1. Giới thiệu dữ liệu

Trong những năm gần đây Bitcoin đã nổi lên mạnh mẻ như một biểu tượng của sự đổi mới trong lĩnh vực tài chính và công nghệ. Sự biến động mạnh mẽ của giá trị Bitcoin không chỉ thu hút sự chú ý của các nhà đầu tư mà còn là chủ đề nghiên cứu của nhiều chuyên gia kinh tế và khoa học dữ liệu. Việc hiểu rõ những yếu tố ảnh hưởng đến giá Bitcoin là chìa khóa để đưa ra các quyết định giao dịch sáng suốt và tối ưu hóa lợi nhuận. Để làm được điều này, dữ liệu lịch sử về Bitcoin đóng vai trò vô cùng quan trọng. Nó cung cấp một cái nhìn toàn diện về xu hướng biến động giá, khối lượng giao dịch và các chỉ số liên quan qua từng ngày. Bộ dữ liệu mà chúng tôi đang sử dụng chính là nền tảng để xây dựng các mô hình phân tích và dự đoán trong tương lai.

Bộ dữ liệu Bitcoin đã được thu thập từ trang web (<https://coinmarketcap.com/currencies/bitcoin/historical-data/>) ghi lại chi tiết về sự thay đổi giá cả, khối lượng giao dịch từ ngày 02/01/2018 đến 02/01/2015. Bộ dữ liệu có 2558 quan trắc, gồm 7 biến:

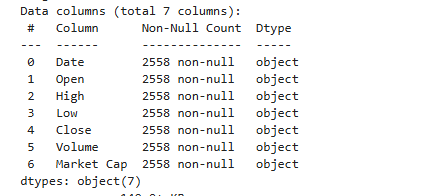
1. Date – Ngày giao dịch của Bitcoin
2. Open - Giá mở cửa của Bitcoin trong ngày
3. High - Giá cao nhất của Bitcoin trong ngày
4. Low - Giá thấp nhất của Bitcoin trong ngày
5. Close - Giá đóng cửa của Bitcoin trong ngày
6. Volume – Khối lượng giao dịch của Bitcoin trong ngày
7. Market Cap – Tổng vốn hóa thị trường của Bitcoin tại thời điểm giao dịch.



Hình : Một vài quan trắc và kích thước của bộ dữ liệu ban đầu

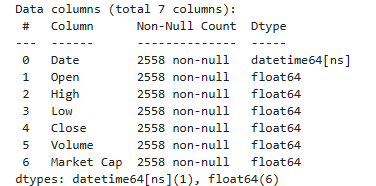
Bộ dữ liệu Bitcoin cung cấp cái nhìn chi tiết về sự biến động của giá và khối lượng giao dịch qua từng ngày. Với 2558 quan sát và 7 cột, bồ dữ liệu cho phép phân tích theo xu hướng dài hạn, giúp nhận diện sự thay đổi giá cả qua thời gian. Biến **Open**, **High**, **Low** và **Close** đại diện cho các mốc giá quan trọng trong một phiên giao dịch, phản ánh rõ ràng sự dao động giá của Bitcoin. Khối lượng giao dịch (**Volume**) là chỉ số thể hiện mức độ sôi động của thị trường, **Market Cap** cho biết tổng giá trị của Bitcoin đang lưu thông tại thời điểm đó. Thông qua việc phân tích các biến này, chúng ta có thể phát hiện các xu hướng, đánh giá rủi ro và dự đoán giá trong tương lai.

1. Phân tích và xử lý dữ liệu:



Hình 2.1: Kiểu dữ liệu sau khi kiểm tra là ‘object’

