




PHÂN TÍCH GIÁ VÀNG THẾ GIỚI

Thực hiện: Phạm Hải Thanh



AGENDA

Giới thiệu đề tài

Các bước thực hiện

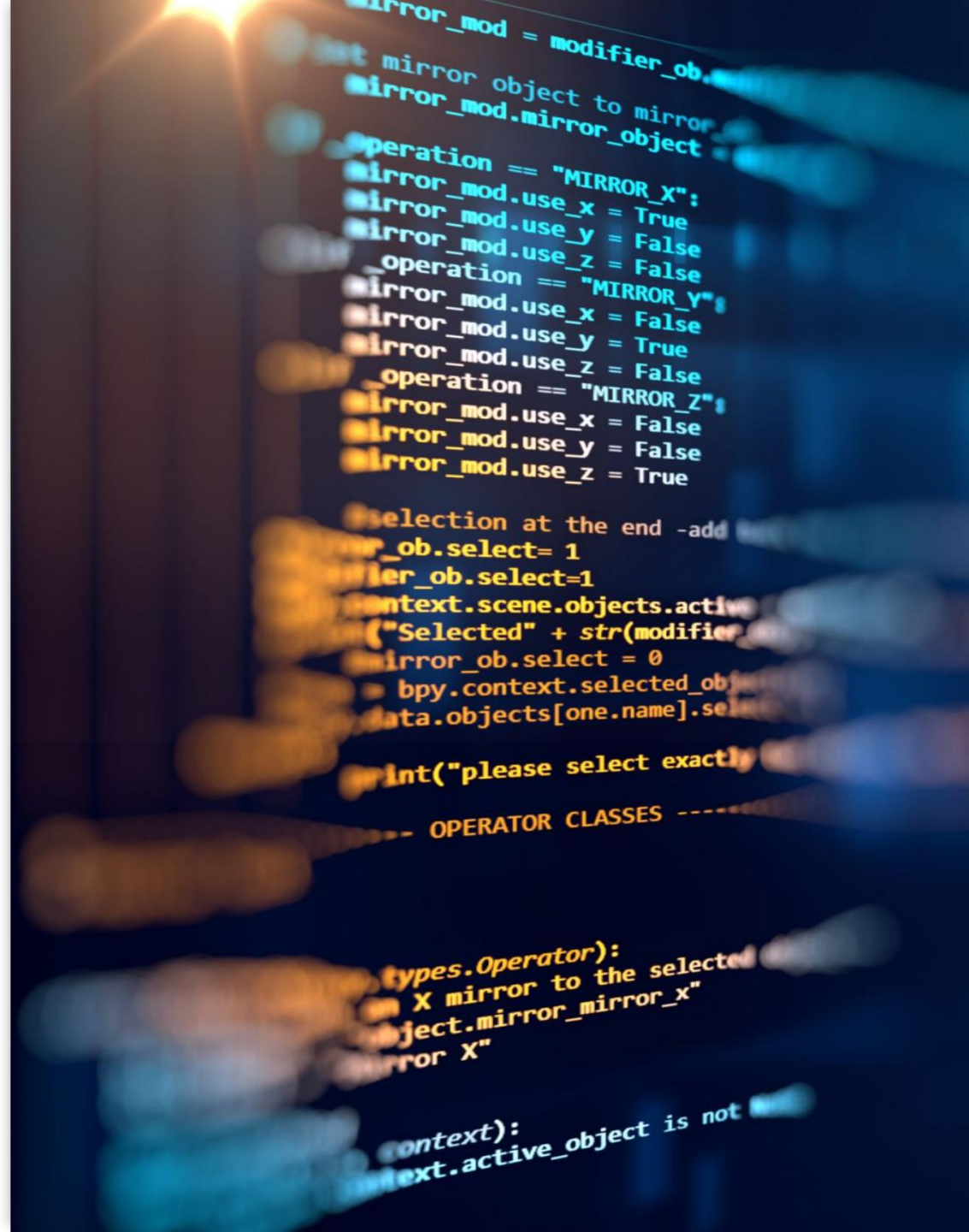
Mô tả bộ dữ liệu

Phân tích dữ liệu

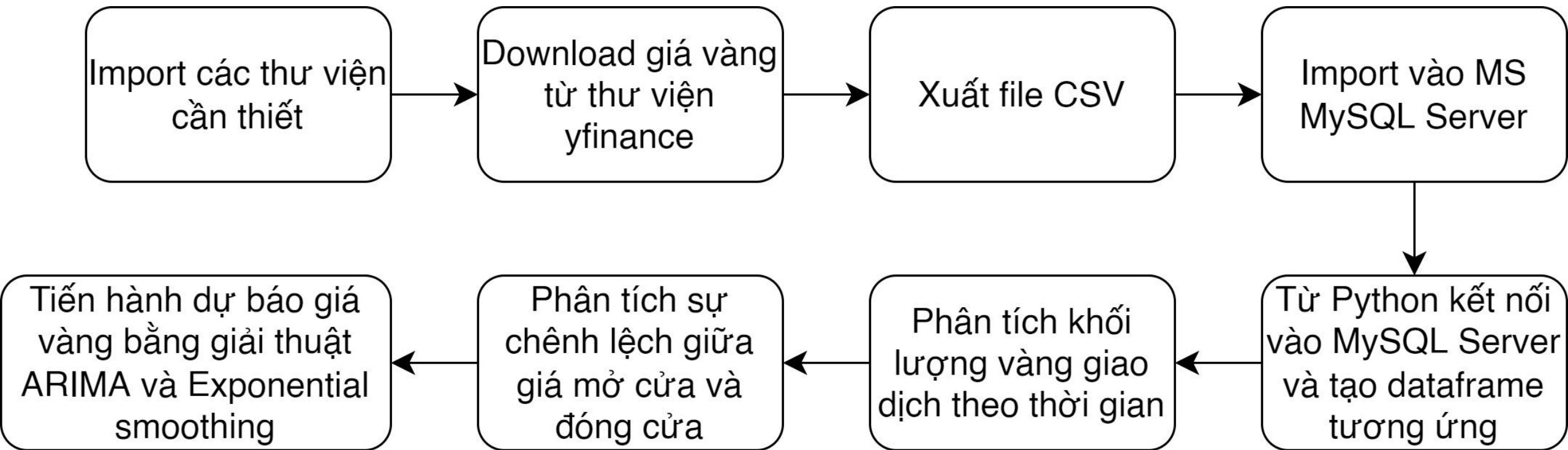
Dự báo giá vàng


Giới thiệu đề tài

Sử dụng vào bộ dữ liệu giá vàng thu thập qua thư viện “yfinance” có sẵn trên Python để dự báo giá vàng thế giới trong 365 ngày kế tiếp



Các bước thực hiện





Mô tả bộ dữ liệu

Dữ liệu giá vàng được tải xuống từ Yahoo Finance thông qua thư viện “yfinance” trong Python. Giá vàng được lấy từ ngày 1 tháng 1 năm 2000 đến ngày 30 tháng 6 năm 2023, bao gồm các cột như Date, Open, High, Low, Close, Adj Close và Volume với mô tả như dưới đây:

- Date: Ngày ghi nhận giá vàng.
- Open: Giá mở cửa vào ngày ghi nhận.
- High: Giá cao nhất đạt được trong ngày.
- Low: Giá thấp nhất đạt được trong ngày.
- Close: Giá đóng cửa vào ngày ghi nhận.
- Adj Close: Giá đóng cửa được điều chỉnh cho bất kỳ lý do nào.
