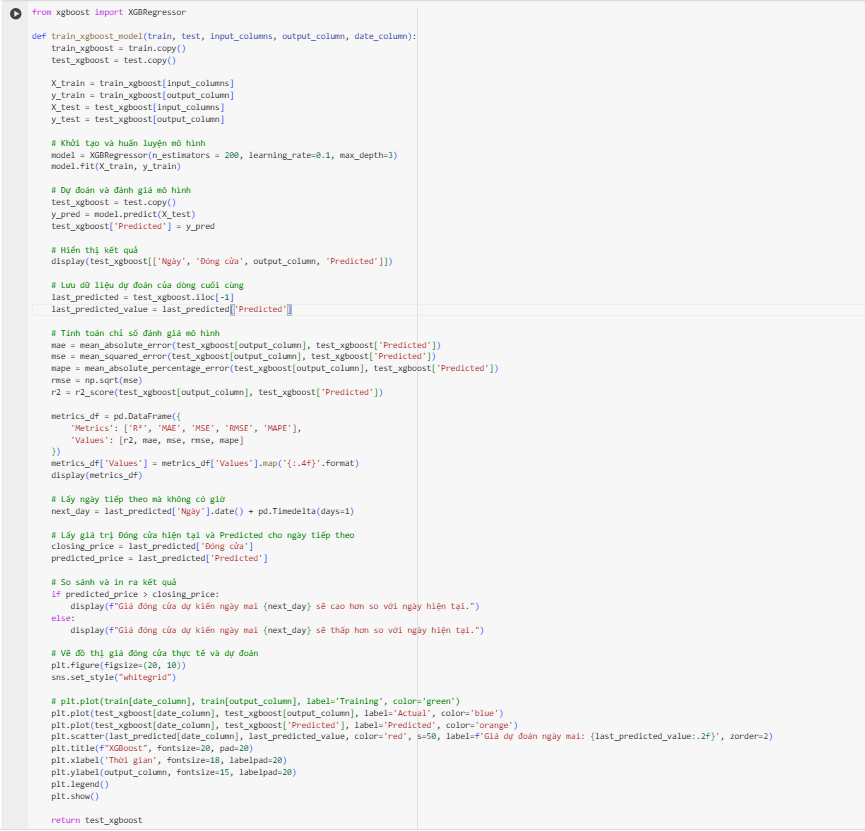
Description automatically generated**

Hình 36: Code chọn các chỉ số phù hợp với XGBoost

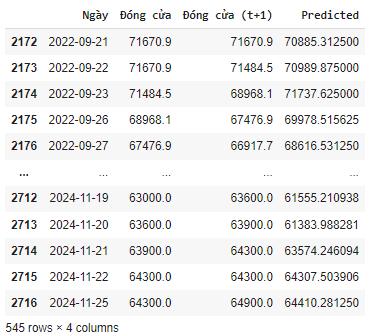
Các chỉ số phù hợp với dữ liệu hiện tại được tìm ra là learning\_rate: = 0.1, max\_depth = 3, n\_estimators = 200.

### 4.2.2. Xây dựng mô hình

Viết hàm xây dựng và dự đoán giá sử dụng thư viện XGRegressor. Sau đó in ra giá dự đoán và vẽ biểu đồ giữa giá thực tế và giá dự đoán để trực quan mức độ chính xác của dự đoán.

****

Hình 37: Code xây dựng mô hình XGBoost

****

Hình 38: Kết quả dự đoán của XGBoost

**A graph with blue and orange lines

Description automatically generated**

Hình 39: Biểu đồ so sánh giá dự đoán và giá thực tế của XGBoost

## 4.3. Mô hình LSTM

### 4.3.1. Chuẩn bị xây dựng mô hình

#### 4.3.1.1. Chuẩn hóa dữ liệu:

* Chuẩn hóa tất cả các cột số liệu về khoảng [0, 1] bằng phương pháp **Min-Max Scaling**:
  + Điều này giúp cải thiện hiệu quả học của LSTM, tránh trường hợp giá trị lớn lấn át.

#### 4.3.1.2. Tạo tập dữ liệu chuỗi thời gian (Sliding Window):

* Sử dụng **cửa sổ trượt**:
  + Mỗi chuỗi đầu vào gồm 60 giá trị trước đó để dự đoán giá trị ngày tiếp theo.
  + **X**: 60 ngày trước đó (biến đầu vào).
  + **y**: Giá đóng cửa ngày thứ 61 (biến mục tiêu)

#### 4.3.1.3. Chia tập dữ liệu

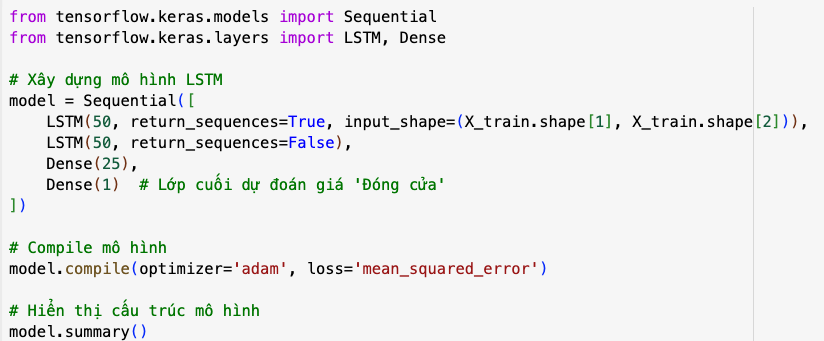
* Chia tập dữ liệu thành:
  + Tập huấn luyện (Training set):
    - 80% dữ liệu đầu tiên.
    - Dùng để huấn luyện mô hình.
  + Tập kiểm tra (Test set):
    - 20% dữ liệu cuối cùng.
    - Dùng để đánh giá hiệu suất của mô hình.

