TRƯỜNG ĐẠI HỌC THỦY LỢI

KHOA CÔNG NGHỆ THÔNG TIN



BÁO CÁO BÀI TẬP LỚN MÔN HỌC MÁY

***Đề tài:***

**Dự đoán giá vàng**

**Nhóm sinh viên thực hiện:** Nhóm 8

**Thành viên nhóm:**  Lê Hải Anh

Nguyễn Anh Dũng

**Giảng viên phụ trách môn học:**

**PGS.TS Nguyễn Hữu Quỳnh**

**MỤC LỤC**

[**1.Giới thiệu**](#_kg29e1pxsfb8) **2**

[1.1  Giới thiệu về học máy](#_hxyyvmdr49iq) 2

[1.2  Giới thiệu về thuật toán Linear Regression](#_93vusxxuamhw) 5

[1.3 Trình bày bài toán](#_dyx1a6nt2cwb) 6

[**2.Thực nghiệm**](#_h2ysg7x53evd) **7**

[**3.Kết luận**](#_hfy4or1ud9uy) **9**

[3.1 Ứng dụng của phương pháp hồi quy tuyến tính (Linear Regression)](#_3j2qqm3) 9

[3.2 Các bài toán có thể giải bằng hồi quy tuyến tính (Linear Regression)](#_1y810tw) 9

[3.3 Hạn chế của phương pháp hồi quy tuyến tính (Linear Regression)](#_4i7ojhp) 10

# 

# 1.Giới thiệu

## **1.1  Giới thiệu về học máy**

Những năm gần đây, **AI - Artificial Intelligence** (Trí Tuệ Nhân Tạo), và cụ thể hơn là **Machine Learning** (Học Máy hoặc Máy Học) nổi lên như một bằng chứng của cuộc cách mạng công nghiệp lần thứ tư (1 - động cơ hơi nước, 2 - năng lượng điện, 3 - công nghệ thông tin).

Trong lĩnh vực AI có một nhánh nghiên cứu về khả năng tự học của máy tính được gọi là học máy (*machine learning*). Hiện nay không có 1 định nghĩa chính thức nào về học máy cả nhưng có thể hiểu rằng nó là các kỹ thuật giúp cho máy tính có thể tự học mà không cần phải cài đặt các luật quyết định. Thường một chương trình máy tính cần các quy tắc, luật lệ để có thể thực thi được một tác vụ nào đó như dán nhãn cho các email là thư rác nếu nội dung email có chứa từ khóa “quảng cáo”. Nhưng với học máy, các máy tính có thể tự động phân lại các thư rác thành mà không cần chỉ trước bất kỳ quy tắc nào cả. Nói hơi khó thoát ý, nhưng có thể hiểu đơn giản là nó giúp cho máy tính có được cảm quan và suy nghĩ được như con người. Còn nếu nói nôm na kỹ thuật một chút thì học máy là phương pháp vẽ các đường thể hiện mối quan hệ của tập dữ liệu. Ví dụ như đường ngăn cách 2 loại dữ liệu cho nhãn khác nhau, đường thể hiện xu hướng của giá nhà phụ thuộc vào diện tích và trí hay các đường phân cụm dữ liệu.

a. Lịch sử ra đời:

- Trước năm 1940, rất nhiều thuật toán đằng sau phép Học máy ngày nay đã xuất hiện từ trước khi chiếc máy tính đầu tiên ra đời và bắt nguồn từ phép thống kê. Trong đó, phải kể đến một bước đột phá lớn – công trình của Thomas Bayes vào thế kỷ thứ 18 mà sau đó dẫn tới Định lý Bayes (1812) của Pierre-Simon Laplace. Tương tự, những công trình khác như phép bình phương nhỏ nhất cho dữ liệu hồi quy (1805) của Adrien-Marie Legendre, và kỹ thuật phân tích “Chuỗi Markov” (1913) của Andrey Markov cũng đã góp vai trò to lớn trong sự phát triển của Học máy hiện đại.