- Volume: Khối lượng vàng (tính theo Ounce) được giao dịch vào ngày ghi nhận.

Bộ dữ liệu này chứa tổng cộng 5.729 hàng, với mỗi hàng đại diện cho dữ liệu giá vàng vào một ngày cụ thể. Dữ liệu được sắp xếp theo thứ tự thời gian với ngày xa nhất ở đầu và ngày gần hiện tại nhất ở dưới cùng.

Cách thu thập và xử lý dữ liệu

1. Sử dụng thư viện yfinance để tải xuống dữ liệu giá vàng từ Yahoo Finance trong giai đoạn từ ngày 1 tháng 1 năm 2000 đến ngày 30 tháng 6 năm 2023
2. Kết nối với cơ sở dữ liệu Microsoft SQL Server bằng thư viện pyodbc và truy xuất dữ liệu từ một bảng có tên gold_price được tạo trước đó.
3. Phương thức read_sql() có sẵn trong pandas được sử dụng để thực thi truy vấn SQL và lưu trữ kết quả trong một DataFrame Pandas.
4. Chuyển đổi cột "date" trong DataFrame ở bước 3 từ dạng "object" sang một kiểu dữ liệu datetime bằng phương thức to_datetime() có sẵn trong pandas. Điều này là cần thiết cho nhiều tác vụ phân tích và dự đoán, vì các giải thuật áp dụng sau này thực hiện trên dữ liệu chuỗi thời gian.