#### 4.3.1.4. Thư viện và công cụ

* Thư viện chính:
  + TensorFlow/Keras: Xây dựng và huấn luyện mô hình LSTM.
  + Pandas, NumPy: Xử lý dữ liệu.
  + Matplotlib, Seaborn: Trực quan hóa kết quả.
* Môi trường thực thi:
  + Google Colab hoặc Jupyter Notebook với GPU để tăng tốc độ huấn luyện.

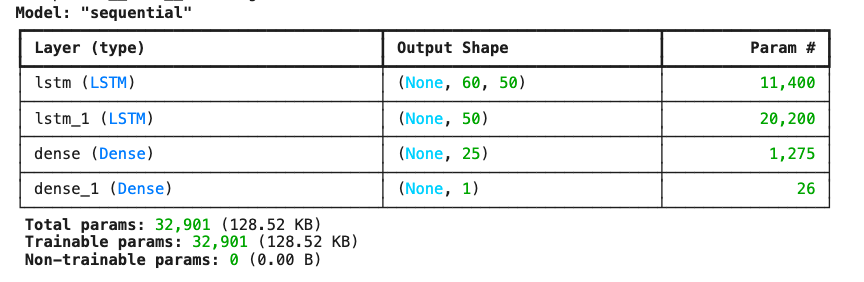
### 4.3.2. Xây dựng mô hình

* **2 lớp LSTM**:
  + Lớp 1: *return\_sequences=True* giữ lại chuỗi đầu ra để chuyển sang lớp LSTM thứ 2.
  + Lớp 2: *return\_sequences=False* vì đây là lớp cuối cùng của LSTM.
* **2 lớp Dense**:
  + Dense(25): Lớp trung gian với 25 neurons.
  + Dense(1): Lớp cuối cùng với 1 neuron dự đoán giá trị "Đóng cửa".

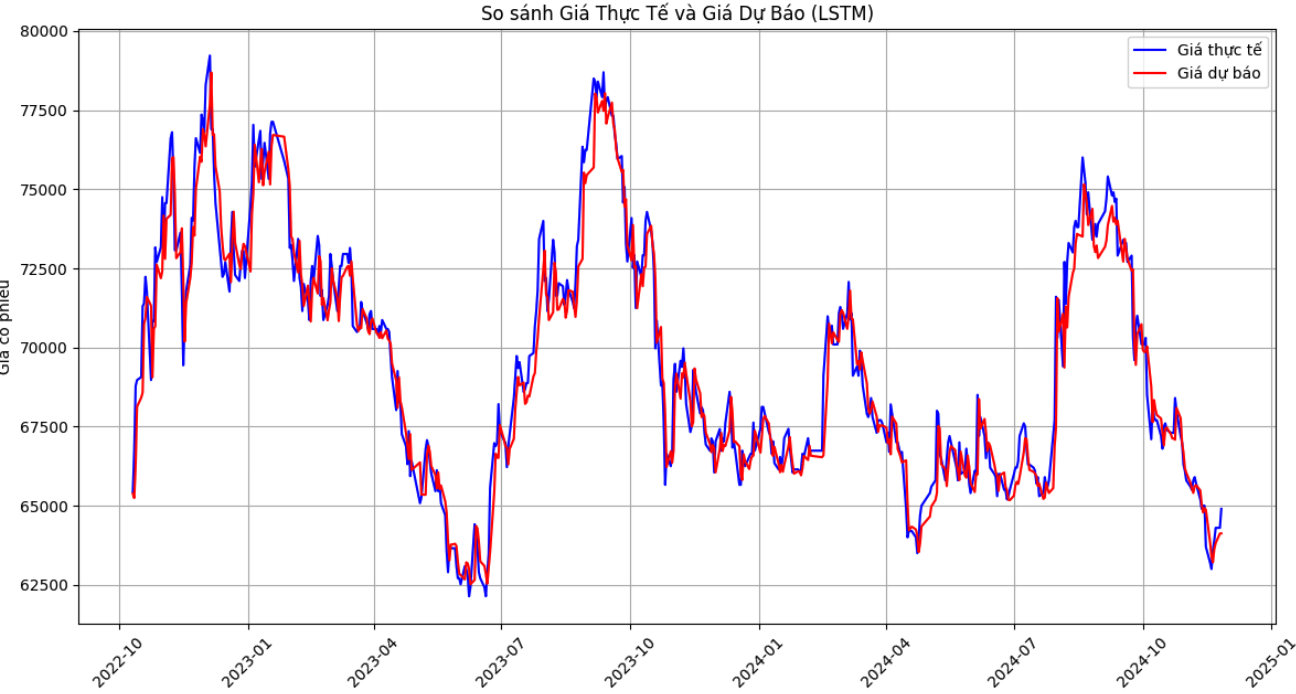


Hình 40: Code xây dựng mô hình

* **Biên dịch mô hình**:
  + Sử dụng Adam optimizer: Phù hợp cho LSTM với tốc độ hội tụ tốt.
  + Hàm mất mát: Mean Squared Error (MSE) để đo lường sai số
* **Cấu trúc mô hình:**