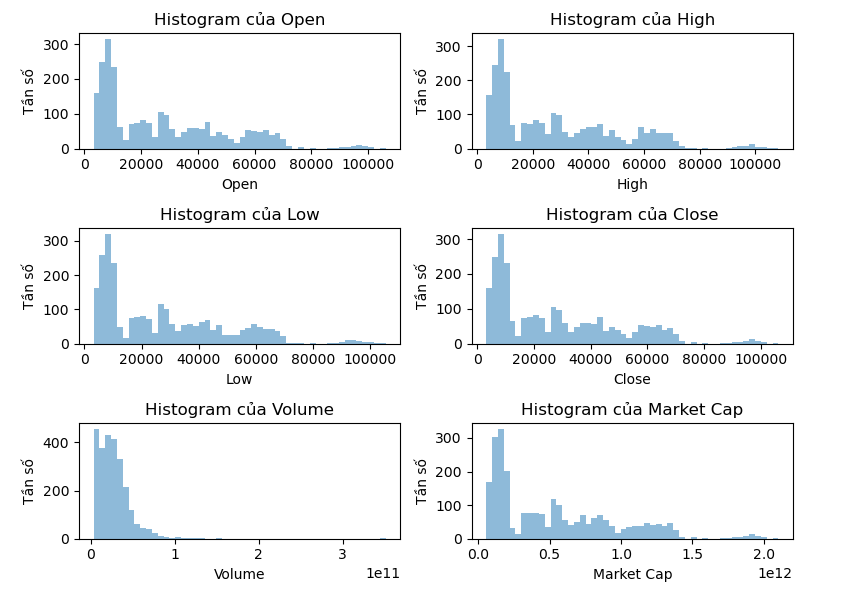
Khi kiểm tra dữ liệu ban đầu, có thể nhận thấy tất cả các cột trong bộ dữ liệu đều thuộc kiểu ‘object’. Điều này chỉ ra rằng các giá trị trong các cột **Open, High, Low, Close, Volume** và **Market Cap** đang được lưu dưới dạng văn bản thay vì dạng số. Việc để dữ liệu ở dạng chuỗi sẽ gây ra khó khăn trong các phép tính toán và trực quan hóa. Nên em đã chuyển đổi sang kiểu số cũng như loại bỏ ký tự ‘$’ và dấu phẩy ‘ , ’, kết quả được thể hiện ở các hình dưới đây.



Hình 2.2: Sau khi chuyển đổi dạng chuỗi về dạng số



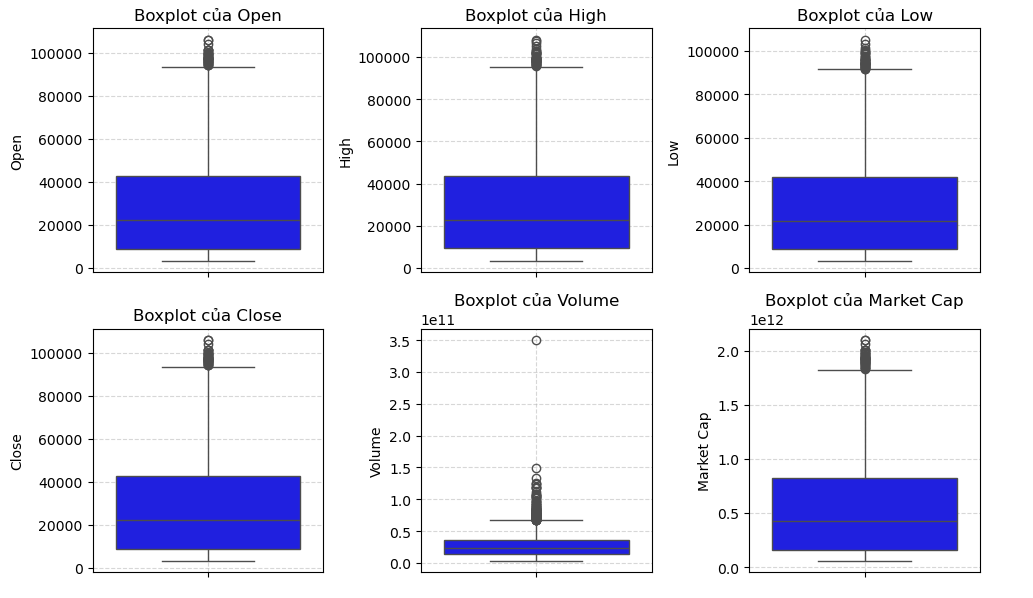
Hình 2.3: Dữ liệu sau khi loại bỏ ký tự và dấu phẩy