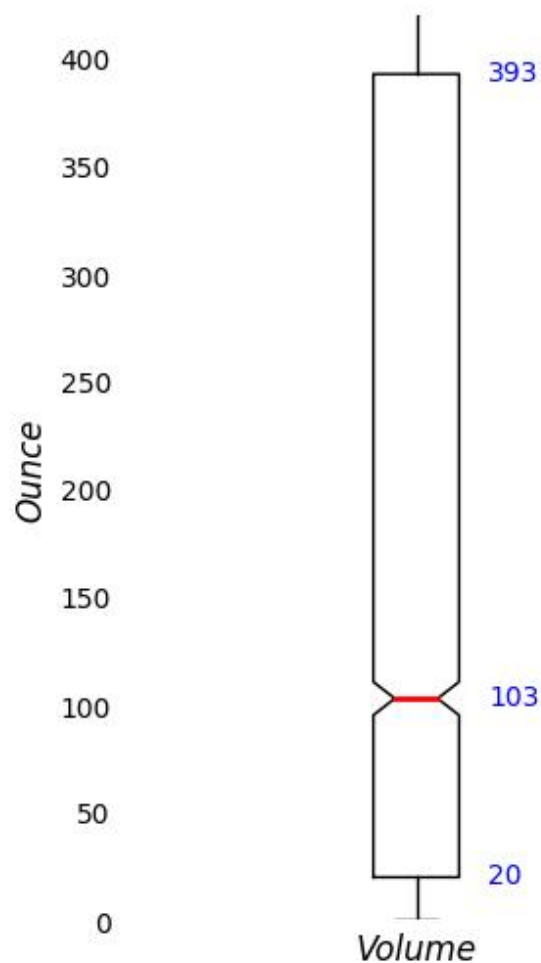


Phân tích khối lượng vàng

BIỂU ĐỒ PHÂN BỐ KHỐI LƯỢNG VÀNG ĐƯỢC GIAO DỊCH MỖI NGÀY

Tổng khối giao dịch: 24,581,133 ounces

Số lượng điểm ngoại vi (lệch phải): 732 điểm



400,000

350,000

300,000

250,000

200,000

150,000

100,000

50,000

0

386,334

count: 5,729

mean: 4,291

std: 24,412

min: 0

25%: 20

50%: 103

75%: 393

max: 386,334

Volume

Phân tích khối lượng vàng

Khối lượng vàng giao dịch mỗi ngày (Ounce) có giá trị được phân bố rộng, từ 0 đến 386,334. Khối lượng giao dịch trung bình là 4,291 cho thấy giá trị trung bình tương đối thấp so với giá trị tối đa. Độ lệch chuẩn là 24,412 cho thấy các điểm dữ liệu được phân tán khá rộng so với giá trị trung bình.

Giá trị nhỏ nhất là 0 cho thấy có một số ngày không phát sinh bất kỳ giao dịch nào (hoặc cũng có thể vào những ngày này chưa được cập nhật dữ liệu). Bách phân vị thứ 25 là 20, nghĩa là 25% trong 5.729 ngày ghi nhận có khối lượng vàng được giao dịch từ 20 Ounce trở xuống. Trung vị (50%) là 103, cho thấy một nửa trong tổng số ngày ghi nhận có khối lượng vàng giao dịch là từ 103 Ounce hoặc thấp hơn. Bách phân vị thứ 75 là 393, nghĩa là 75% trong tổng số ngày ghi nhận có khối lượng vàng được giao dịch từ giá trị 393 hoặc thấp hơn. Giá trị lớn nhất là 386,334, giá trị này có thể là ngoại lệ hoặc đại diện cho một nhóm cụ thể trong tập dữ liệu.

Giá trị trung bình thấp hơn nhiều so với giá trị lớn nhất và toàn bộ giá trị ngoại vi (732 điểm giá trị) nằm ở phía bên phải, cho thấy một phân phối lệch phải với một số các giá trị ngoại vi lớn. Độ lệch chuẩn tương đối cao, cho thấy các điểm dữ liệu được phân tán rộng rãi từ giá trị trung bình. Ngoài ra, cũng quan sát thấy vào một số ngày không có giao dịch nào được thực hiện (hoặc không/chưa được ghi nhận).