Hình 41: Cấu trúc mô hình LSTM



Hình 42: Biểu đồ so sánh giá dự đoán và giá thực tế của LSTM

## 4.4. Mô hình đưa ra quyết định và trọng số của các đặc trưng trong từng quyết định mua, bán và giữ.

### 4.4.1. Random Forest Classifier

#### 4.4.1.1. Khái niệm

Random Forest là một thuật toán ensemble learning được sử dụng cho cả bài toán phân loại (classification) và hồi quy (regression).

Nó hoạt động bằng cách kết hợp nhiều mô hình decision trees (cây quyết định) để tạo ra một mô hình mạnh mẽ và chính xác hơn.

#### 4.4.1.2. Cách hoạt động

* Bootstrap Sampling:
  + Random Forest tạo ra nhiều mẫu dữ liệu khác nhau từ tập dữ liệu gốc bằng cách sử dụng phương pháp bootstrap sampling (lấy mẫu có hoàn lại).
* Xây dựng Decision Trees:
  + Với mỗi mẫu dữ liệu, một cây quyết định (decision tree) được xây dựng.
  + Tại mỗi nút của cây, chỉ một tập hợp con ngẫu nhiên các feature được xem xét để tìm ra điều kiện phân tách tốt nhất.
* Kết hợp kết quả:
  + Phân loại: Kết quả đầu ra của Random Forest là bỏ phiếu đa số từ các cây (majority voting).
  + Hồi quy: Đầu ra là trung bình của các dự đoán từ các cây.

#### 4.4.1.3. Vai trò trong bài toán

* Random Forest Classifier được sử dụng để phân loại các quyết định Buy (1), Sell

(-1), và Hold (0) dựa trên các chỉ số kỹ thuật (MA, Volatility, Trend, Seasonality).

* Nó tận dụng dữ liệu huấn luyện để xây dựng mô hình dự đoán đáng tin cậy.

### 4.4.2. SHAP (SHapley Additive exPlanations)

#### 4.4.2.1. Khái niệm

SHAP là một phương pháp giải thích các mô hình máy học dựa trên lý thuyết trò chơi (game theory).

Mục tiêu của SHAP là cung cấp một cách tiếp cận công bằng để đo lường mức độ đóng góp của từng feature đối với dự đoán của mô hình.

#### 4.4.2.2. Cách hoạt động

* Shapley Value (Giá trị Shapley):

SHAP tính toán mức độ đóng góp của từng feature bằng cách xem xét tất cả các tổ hợp có thể của các feature khác.

Ví dụ: Với 4 feature (A, B, C, D), SHAP sẽ xem xét đóng góp của A trong các tổ hợp như {}, {B}, {C}, {B, C}, {B, C, D},...

* Linear Additivity:

Tổng của tất cả giá trị SHAP bằng tổng dự đoán của mô hình, đảm bảo tính minh bạch và giải thích được:

A white background with black text

Description automatically generated

#### 4.4.2.3. Vai trò trong bài toán:

* SHAP được sử dụng để giải thích mô hình Random Forest.
* Nó đo lường mức độ đóng góp của từng feature (MA, Volatility, Trend, Seasonality) vào các quyết định Buy, Sell, Hold.
* Biểu đồ SHAP giúp biết được:
  + Feature nào quan trọng nhất.
  + Feature nào có ảnh hưởng tích cực hay tiêu cực đến từng quyết định.

# CHƯƠNG 5: ĐÁNH GIÁ MÔ HÌNH

## 5.1. Mô hình Hồi quy tuyến tính

A screenshot of a calculator

Description automatically generated

Hình 43: Các chỉ số đánh giá của mô hình HQTT

Dưới đây là nhận xét về các chỉ số đánh giá của mô hình Hồi quy tuyến tính:

1. R² (Hệ số xác định): 0.9337

* Với giá trị 0.9337 (93.37%), mô hình giải thích rất tốt sự biến động của dữ liệu. Chỉ 6.63% phương sai chưa được mô hình giải thích.

2. MAE (Mean Absolute Error): 748.2008

* Sai số trung bình là 748.2 đơn vị. Đây là một mức sai số chấp nhận được trong bối cảnh dự báo giá cổ phiếu, nhưng có thể xem xét giảm thêm.

3. MSE (Mean Squared Error): 955060.3298

* Giá trị 955060.33 khá lớn, chủ yếu do sai số lớn được bình phương. Điều này là bình thường khi đơn vị giá cổ phiếu lớn. Tuy nhiên, MSE nên được kết hợp với RMSE để đánh giá chi tiết hơn.