Hình 2.4: Quan sát phân bố của từng biến

Quan sát hình 2.4 có nhận xét sau:

* Open, High, Low, Close: các biến liên quan đến giá đều có phân phối lệch phải. Điều này cho thấy phần lớn giá Bitcoin giao dịch ở mức thấp, nhưng có những thời điểm giá tăng cao đột biến, thể hiện qua phần đuôi kéo dài của biểu đồ. Các mức giá vượt quá 60,000 USD có tần suất thấp, chứng tỏ các phiên giao dịch đạt đỉnh không phổ biến.
* Volume: phân phối khối lượng giao dịch cũng có đuôi dài, phản ánh một số giao dịch có khối lượng lớn hơn nhiều so với phần còn lại. Phần lớn khối lượng giao dịch nằm ở mức thấp dưới 50 tỷ USD, nhưng có những điểm vượt quá 100 tỷ USD, thể hiện sự tham gia của các nhà đầu tư tổ chức hoặc các đợt biến động lớn trên thị trường.
* Market Cap: phần lớn tập trung dưới 500 tỷ USD, một số điểm dữ liệu cho thấy vốn hóa (Market Cap) vượt quá 1 nghìn tỷ USD, phản ánh các giai đoạn Bitcoin bùng nổ và chiếm lĩnh thị trường tiền điện tử.



Hình 2.4: Boxplot kiểm tra outliers của các biến

Quan sát hình 2.4 có nhận xét sau:

* Open, High, Low, Close: phần lớn các giá trị tập trung trong khoảng nhất định, nhưng có nhiều điểm ngoại lai phía trên, phản ánh các giai đoạn tăng giá đột biến của Bitcoin
* Volume, Market Cap: cũng có sự phân tán lớn, với nhiều điểm vượt xa phần còn lại, cho thấy sự xuất hiện của các phiên giao dịch bất thường hoặc các đợt tăng trưởng mạnh

Từ hình 2.3 và hình 2.4 ta có thể kết luận cho các biến như sau:

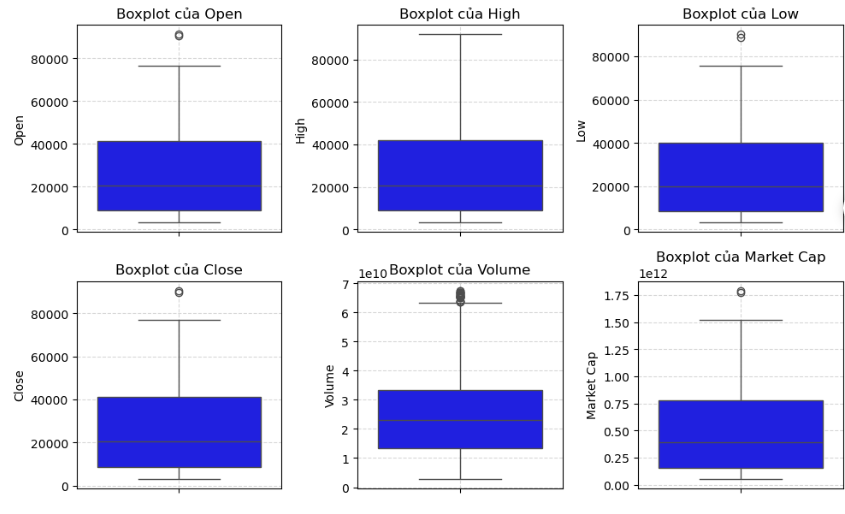
* Open, High, Low, Close: có xu hướng tập trung ở mức thấp trong phần lớn thời gian, nhưng cả hai biểu đồ đều cho thấy sự tăng giá đột biến trong một số giai đoạn, được thể hiện qua các điểm ngoại lai. Điều này phản ánh đặc điểm của thị trường Bitcoin thường xuyên biến động mạnh với những thời kỳ tăng trưởng bùng nổ.
* Volume và Market Cap: có sự phân bố không đồng đều. Các giá trị thấp chiếm đa số, trong khi các điểm cao vượt trội cho thấy có những phiên giao dịch lớn, có thể do sự tham gia của các tổ chức hoặc các sự kiện tác động mạnh đến thị trường.
* Sự tồn tại của nhiều outliers trong cả hai biểu đồ nhấn mạnh tính không ổn định và biến động cao của thị trường Bitcoin. Điều này vừa là rủi ro nhưng cũng là cơ hội lớn cho các nhà đầu tư trong các giai đoạn tăng giá mạnh.