Biến động giá đóng cửa

BIẾN ĐỘNG GIÁ ĐÓNG CỬA THEO THỜI GIAN



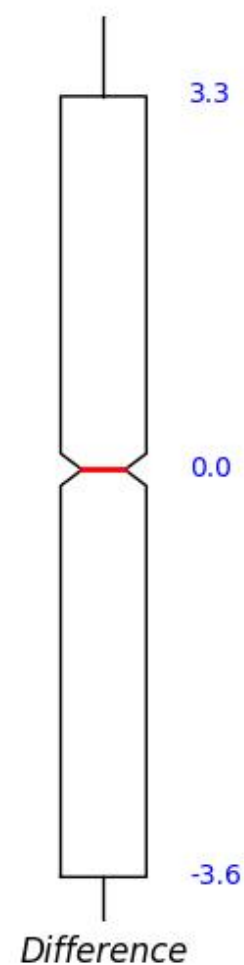
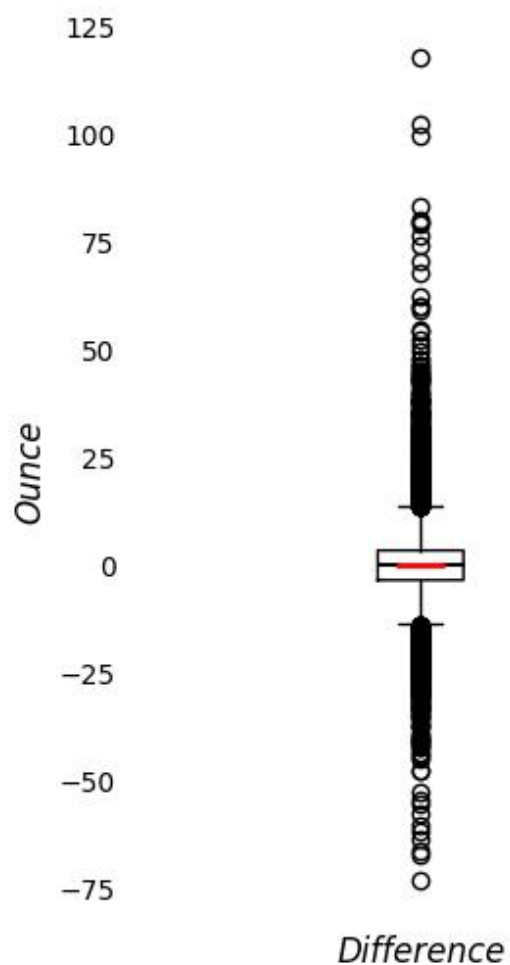
Biến động giá đóng cửa

Biểu đồ cho thấy sự biến động của giá vàng với sự biến đổi đáng kể qua 23 năm. Cụ thể, có thể dễ dàng quan sát thấy các đỉnh và đáy rõ ràng trong biểu đồ, cho thấy các giai đoạn tăng mạnh và giảm nhanh chóng. Tuy nhiên, xu hướng có vẻ chung chung tăng lên trong dài hạn.

Đường xu hướng có phương trình bậc nhất như sau: $y = 0,3x + 275,1$ đại diện cho sự biến động của giá đóng cửa trong 23 năm qua. Hệ số góc 0,3 cho biết độ dốc của đường xu hướng này, hoặc tốc độ tăng lên của giá vàng theo thời gian. Với hệ số góc là 0,3 có thể thấy giá vàng đã tăng với tốc độ tương đối ổn định. Tung độ góc của đường xu hướng là 275,1 đại diện cho giá vàng ước tính tại thời điểm bắt đầu của giai đoạn phân tích (ngày 1 tháng 1 năm 2000).

Chênh lệch giá mở cửa và đóng cửa

BIỂU ĐỒ PHÂN BỐ SỰ CHÊNH LỆCH GIỮA GIÁ MỞ CỬA
VÀ GIÁ ĐÓNG CỬA THEO THỜI GIAN



Số lượng điểm ngoại vi: 817 điểm

count: 5,729.0

mean: 0.1

std: 11.2

min: -73.5

25%: -3.6

50%: 0.0

75%: 3.3

max: 117.6

Chênh lệch giá mở cửa và đóng cửa

Lấy hiệu số giữa giá mở cửa và giá đóng cửa, ta được một dãy số và việc tiến hành phân tích dãy số này cho ta các thông tin chi tiết có giá trị về phân phối của các sự chênh lệch giữa 02 mức giá. Sự khác biệt trung bình là 0,1 cho thấy rằng không có nhiều khác biệt giữa giá mở cửa và giá đóng cửa. Tuy nhiên, độ lệch chuẩn 11,2 cho thấy có sự biến đổi đáng kể trong sự chênh lệch giữa 02 mức giá. Giá trị nhỏ nhất là -73,5, cho thấy trong một số trường hợp, giá đóng cửa có thể thấp hơn đáng kể so với giá mở cửa. Và giá trị lớn nhất là 117.6 cho thấy có một số trường hợp giá đóng cửa cao hơn đáng kể so với giá mở cửa.