4. RMSE (Root Mean Squared Error): 977.2719

* RMSE là 977.27, lớn hơn MAE (748.2), cho thấy tồn tại một số sai số lớn ảnh hưởng đến độ chính xác của mô hình.

5. MAPE (Mean Absolute Percentage Error): 0.0107

* Với MAPE chỉ 1.07%, sai số dự báo của mô hình là rất thấp. Giá trị này khẳng định độ chính xác cao của mô hình.

**Nhận xét tổng thể:**

Mô hình hồi quy tuyến tính thể hiện hiệu quả cao khi dự báo giá cổ phiếu, thể hiện qua R² = 93.37% và MAPE = 1.07%, cho thấy mô hình giải thích được phần lớn sự biến động và có độ sai lệch trung bình rất nhỏ.

Tuy nhiên, RMSE và MSE khá lớn, phản ánh sự ảnh hưởng của một số sai số lớn trong dự báo. Điều này cần cân nhắc thêm nếu có các ngoại lệ hoặc dữ liệu bất thường trong tập dữ liệu.

Với kết quả này, mô hình hồi quy tuyến tính đã hoạt động tốt, nhưng có thể cải thiện thêm bằng cách thử nghiệm các mô hình phức tạp hơn như XGBoost hoặc LSTM để so sánh kết quả.

## 5.2. Mô hình XGBOOST

**A screenshot of a graph

Description automatically generated**

Hình 44: Các chỉ số đánh giá của mô hình XGBoost

Dưới đây là nhận xét về các chỉ số đánh giá của mô hình XGBoost:

1. R² (Hệ số xác định): 0.8928

* Đây là một chỉ số cao, cho thấy mô hình giải thích được 89.28% phương sai của dữ liệu.
* Điều này chứng tỏ mô hình có khả năng dự đoán tốt và phù hợp với dữ liệu.

2. MAE (Mean Absolute Error): 922.4586

* Với MAE khoảng 922, sai số dự đoán trung bình nằm trong khoảng 922 đơn vị, tương đối nhỏ.

3. MSE (Mean Squared Error): 1542938.8583

* Với giá trị 1,542,938.8583, sai số này khá lớn so với khoảng giá từ 65,000 đến 90,000.

4. RMSE (Root Mean Squared Error): 1242.1509

* Với 1242, sai số này vẫn chấp nhận được trong bối cảnh dữ liệu chứng khoán.
* RMSE lớn hơn MAE một chút, điều này cho thấy có thể tồn tại một số điểm dự đoán lệch lớn (outliers).

5. MAPE (Mean Absolute Percentage Error): 0.0133

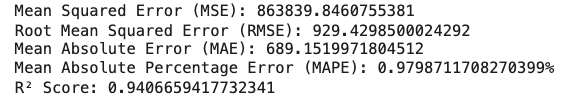
* Đây là một chỉ số rất tốt, cho thấy sai số trung bình chỉ chiếm khoảng 1.33% giá trị thực tế.
* Điều này chứng tỏ mô hình dự đoán có độ chính xác cao.

**Nhận xét tổng thể:**

Mô hình XGBoost hoạt động rất tốt với R² = 0.8928 và MAPE = 1.33%.

Sai số tuyệt đối và sai số bình phương ở mức chấp nhận được, phù hợp với dữ liệu dự đoán giá chứng khoán..

## 5.3. Mô hình LSTM



Hình 45: Các chỉ số đánh giả của mô hình LSTM

Dưới đây là nhận xét về các chỉ số đánh giá của mô hình LSTM:

1. Mean Squared Error (MSE):

* MSE = 863,839.85
* MSE càng nhỏ, mô hình càng tốt. Tuy nhiên, vì MSE sử dụng bình phương nên giá trị có thể lớn hơn các chỉ số khác.

2. Root Mean Squared Error (RMSE):

* RMSE = 929.43
* Với RMSE ~ 929, mô hình có sai số trung bình khoảng 929 đơn vị (tương ứng với đơn vị giá cổ phiếu). Đây là sai số tương đối nhỏ so với giá cổ phiếu thường có giá trị lớn (ví dụ, vài chục nghìn đồng).

3. Mean Absolute Error (MAE):

* MAE = 689.15
* So với RMSE, MAE ít bị ảnh hưởng bởi các outliers (giá trị dự đoán lệch nhiều so với thực tế). Giá trị 689.15 thể hiện sai số trung bình trong dự đoán khoảng 689 đơn vị, là một mức sai số thấp.

4. Mean Absolute Percentage Error (MAPE):

* MAPE = 0.9799%
* Kết quả 0.9799% nghĩa là mô hình dự đoán lệch trung bình chưa đến 1% so với giá trị thực tế, một mức độ sai số rất nhỏ và cho thấy mô hình có hiệu suất rất cao.

5. R² Score (R-squared):

* R² = 0.9407
* Mô hình đã giải thích được 94% sự biến động trong dữ liệu, một kết quả rất tốt cho bài toán dự đoán giá cổ phiếu.