Để đảm bảo dữ liệu ổn định và giảm thiểu sự ảnh hưởng của các giá trị bất thường, phương pháp IQR (Interquartile Range) đã được sử dụng để loại bỏ outliers. Các giá trị được xem là ngoại lai và bị loại bỏ khỏi dữ liệu khi nằm ngoài khoảng:

[Q1 – 1.5 \* iQR, Q3 + 1.5 \* IQR]

Trong đó:

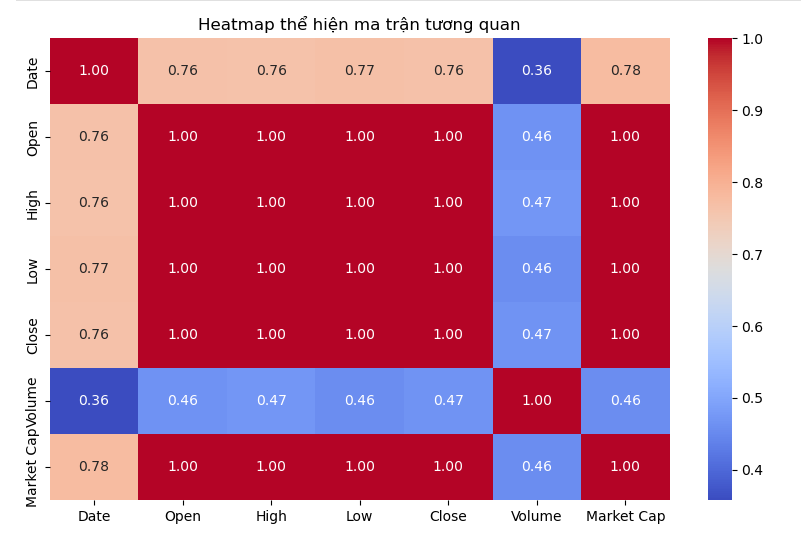
* Q1 (phân vị 25%) và Q3 (phân vị 75%) là các điểm chia dữ liệu thành các phần tư.
* IQR (Interquartile Range) là sự chênh lệch giữa Q3 và Q1, thể hiện độ phân tán của 50% dữ liệu giữa hai phân vị này.



Hình 2.4: Boxplot các biến sau khi loại bỏ outliers

Sau khi xử lý ngoại lai (outliers), các biểu đồ Boxplot (hình 2.4) cho thấy dữ liệu của các biến Open, High, Close và Market Cap đã trở nên ổn định hơn, ít điểm ngoại lai hơn.

Volume có sự phân tán mạnh hơn với một vài điểm ngoại lai, cho thấy sự biến động lớn về khối lượng giao dịch trong một số giai đoạn.



Hình 2.5: Heatmap thể hiện ma trận tương quan

Ma trận tương quan cho thấy:

* Các biến Open, High, Low, Close có mức độ tương quan rất cao, cho thấy sự liên quan chặt chẽ giữa các biến giá trong mỗi phiên giao dịch. Khi giá mở cửa (Open) tăng, giá cao (High) tăng, giá thấp (Low) và giá đóng cửa (Close) cũng có xu hướng tăng theo, cho thấy có hiện tượng đa cộng tuyến
* Volume có tương quan thấp so với các biến còn lại, đặc biệt là với Date, cho thấy khối lượng giao dịch không phải lúc nào cũng diễn ra đồng bộ với biến động giá hoặc thời gian.
* Market Cap: có mức độ tương quan manh với tất cả các biến giá, điều này phản ánh sự ảnh hưởng trực tiếp của giá Bitcoin đến vốn hóa thị trường. Khi giá Bitcoin tăng hoặc giảm, vốn hóa thị trường thay đổi tương ứng.