Bách phân vị thứ 25 là -3,6 và bách phân vị thứ 75 là 3,3 cho thấy một nửa sự chênh lệch đều tập trung trong khoảng khá nhỏ. Giá trị trung vị là 0 và giá trị trung bình là 0.1 cho thấy sự chênh lệch giữa giá mở cửa và giá đóng cửa là không đáng kể, chủ yếu xoay quanh 0. Tuy nhiên, độ lệch chuẩn là 11.2 do bị ảnh hưởng bởi các giá trị ngoại vi có trong dãy số.

Số lượng điểm ngoại vi là 817 điểm, chiếm khoảng 14.3% tổng số điểm dữ liệu cho thấy số lượng điểm ngoại vi chiếm tỷ trọng không lớn càng củng cố cho sự chênh lệch giữa giá mở cửa và giá đóng cửa là không đáng kể.

Giải thuật ARIMA

ARIMA là viết tắt của Autoregressive Integrated Moving Average (Dịch thô là: Tự hồi quy tích hợp trung bình động). Dưới đây là các khái niệm cốt lõi trong ARIMA:

- Tự hồi quy (AR): Phần "tự hồi quy" của ARIMA đề cập đến ý tưởng rằng các giá trị trong tương lai của một biến phụ thuộc vào các giá trị quá khứ của nó. Nói cách khác, nó giả định rằng giá trị tại bất kỳ thời điểm nào đều bị ảnh hưởng bởi các giá trị trước đó của nó.
- Tích hợp (I): Phần "tích hợp" của ARIMA liên quan đến việc chênh lệch dữ liệu chuỗi thời gian để làm cho nó có tính ổn định. Dữ liệu ổn định có nghĩa là các thuộc tính thống kê của dữ liệu không thay đổi theo thời gian. Chênh lệch giúp loại bỏ xu hướng hoặc đặc tính theo mùa trong dữ liệu, làm cho nó dễ mô hình hóa hơn.
- Trung bình động (MA): Phần "trung bình động" của ARIMA đề cập đến việc sử dụng các lỗi dự báo trong quá khứ để dự đoán các giá trị trong tương lai. Nó giả định rằng các lỗi được thực hiện trong các dự báo trước đó có thể giúp cải thiện các dự báo trong tương lai.

Tuy nhiên, cần lưu ý rằng ARIMA giả định rằng các mẫu cơ bản trong dữ liệu lịch sử sẽ tồn tại trong tương lai. Tuy nhiên, nó có thể không phù hợp với tất cả các loại dữ liệu, đặc biệt nếu các mẫu là phức tạp hoặc phi tuyến.

Giải thuật ARIMA

Tách bộ dữ liệu ra làm 02 bộ: bộ huấn luyện bao gồm 80% số lượng điểm dữ liệu và bộ thử gồm 20% số lượng điểm dữ liệu còn lại. Sử dụng bộ huấn luyện để xây dựng mô hình dự báo dựa trên giải thuật ARIMA và dùng mô hình đã xây dựng áp dụng vào bộ thử để đánh giá hiệu quả dự báo.

Xây dựng mô hình dự báo với các tham số mô hình như dưới đây:

```
ARIMA(train_data['close'], order=(2,2,6))
```

- Tham số tự hồi quy - AR: cài đặt là 2 có nghĩa là giá trị hiện tại của chuỗi thời gian phụ thuộc vào hai quan sát gần đây nhất của nó;
- Tham số liên quan đến sự chênh lệch – I: cài đặt là 2 có nghĩa là dữ liệu được chênh lệch hai lần để làm cho nó ổn định
- Tham số trung bình độ - MA: cài đặt là 6 có nghĩa là mô hình sử dụng sáu lỗi dự báo trước đó để đưa ra dự đoán trong tương lai.