## 5.4. Đánh giá và so sánh 3 mô hình

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **R2** | **MSE** | **MAE** | **RMSE** | **MAPE** |
| **Hồi quy tuyến tính** | 0.9337 | 748.2008 | 955060.3 | 977.2 | 1.1% |
| **XGBoost** | 0.8928 | 922.4586 | 1542938.8 | 1242.1 | 1.33% |
| **LSTM** | 0.94 | 863.8398 | 689151.9 | 929.4 | 0.97% |

Bảng 1: So sánh các chỉ số đánh giá của 3 mô hình

**Hiệu suất tổng thể**:

* + **LSTM** là mô hình tốt nhất khi xét tổng hợp các chỉ số. Mô hình này đạt:
    - **R² cao nhất (0.94)**.
    - **MAE, RMSE, và MAPE nhỏ nhất**, cho thấy dự đoán chính xác hơn và sai số thấp hơn.
  + **Hồi quy tuyến tính** có kết quả khá tốt nhưng vẫn không bằng LSTM, đặc biệt ở các chỉ số MAE và MAPE.
  + **XGBoost** kém hiệu quả nhất, với sai số lớn hơn đáng kể ở tất cả các chỉ số.

# CHƯƠNG 6: TRÍCH XUẤT THÔNG TIN TỪ DỮ LIỆU THỜI GIAN

6.1. Chỉ số MA 50 và MA 200

MA (Moving Average) là chỉ số trung bình động, thể hiện trung bình giá cổ phiếu trong một khoảng thời gian xác định.

* **MA 50** giúp xác định xu hướng **ngắn hạn đến trung hạn** của giá cổ phiếu.
* **MA 200** thể hiện xu hướng **dài hạn**, phản ánh sức khỏe chung của thị trường cổ phiếu.
* Khi **MA 50 cắt lên MA 200** → Tín hiệu mua (xu hướng tăng giá).
* Khi **MA 50 cắt xuống MA 200** → Tín hiệu bán (xu hướng giảm giá).

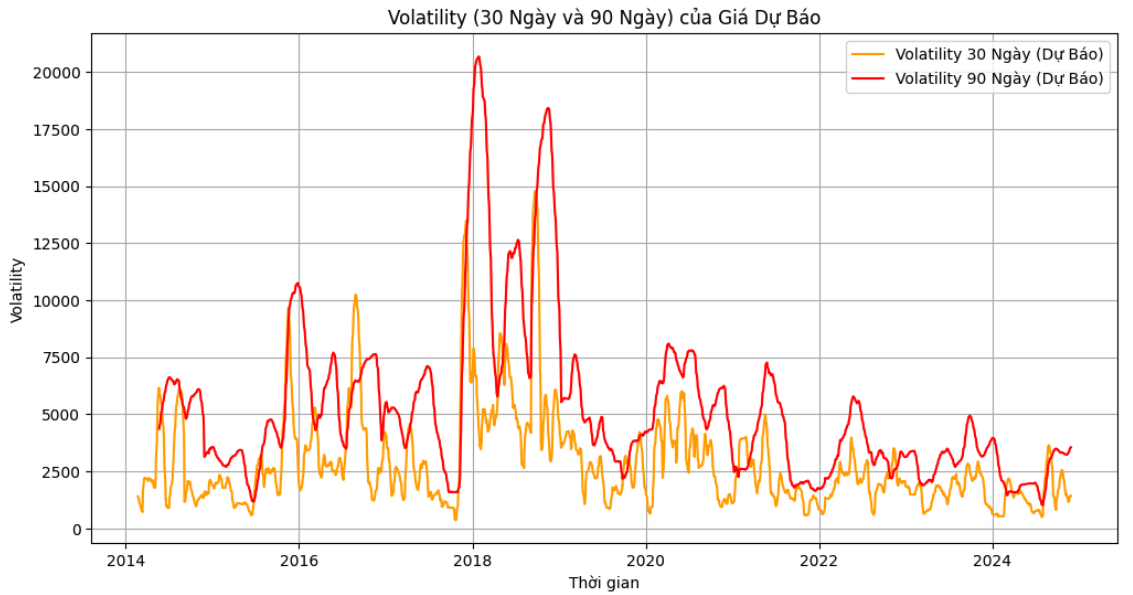


Hình 46: Biểu đồ đường của giá dự báo so với MA50 dự báo và MA200 dự báo

6.2. Chỉ số Valatility 30 và Volatility 90

Volatility (**Độ biến động**) đo lường mức độ dao động của giá cổ phiếu trong khoảng thời gian nhất định.

* **Volatility cao**: Giá cổ phiếu biến động mạnh → Rủi ro cao nhưng cũng tạo cơ hội lợi nhuận lớn.
* **Volatility thấp**: Giá cổ phiếu ít biến động → Thị trường ổn định, rủi ro thấp.
* So sánh **Volatility 30** và **Volatility 90** để đánh giá xu hướng biến động ngắn hạn và dài hạn.