Để giải quyết vấn đề đa cộng tuyến và giảm bớt sự dư thừa thông tin giữa các biến có tương quan cao, em đã sử dụng PCA (Principal Component Analysis), phương pháp này không chỉ giảm chiều dữ liệu mà còn giữ lại được thông tin quan trọng, giúp mô hình trở nên ổn định và hiểu quả hơn trong việc dự đoán.

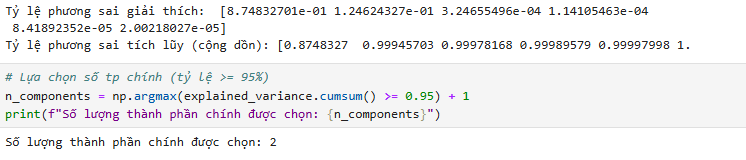


Hình 2.6 : Kết quả tỷ lệ phương sai giải thích và tỷ lệ phương sai tích lũy

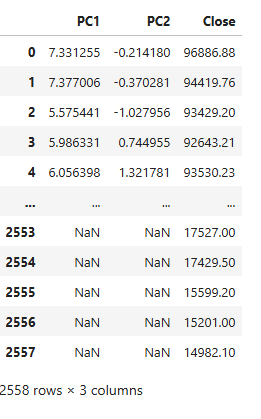
Các giá trị tỷ lệ phương sai giải thích cho thấy mức độ quan trọng của từng thành phần chính. Cụ thể, các thành phần chính đầu tiên giữ lại phần lớn sự biên thiên trong dữ liệu, trong khi các thành tiếp theo đóng góp ít hơn. Tỷ lệ phương sai giải thích của các thành phần đầu tiên là rất cao, cho thấy hiệu quả của PCA trong việc rút gọn dữ liệu mà vẫn duy trì được tính đại diện của dữ liệu gốc.

Tỷ lệ phương sai tích lũy cung cấp tổng cộng tỷ lệ phương sai giải thích của các thành phần chính, giúp đánh giá sự đóng góp của các thành phần vào việc giữ lại thông tin của dữ liệu.

Tiếp theo, mã lệnh kiểm tra tổng phương sai tích lũy và lựa chọn số lượng thành phần chính sao cho tổng phương sai tích lũy đạt ít nhất 95% (Hình 2.7). Điều này đảm bảo rằng 95% thông tin quan trọng của dữ liệu được giữ lại trong mô hình, nhưng chỉ với một số ít các thành phần chính. Số thành phần chính được chọn có kết quả là 2

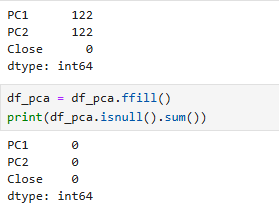


Hình 2.7: Số thành phần chính được chọn

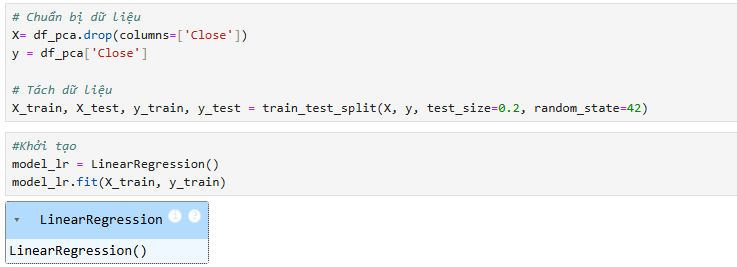


Hình 2.8: Kết quả sau khi giảm chiều dữ liệu với PCA

Sau khi giảm chiều dữ liệu, kết quả của PCA được kết hợp với biến Close từ dữ liệu gốc. Tạo thành một DataFrame mới (df\_pca) với ba cột: PC1, PC2, và Close (hình 2.8). Khi kiểm tra dữ liệu thiếu (hình 2.9), ta nhận thấy giá trị trong cột PC1 và PC2 có 122 giá trị thiếu. Để xử lý vấn đề này, phương pháp *ffill()* (forward fill) được sử dụng để thay thế các giá trị NaN bằng giá trị không bị thiếu từ dòng trước đó. Sau khi áp dụng *ffill()*, dữ liệu không còn thiếu giá trị (hình2.9).