Giải thuật ARIMA

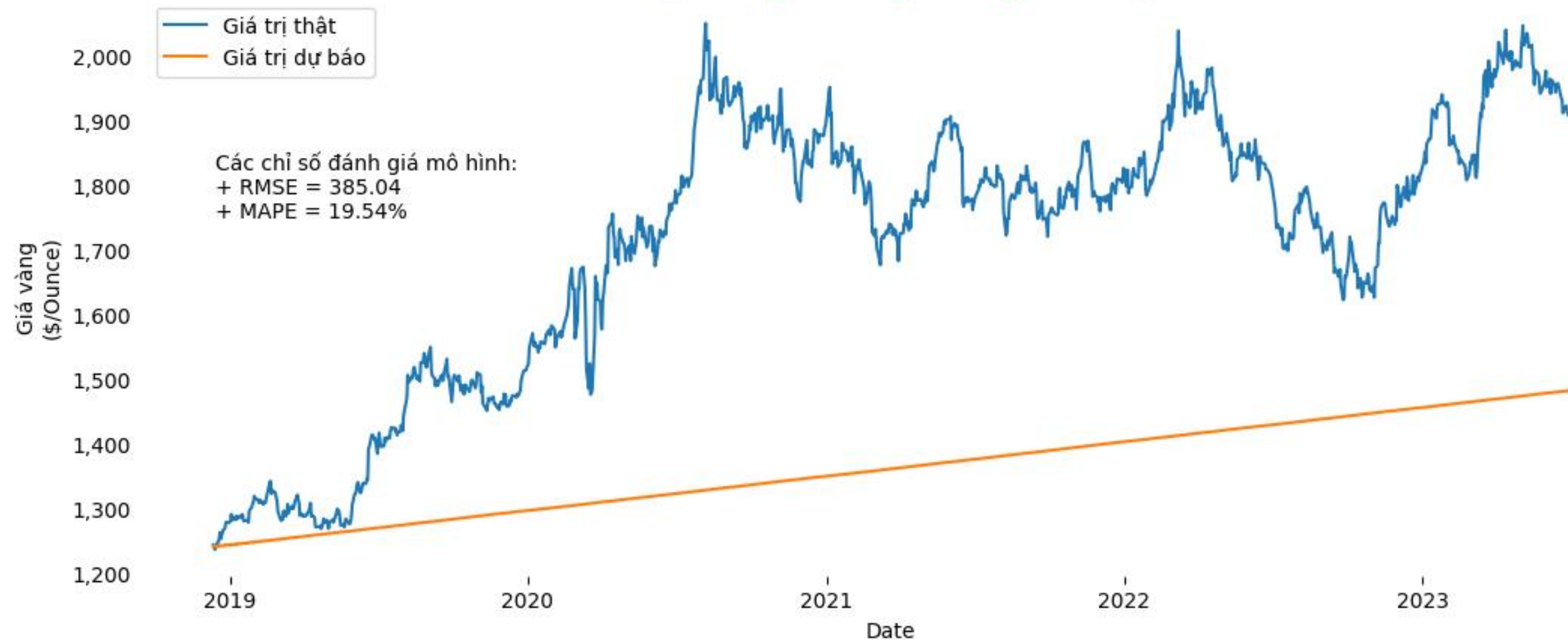
Tiến hành dự báo:

```
model_fit.predict(start=len(train_data), end=len(df)-1, typ='levels')
```

Cài đặt tham số “typ” là “levels” có nghĩa là các giá trị dự đoán sẽ ở cùng một thang đo so với dữ liệu gốc, hay có thể hiểu là tham số typ = “levels” đảm bảo rằng các giá trị dự báo được cung cấp dưới dạng các giá trị thực tế của biến chứ không phải là các chênh lệch hoặc tỷ lệ

Giải thuật ARIMA

Mô hình dự báo giá vàng theo giải thuật ARIMA



Giải thuật Exponential Smoothing

Exponential Smoothing (dịch thô là: làm trơn hàm số mũ) là một phương pháp dự báo chuỗi thời gian giúp dự đoán các giá trị trong tương lai dựa trên các mẫu quan sát được trong dữ liệu lịch sử.