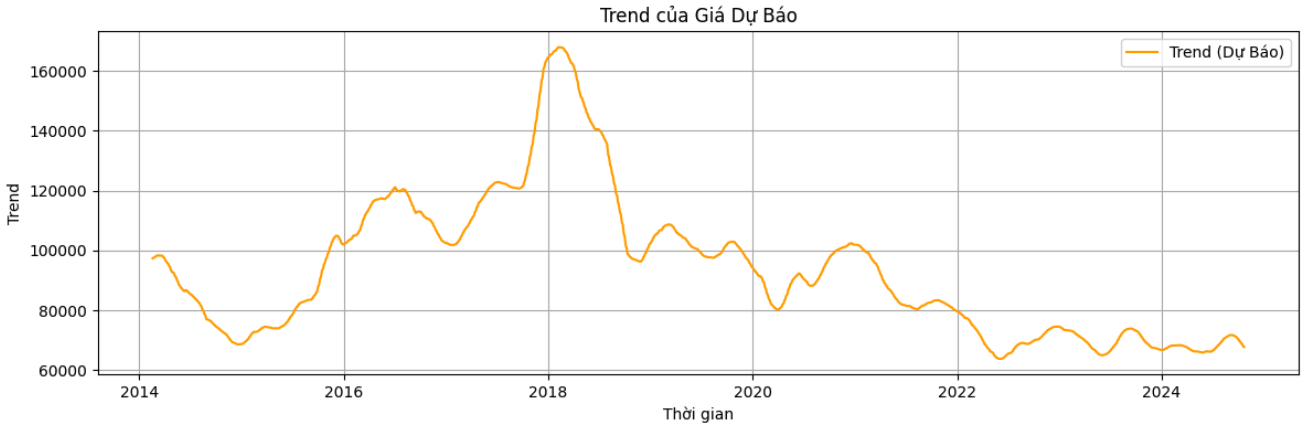


Hình 47: Biểu đồ đường thể hiện độ biến động 30 ngày và 90 ngày của giá dự báo

## 6.3. Time series Decomposition (Trend và Seasonal):

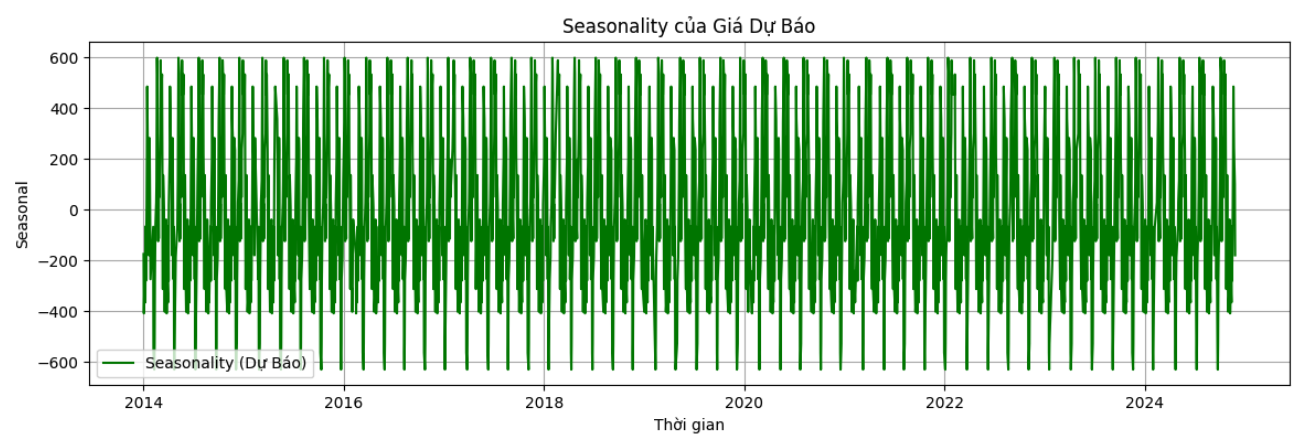
Là phương pháp phân tích chuỗi thời gian bằng cách tách nó thành các thành phần cơ bản:

* **Trend (Xu hướng):**
  + Xu hướng dài hạn thể hiện hướng đi của dữ liệu qua thời gian: **tăng, giảm hoặc ổn định**.
* Nếu **trend tăng ổn định** → Tín hiệu tích cực để đầu tư dài hạn.
* Nếu **trend giảm liên tục** → Thị trường đang suy yếu.



Hình 48: Biểu đồ thể hiện Trend của giá dự báo

* **Seasonal (Mùa vụ):**
  + Biến động lặp lại theo chu kỳ, thường do yếu tố thời vụ hoặc sự kiện có tính chất định kỳ (mùa lễ hội, chu kỳ kinh tế, v.v.).
* Giúp nhà đầu tư dự đoán được các thời điểm giá có khả năng tăng/giảm dựa trên tính mùa vụ.
* Có thể kết hợp seasonal với giá dự báo để tìm **điểm vào/thoát lệnh** tối ưu.