Hình 2.9: Kết quả kiểm tra dữ liệu thiếu



Hình 2.10: Chuẩn bị dữ liệu và huấn luyện mô hình hồi quy tuyến tính

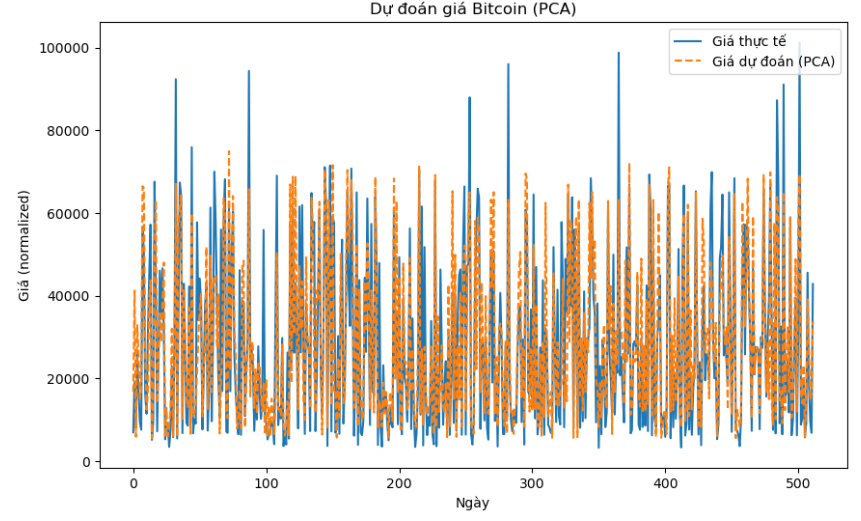
Dữ liệu được chuẩn bị bằng cách loại bỏ biến Close từ DataFrame df\_pca để tạo ra X (các đặc trưng) và y (biến mục tiêu là Close). Biến X sẽ bao gồm các thành phần chính (PC1, PC2) từ PCA. Dữ liệu được chia thành hai phần: tập huấn luyện (80% dữ liệu) và tập kiểm tra (20% dữ liệu) thông qua hàm *train\_test\_split()*. Điều này giúp mô hình học từ tập huấn luyện và được đánh giá trên tập kiểm tra. Một mô hình hồi quy tuyến tính (LinearRegression()) được khởi tạo và huấn luyện bằng cách sử dụng tập huấn luyện (X\_train, y\_train).



Hình 2.10: Kết quả đánh giá mô hình mô quy tuyến tính

Sau khi chạy mô hình ta có các kết quả như sau:

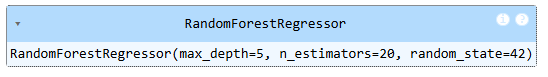
* MAE (Mean Absolute Error): trung bình độ sai số tuyệt đối, cho thấy mức độ sai lệch trung bình giữa các dự đoán và giá tri thực tế là khoảng 7895.89
* RMSE (Root Mean Squared Error): Căn bậc hai của sai số bình phương trung bình, dùng để đo mức độ chính xác của mô hình với độ nhạy cảm với sai số lớn, có mức sai số căn bậc hai là 11077.21, cho thấy mô hình có độ sai lệch khá lớn, nhưng vẫn có thể chấp nhận được.
* R² (R-squared): Chỉ số xác định độ phù hợp của mô hình, cho biết tỷ lệ biến động của dữ liệu được giải thích bởi mô hình khoảng 73% biến động trong giá Close, đây là một mức độ tương đối tốt.