Dưới đây là cách thức hoạt động của giải thuật Exponential Smoothing theo các bước đơn giản:

1. Khởi tạo: gán một giá trị ban đầu cho điểm dữ liệu dự báo. Giá trị ban đầu này thường dựa trên điểm dữ liệu quan sát đầu tiên trong chuỗi thời gian.
2. Làm trơn dữ liệu: giải thuật sẽ tiếp tục xử lý từng điểm dữ liệu kế tiếp, tại mỗi điểm dữ liệu mới, nó sẽ tính một trung bình có trọng số của giá trị dự báo trước đó và giá trị quan sát thực tế. Trọng số được gán cho giá trị dự báo trước đó giảm theo cấp số mũ khi chúng ta di chuyển xa hơn về quá khứ trong chuỗi thời gian.
3. Cập nhật dự báo: sau khi làm trơn điểm dữ liệu hiện tại, thuật toán sẽ cập nhật giá trị dự báo cho giai đoạn tiếp theo dựa trên giá trị trung bình có trọng số.
4. Lặp lại quá trình: Thuật toán tiếp tục quá trình này cho mỗi điểm dữ liệu tiếp theo trong chuỗi thời gian, cập nhật giá trị dự báo cho kỳ tiếp theo dựa trên dự báo trước đó và giá trị quan sát mới.

Giải thuật Exponential Smoothing

Xây dựng mô hình:

```
ExponentialSmoothing(train_data['close'], seasonal_periods=12, trend='add',  
seasonal='mul')
```

Mô hình dự báo dựa trên giải thuật Exponential Smoothing áp dụng trên bộ dữ liệu huấn luyện và bộ thử với tỷ lệ: 80% huấn luyện và 20% thử.

Tham số “seasonal_periods” được cài đặt thành 12, cho biết mô hình đang xem xét một mẫu theo mùa có chu kỳ 12 thời kỳ (ví dụ: tháng nếu dữ liệu là hàng tháng).

Tham số “trend” được đặt thành 'add', có nghĩa là mô hình sẽ xem xét một thành phần xu hướng cộng thêm. Điều này ngụ ý rằng xu hướng sẽ được thêm vào mức của chuỗi thời gian.

Tham số seasonal được đặt thành 'mul', cho biết mô hình sẽ xem xét một thành phần theo mùa nhân lên. Điều này có nghĩa là mẫu theo mùa sẽ được nhân với mức của chuỗi thời gian.

Giải thuật Exponential Smoothing

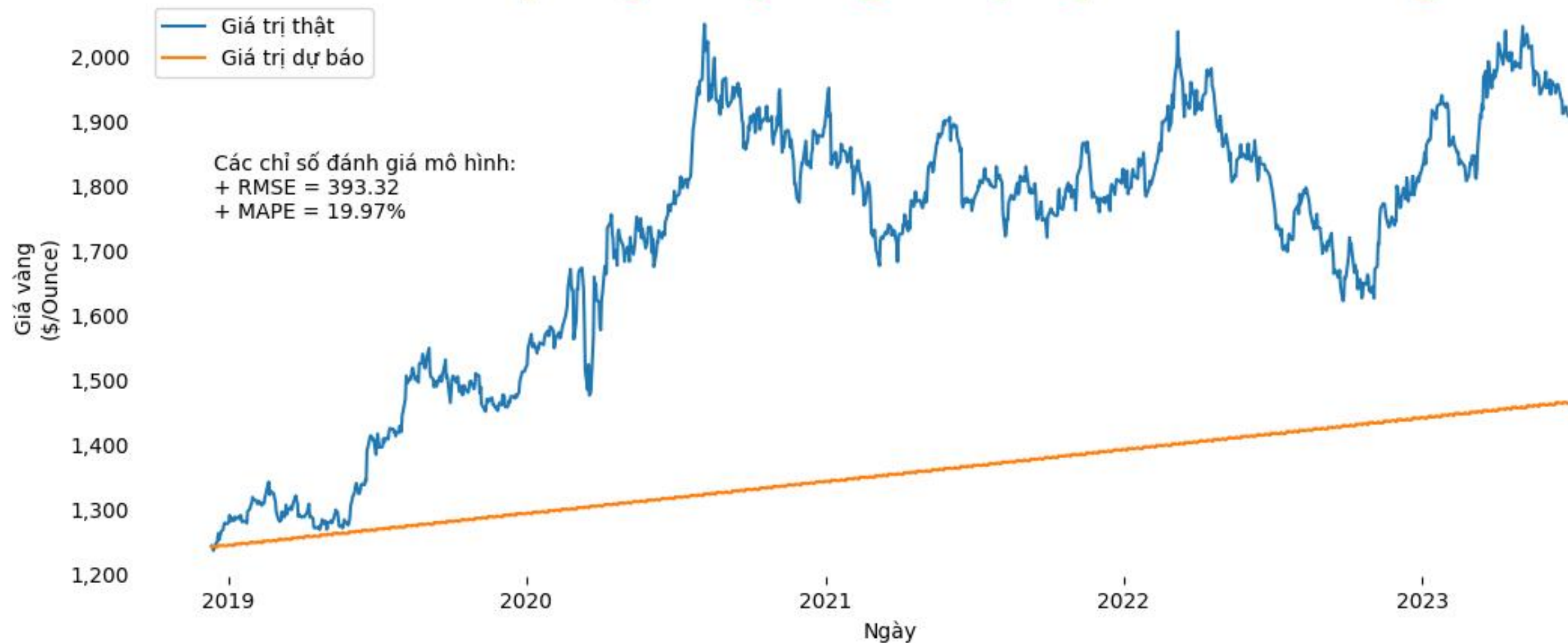
Tiến hành dự báo:

```
fit_model.forecast(len(test_data))
```