Hình 49: Biểu đồ thể hiện Seasonal của giá dự báo

# CHƯƠNG 7: BỘ HỖ TRỢ QUYẾT ĐỊNH

## 7.1. Quyết định mua (Mở vị thế mua)

Nhà đầu tư nên **mua vào** khi thỏa mãn các điều kiện sau:

* **Xu hướng (Trend)**:
  + Giá dự báo cho thấy xu hướng **tăng dần** trong dài hạn (Trend tích cực).
* **MA (Moving Averages)**:
  + Giá dự báo nằm **dưới MA 50 hoặc MA 200** nhưng đang có xu hướng tăng vượt lên.
    - Điều này phản ánh tín hiệu "cổ phiếu đang ở vùng chiết khấu" và chuẩn bị bứt phá.
* **Volatility (Độ biến động)**:
  + **Volatility 30** giảm thấp hơn **Volatility 90**, cho thấy rủi ro ngắn hạn đang giảm.
* **Seasonality**:
  + Dữ liệu mùa vụ cho thấy đây là thời điểm giá cổ phiếu thường có xu hướng tăng (ví dụ: vào cuối quý hoặc cuối năm).

## 7.2. Quyết định bán (Vị thế hoặc chốt lời)

Nhà đầu tư nên **bán ra** khi thỏa mãn các điều kiện sau:

* **Xu hướng (Trend)**:
  + Giá dự báo có dấu hiệu **giảm dần** trong dài hạn (Trend tiêu cực).
* **MA (Moving Averages)**:
  + Giá dự báo đã vượt **MA 50 hoặc MA 200** và bắt đầu có dấu hiệu giảm.
    - Đây là dấu hiệu cổ phiếu có thể đạt đỉnh và sắp giảm giá.
* **Volatility (Độ biến động)**:
  + **Volatility 30** tăng mạnh hơn **Volatility 90**, cho thấy rủi ro ngắn hạn gia tăng và thị trường không ổn định.
* **Seasonality**:
  + Dữ liệu mùa vụ cho thấy đây là thời điểm giá cổ phiếu có xu hướng giảm (ví dụ: giai đoạn thấp điểm kinh doanh).

## 7.3. Quyết định giữ (không bán, cũng không mua thêm)

* **Xu hướng giá (Trend)**: **Tăng nhẹ hoặc đi ngang ổn định**.
  + Không có dấu hiệu rõ ràng của xu hướng giảm.
  + Xu hướng dài hạn vẫn tích cực, nhưng ngắn hạn không có biến động mạnh.
* **Giá nằm giữa MA 50 và MA 200**: **Chưa thấy tín hiệu rõ ràng**
  + Giá chưa cắt lên **MA 50** (tín hiệu mua), nhưng cũng chưa cắt xuống **MA 200** (tín hiệu bán).
* **Volatility 30 ≈ Volatility 90**: **thấp và ổn định**
  + Độ biến động trong ngắn hạn và dài hạn không có nhiều khác biệt.
  + Thị trường đang trong trạng thái **ổn định** và ít rủi ro.
* **Tính mùa vụ (Seasonality) chưa đến thời điểm cao hoặc thấp**
  + Mùa vụ giá chưa có dấu hiệu tăng mạnh (cuối năm, quý 3...) hoặc giảm mạnh (đầu năm sau Tết).

|  |  |  |  |
| --- | --- | --- | --- |
| **Yếu tố** | **Quy tắc MUA** | **Quy tắc BÁN** | **Quy tắc GIỮ** |
| **Xu hướng (Trend)** | Trend tích cực (tăng dần). | Trend tiêu cực (giảm dần). | Xu hướng giá đi ngang hoặc tăng nhẹ, không có dấu hiệu giảm mạnh. |
| **MA 50 & MA 200** | Giá dưới MA 50/200 nhưng có xu hướng cắt lên. | Giá trên MA 50/200 và có dấu hiệu cắt xuống. | Giá nằm giữa MA 50 và MA 200, chưa có tín hiệu rõ ràng. |
| **Volatility** | Volatility 30 < Volatility 90 (rủi ro giảm). | Volatility 30 > Volatility 90 (rủi ro tăng). | Volatility 30 ≈ Volatility 90 → Biến động thấp, thị trường ổn định. |
| **Seasonality** | Thời điểm mùa vụ giá có xu hướng tăng. | Thời điểm mùa vụ giá có xu hướng giảm | |  | | --- | |  | | Chưa đến giai đoạn cao điểm hoặc giảm giá của mùa vụ. | |

Bảng 2: Quy tắc mua, bán và giữ

## 7.4. Trọng số của các thuộc tính trong các quyết định mua, bán và giữ

### 7.4.1. Quyết định bán

A graph of blue rectangular objects

Description automatically generated with medium confidence

Hình 50: Trọng số của các thuộc tính cho quyết định bán

**MA (Moving Average):** Là yếu tố quan trọng nhất trong quyết định bán cổ phiếu, với trọng số 0.213560. Điều này có nghĩa là xu hướng giá trung bình dài hạn (MA50 và MA200) đóng vai trò lớn trong việc xác định khi nào nên bán.

**Volatility:** Cũng có mức ảnh hưởng cao (0.208653), cho thấy biến động giá cổ phiếu trong ngắn hạn là yếu tố quan trọng để quyết định bán khi thị trường có biến động mạnh.