Hình 2.11: So sánh giá thực tế và giá dự đoán Bitcoin bằng mô hình hồi quy tuyến tính

Biểu đồ so sánh giữa giá trị thực tế và giá dự đoán có thể thấy:

* Giá thực tế (màu xanh dương) có sự biến động lớn, phản ánh các thay đổi mạnh mẽ trong giá trị của Bitcoin trong suốt thời gian.
* Giá dự đoán (đường nét đứt màu cam) theo xu hướng cũng biến động khá gần với giá thực tế, tuy nhiên, có sự chệch lệch nhất định trong một số khoảng thời gian, đặc biệt là vào những đỉnh cao hoặc đáy sâu của giá.



Hình 2.12: Khởi tạo mô hình Random Forest

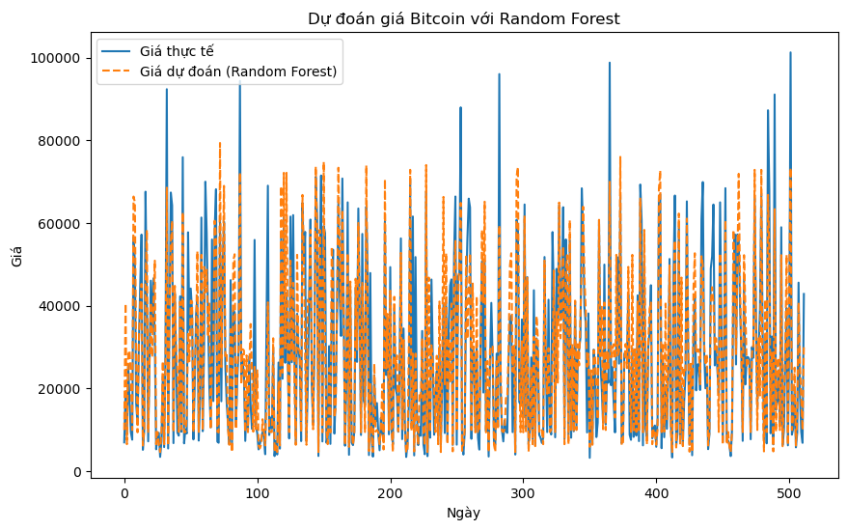
Em đã thực hiện thêm mô hình Random Forest Regressor (hình 2.12) để cải thiện độ chính xác trong việc dự đoán giá Bitcoin. Mô hình Random Forest được khởi tạo với 20 cây quyết định (n\_estimators=20) và độ sâu của mỗi cây là 5 (max\_depth=5). Các tham số này giúp hạn chế việc quá khớp (overfitting) và đảm bảo mô hình không quá phức tạp.



Hình 2.13: Kết quả đánh giá mô hình Random Forest

Mô hình Random Forest có kết quả như sau:

* MAE (Mean Absolute Error): trung bình độ sai số tuyệt đối, cho thấy mức độ sai lệch trung bình giữa các dự đoán và giá tri thực tế là khoảng 6450,71
* RMSE (Root Mean Squared Error): Căn bậc hai của sai số bình phương trung bình, dùng để đo mức độ chính xác của mô hình với độ nhạy cảm với sai số lớn, có mức sai số căn bậc hai là 9638.96, cho thấy mô hình có độ sai lệch đáng kể, nhưng vẫn có thể chấp nhận được.
* R² (R-squared): Chỉ số xác định độ phù hợp của mô hình, cho biết tỷ lệ biến động của dữ liệu được giải thích bởi mô hình khoảng 80% biến động trong giá Close, đây là một mức độ tương tốt.



Hình 2.14: So sánh giá thực tế và giá dự đoán Bitcoin bằng mô hình Random Forest

Biểu đồ trên so sánh giá thực tế và giá dự đoán sử dụng mô hình Random Forest:

* Giá thực tế (màu xanh dương) cho thấy sự biến động mạnh mẽ, phản ánh các dao động không thể đoán trước của giá Bitcoin trong thời gian. Các đỉnh và đáy của giá thường xuyên thay đổi.
* Giá dự đoán (đường nét đứt màu cam) từ mô hình Random Forest có xu hướng bám sát khá gần giá thực tế, nhưng vẫn có một số sai lệch ở các đỉnh và đáy. Mô hình dự đoán có thể không bắt kịp chính xác những biến động lớn trong một số khoảng thời gian, nhưng nhìn chung vẫn thể hiện xu hướng chung của giá Bitcoin.