- Năm 1948 (máy tính có chương trình lưu trữ sẵn): Vào cuối những năm 1940, các máy tính với chương trình được lưu trữ sẵn đã có thể lưu trữ chương trình và dữ liệu trong cùng một bộ nhớ. Những chiếc máy tính này đã mở đầu cách mạng máy tính hiện đại đầu tiên, với chiếc Manchester Small-Scale Experimental Machine (máy thử nghiệm cỡ nhỏ Manchester, còn gọi là “Baby”) được tạo ra bởi EDSAC, Đại học Cambridge vào năm 1948, và chiếc Manchester Mark 1 tạo ra bởi EDVAC, Đại học Pennsylvania vào năm 1949

- 1950: Máy tính và trí thông minh. Nghiên cứu “Computing Machinery and Intelligence” (Máy tính và Trí thông minh), được xuất bản năm 1950 bởi Alan Turing đã đưa ra một câu hỏi mà cho tới hiện tại vẫn chưa được giải đáp: “*Liệu* *máy móc có thể suy nghĩ?*”. Nghiên cứu đã đi vào sức mạnh của máy tính, và là một trong những bài viết học thuật đầu tiên tập trung mô tả cách phát triển trí tuệ nhân tạo. Một trong những thử nghiệm nổi tiếng của nghiên cứu là “trò chơi mô phỏng” (imitation game), trong đó mức độ thông minh của máy tính sẽ được đánh giá bằng khả năng phân biệt người thật và máy tính khác qua tin nhắn. BBC đã mời Turing tới chia sẻ về thử nghiệm này trong radio “Third Programme” lần lượt vào năm 1951 và 1952.

- 1951: Mạng nơ-ron thần kinh đầu tiên. Marvin Minsky và Dean Edmonds đã xây dựng mạng nơ-ron thần kinh đầu tiên (một hệ thống máy tính mô phỏng hoạt động của não bộ. Với khả năng học hỏi và rút kinh nghiệm qua nhiều lần thử, máy SNARC (Stochastic Neural Analog Reinforcement Computer) có thể di chuyển trong các mê cung giống như chuột thí nghiệm. Máy tính này được xây dựng dựa trên các nguyên lý liên kết – mô phỏng một mạng lưới các đơn vị cơ bản, rồi từ đó xây dựng “trí thông minh”. Sau phát minh này, Minsky chuyển tới làm việc tại Phòng thí nghiệm Trí tuệ nhân tạo tại MIT, và tiếp tục gây dựng những bước tiến lớn cho sự phát triển của trí tuệ nhân tạo. Ông cũng đã tham gia cố vấn chuyên môn cho bộ phim 2001: A Space Odyssey.

- 1974: “Kỷ băng hà” AI đầu tiên. Trái với kỳ vọng, những nghiên cứu về AI trong những năm 1950 và 1960, không đem lại những kết quả lớn. Sự thất bại trong dịch máy (Machine Translation), cũng như những kỳ vọng quá xa vời về AI đã dẫn tới ngân sách cho nghiên cứu bị cắt giảm nghiêm trọng. Trong Báo cáo Lighthill vào năm 1973, tình trạng nghiên cứu đình trệ và thiếu hụt kết quả đã được nhấn mạnh, và chỉ đến cuối những năm 1980, tình trạng này mới được thay đổi. Những phương pháp tiếp cận mới, các hệ thống mới đã chỉ ra những ứng dụng cho các ý tưởng cũ, và từ đây, AI đã xuất hiện các thuật ngữ mới như Tin học (Informatics), Học máy (Machine Learning), và Trí thông minh máy tính (Computational intelligence).

- 1996: Deep Blue đánh bại Garry Kasparov. Thuật ngữ Ai dần trở nên phổ biến với công chúng qua sự kiện máy tính Deep Blue của IBM đánh bại kiện tướng cờ vua thế giới Garry Kasparov. Kasparov đã thắng vào năm 1996, nhưng lại bị đánh bại vào năm 1997 bởi một phiên bản cái tiến của Deep Blue. Để giành được chiến thắng này, Deep Blue đã phải sử dụng rất nhiều khả năng tính toán, với 480 con chip đặc biệt chỉ dùng để chơi cờ, giúp cỗ máy này tìm kiếm từ 6-20 nước cờ trước cho mỗi vị trí, “học hỏi” được qua quan sát hàng nghìn ván cờ.