**Trend:** Có ảnh hưởng rất thấp (0.007083), nghĩa là chênh lệch giữa giá dự báo và giá hiện tại không phải là yếu tố quyết định chính trong việc bán cổ phiếu.

### 7.4.2. Quyết định giữ

**A graph of blue rectangular objects

Description automatically generated with medium confidence**

Hình 51: Trọng số của các thuộc tính cho quyết định giữ

**Volatility:** Là yếu tố quan trọng nhất (0.019801) khi quyết định giữ cổ phiếu. Điều này có nghĩa là mức độ biến động thấp có thể khiến nhà đầu tư tiếp tục giữ cổ phiếu.

**MA (Moving Average):** Có trọng số 0.014179, cho thấy xu hướng dài hạn vẫn có ảnh hưởng nhất định trong việc quyết định giữ cổ phiếu.

**Trend:** Có mức ảnh hưởng thấp nhất (0.006031), cho thấy chênh lệch giữa giá dự báo và giá hiện tại không phải là yếu tố chính để quyết định giữ cổ phiếu.

### 7.4.3. Quyết định bán

**A graph of blue rectangular objects

Description automatically generated with medium confidence**

Hình 52: Trọng số của các thuộc tính cho quyết định mua

**Trend:** Là yếu tố quan trọng nhất (0.238690) trong quyết định mua cổ phiếu. Điều này cho thấy chênh lệch giữa giá dự báo và giá hiện tại đóng vai trò lớn trong việc nhận diện cơ hội mua.

**Volatility:** Có trọng số cao thứ hai (0.226057), cho thấy biến động giá cổ phiếu cũng được cân nhắc nhiều khi quyết định mua, đặc biệt trong thị trường biến động mạnh.

**MA (Moving Average):** Có mức ảnh hưởng thấp nhất (0.015461), nghĩa là xu hướng trung bình không phải là yếu tố quan trọng khi ra quyết định mua cổ phiếu.

# CHƯƠNG 8: KẾT LUẬN

Qua quá trình nghiên cứu và phân tích dữ liệu giá cổ phiếu Vinamilk (VNM) từ năm 2014 đến 2024, đề tài đã lựa chọn được mô hình dự báo tối ưu nhất là LSTM. Đây là mô hình đạt hiệu suất vượt trội với R² cao nhất (0.94), cùng các chỉ số MAE, RMSE và MAPE nhỏ nhất, cho thấy khả năng dự báo chính xác và ổn định hơn so với các phương pháp khác. LSTM chứng minh được ưu thế trong việc xử lý dữ liệu chuỗi thời gian phức tạp, đặc biệt là các mối quan hệ phi tuyến và sự biến động của giá cổ phiếu Vinamilk. Mô hình này không chỉ giúp dự báo sát thực tế mà còn có khả năng ghi nhớ mối quan hệ dài hạn, từ đó cung cấp những tín hiệu dự báo tin cậy.

Dựa trên kết quả dự báo từ mô hình LSTM, nhóm đã đưa ra được bộ quy tắc đầu tư nhằm hỗ trợ nhà đầu tư ra quyết định. Bộ quy tắc đề xuất các tín hiệu rõ ràng về việc mua, bán hoặc giữ cổ phiếu Vinamilk dựa trên xu hướng biến động giá và các yếu tố thị trường. Quy trình này giúp tối ưu hóa lợi nhuận và giảm thiểu rủi ro đầu tư, từ đó hỗ trợ nhà đầu tư đưa ra những quyết định hợp lý và mang lại hiệu quả tài chính tốt nhất. Mô hình và quy tắc đầu tư này có tiềm năng ứng dụng rộng rãi trong các bài toán tài chính tương tự, góp phần nâng cao hiệu quả và sự bền vững của các quyết định đầu tư trong tương lai.

*TÀI LIỆU THAM KHẢO*

[1] A. Hayes, "Volatility: Meaning in Finance and How It Works With Stocks," *Investopedia*, 2003.

[2] "Golden Cross và Death Cross là gì? Ý nghĩa Điểm cắt vàng, Điểm cắt tử thần trong giao dịch chứng khoán," *DSC*, 2023.

[3] "Time Series Decomposition Techniques," *GeeksforGeeks*, 2023.

[4] M. Peixeiro, "The Complete Guide to Time Series Models," *Built In*, 2023.

[5] S. Pandian, "Time Series Analysis: Definition, Components, Methods, and Applications," *Analytics Vidhya*, 2024.

[6] "Step by Step Guide to Modeling Time Series Data Using Linear Regression," *GeeksforGeeks*, 2023.

[7] R. Adusumilli, "Machine Learning to Predict Stock Prices," *Towards Data Science*, 2019.

[8] Y. Zhang, "Stock Price Prediction Method Based on XGBoost Algorithm," *Lanzhou University of Technology, Qi Lihe, Lan Zhou, China*, 2023.

[9] C. Molnar, "SHAP (SHapley Additive exPlanations)," *Interpretable Machine Learning*, 2023.

[10] "sklearn.ensemble.RandomForestClassifier," *scikit-learn 1.5.2 documentation*, 2023.