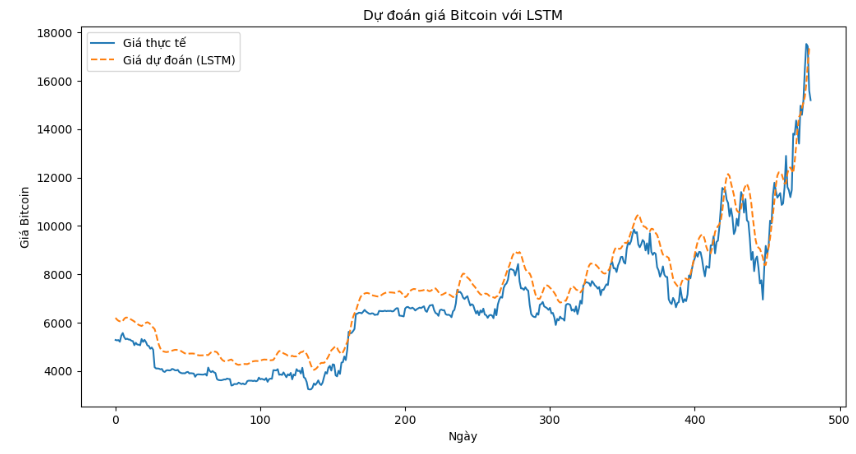
Mô hình Random Forest đã có cải thiện hơn so với mô hình hồi quy tuyến tính, nhưng để tiếp tục nâng cao độ chính xác và khả năng dự đoán với dữ liệu chuỗi thời gian như giá Bitcoin, em đã thử áp dụng mô hình LSTM (Long Short-Term Memory). Đây là một mô hình học sâu, thuộc nhóm mạng nơ-ron hồi quy (RNN), chuyên dụng cho dữ liệu có tính tuần tự và phụ thuộc vào quá khứ, giúp mô hình học và dự đoán các xu hướng trong chuỗi thời gian một cách hiệu quả hơn. Kết quả đã có cải thiện so với hai mô hình trước đó (hình 2.15).



Hình 2.15: Kết quả đánh giá mô hình LSTM

Mô hình LSTM có kết quả như sau:

* MAE (Mean Absolute Error): trung bình độ sai số tuyệt đối, cho thấy mức độ sai lệch trung bình giữa các dự đoán và giá tri thực tế là khoảng 768.45, đây là một sai số chấp nhận được.
* RMSE (Root Mean Squared Error): Căn bậc hai của sai số bình phương trung bình, dùng để đo mức độ chính xác của mô hình với độ nhạy cảm với sai số lớn, có mức sai số căn bậc hai là 836.66, cho thấy mô hình có sai lệch nhỏ nhưng vẫn có thể cải thiện hơn nữa.
* R² (R-squared): Chỉ số xác định độ phù hợp của mô hình, cho biết tỷ lệ biến động của dữ liệu được giải thích bởi mô hình khoảng 89% biến động trong giá Close, cho thấy mô hình rất tốt trong việc dự đoán giá Bitcoin.



Hình 2.16: So sánh giá thực tế và giá dự đoán Bitcoin bằng mô hình LSTM

Biểu đồ trên so sánh giá thực tế và giá dự đoán sử dụng mô hình LSTM:

* Giá thực tế (màu xanh dương) có sự biến động mạnh mẽ, đặc biệt là những đợt tăng giá đột ngột về cuối biểu đồ, phản ánh các biến động khó lường của thị trường Bitcoin.
* Giá dự đoán (đường nét đứt màu cam) từ mô hình LSTM theo xu hướng bám sát giá thực tế khá chặt chẽ, nhưng vẫn có một số sai lệch ở những điểm cực trị và sự biến động mạnh trong một số giai đoạn.

Qua ba mô hình trên thì mô hình LSTM là mô hình hiệu quả nhất với khả năng dự đoán cao nhất và sai số thấp nhất. Mô hình này nắm được xu hướng biến động của giá Close của Bitcoin rất tốt, đặc biệt là trong các giai đoạn biến động mạnh.