Method `fit_model()` tự động tìm các tham số tốt nhất cho mô hình Exponential Smoothing bằng cách tối ưu hóa hiệu suất của mô hình dựa trên bộ huấn luyện. Sau khi đưa mô hình vào, sử dụng method `forecast()` để tạo dự báo cho độ dài của bộ thử.

Giải thuật Exponential Smoothing

Mô hình dự báo giá vàng theo giải thuật Exponential Smoothing



Các chỉ số đánh giá mô hình

Root Mean Square Error – RMSE: đo lường mức độ chênh lệch trung bình giữa các giá trị dự đoán và các giá trị thực tế. RMSE càng thấp càng cho thấy hiệu suất dự đoán tốt hơn. RMSE thường được coi là một chỉ số tốt cho hiệu suất dự đoán nếu nó dưới 10% độ lệch chuẩn của dãy số cần dự báo. Ví dụ, nếu độ lệch chuẩn của các giá trị thực tế là 10, thì mô hình có $RMSE = 1$ sẽ được coi là một mô hình hiệu suất tốt.

Mean Absolute Percentage Error – MAPE: đo lường mức độ chênh lệch phần trăm trung bình giữa các giá trị dự đoán và các giá trị thực tế. MAPE càng thấp càng cho thấy hiệu suất dự đoán tốt hơn. MAPE thường được coi là một chỉ số tốt cho hiệu suất dự đoán nếu nó dưới 20% (ở một vài góc nhìn khác, nó được yêu cầu phải dưới 10%).

Giải thuật Prophet

Prophet là một thuật toán dự báo chuỗi thời gian do nhóm Core Data Science của Facebook phát triển. Nó được giới thiệu vào năm 2017 và được thiết kế đặc biệt cho việc dự báo một cách chính xác và hiệu quả trên dữ liệu chuỗi thời gian với nhiều xu hướng và biến động theo mùa.

Thuật toán Prophet là một mô hình tự hồi quy phi tuyến với biến số ngoại lai. Đây là một phương pháp dự báo sử dụng một mô hình cộng hưởng để phân rã chuỗi thời gian thành ba thành phần: xu hướng, sự thay đổi theo mùa và các ngày lễ.

Cốt lõi của thuật toán là tổng các xu hướng phi tuyến được tính xấp xỉ bằng các hàm tuyến tính từng phần. Mô hình tổng thể được biểu diễn như sau:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon(t)$$

- $g(t)$ đại diện cho hàm xu hướng nắm bắt các thay đổi lâu dài.
- $s(t)$ đại diện cho các thành phần theo mùa, bao gồm các mẫu hàng năm, hàng tuần và hàng ngày.
- $h(t)$ đại diện cho các theo mùa do người dùng xác định và hiệu ứng ngày lễ.
- $\varepsilon(t)$ là sai số hoặc nhiễu.

Giải thuật Prophet

Xây dựng mô hình theo:

```
model_prophet = Prophet()  
model_prophet.fit(train_data_prophet)
```

Tính toán biến dự báo:

```
predictions_prophet = model_prophet.predict(test_data_prophet[['ds']])
```

Sử dụng method `.predict` có sẵn.

Lưu ý:

- Cột dữ liệu thời gian: phải được đặt tên cột là “ds”;
- Cột dữ liệu cần dự báo (biến đầu ra): phải được đặt tên cột là “y”

Giải thuật Prophet

Mô hình dự báo giá vàng theo giải thuật Prophet