- 2006: Truyền ngược (Backpropagation). Truyền ngược (Backpropagation) là một yếu tố cốt lõi trong Học máy, được sử dụng để đào tạo các hệ nơron thần kinh sâu. Truyền ngược lần đầu xuất hiện vào năm 1960 trong thuyết quản lý hệ nơron thần kinh sâu, nhưng chỉ được ứng dụng sau khi Geoff Hinton và một số nhà nghiên cứu phát hiện ra hiệu quả của nó thông qua các bộ xử lý hiện đại. Tuy nhiên, vào năm 2017, Hinton lại bày tỏ lo ngại về hạn chế của truyền ngược trong xây dựng hệ thống Học máy, và chỉ ra sự cần thiết của một hệ thống mới.

b. Ứng dụng của Học Máy:

* Học máy là sự giao thoa giữa thống kê cổ điển với khoa học máy tính. Một mục tiêu quan trọng của ngành học máy là làm sao để máy tính thông minh hơn, có khả năng học hỏi và hình thành tri thức một cách tự động từ kinh nghiệm và trở nên hữu ích hơn trong giao tiếp với con người.
* Đây cũng là một trong những mục tiêu ban đầu của ngành học máy, vào những năm 1950 trong thế kỷ trước. Giờ đây, sau nhiều năm phát triển và đặc biệt trong quãng 5-10 năm trở lại đây, học máy đã được ứng dụng trong rất nhiều lĩnh vực của xã hội.
* Học máy có một vai trò quan trọng với mọi Chính phủ trên toàn cầu, trong việc phân tích và dự báo dữ liệu thông tin quốc gia từ quốc phòng - an ninh tới y tế, tài chính, kinh doanh…
* Dữ liệu luôn được thu thập liên tục bằng rất nhiều phương tiện trong từng giây. Những bộ dữ liệu quý giá này sẽ được xử lý bởi hệ thống học máy. Thông tin sau khi xử lý, phân tích chính xác và đưa ra những chỉ báo hiệu quả sẽ mang lại rất nhiều giá trị và lợi thế cho quốc gia đó.

đó:

**1. Học tập có giám sát** là nơi bạn có các biến đầu vào (X) và biến đầu ra (Y) và bạn sử dụng thuật toán để tìm hiểu hàm ánh xạ từ đầu vào đến đầu ra. Mục đích là để xây dựng hàm ánh xạ một cách tốt nhất có thể để khi bạn có dữ liệu đầu vào mới (X) và bạn có thể dự đoán các biến đầu ra (Y) cho dữ liệu đó. Mô hình được huấn luyện bằng cách giảm thiểu sai số lỗi **(loss)** của các dự đoán tại các vòng lặp huấn luyện. Sau quá trình huấn luyện. mô hình sẽ có khả năng đưa ra dự đoán về đầu ra với một đầu vào mới gặp (không có trong dữ liệu học). Nếu không gian đầu ra được biểu diễn dưới dạng rời rạc, ta gọi đó là bài toán phân loại **(classification)**. Nếu không gian đầu ra được biểu diễn dưới dạng liên tục, ta gọi đó là bài toán hồi quy **(regression).**

**- Ưu điểm:**

+ Học có giám sát cho phép bạn thu thập dữ liệu hoặc tạo ra đầu ra dữ liệu từ trải nghiệm trước đó giúp bạn tối ưu hóa tiêu chí hiệu suất bằng kinh nghiệm.

+ Học máy được giám sát giúp bạn giải quyết các loại vấn đề tính toán trong thế giới thực.

**- Nhược điểm:**

+ Ranh giới quyết định có thể được tập trung quá mức nếu tập huấn luyện của bạn không có ví dụ mà bạn muốn có trong một lớp

+ Bạn cần chọn nhiều ví dụ hay từ mỗi lớp trong khi bạn đang đào tạo trình phân loại.

+ Phân loại dữ liệu lớn có thể là một thách thức thực sự.

+ Đào tạo cho việc học có giám sát cần rất nhiều thời gian tính toán.

**2. Học không giám sát:** Là cách huấn luyện một mô hình trong đó dữ liệu học chỉ bao gồm đầu vào mà không có đầu ra. Mô hình sẽ được huấn luyện cách để tìm cấu trúc hoặc mối quan hệ giữa các đầu vào(ví dụ như phân cụm,…)

* + **Ưu điểm:**

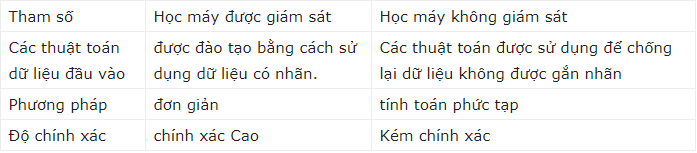
+ Có hữu ích cho việc nén dữ liệu.

* + **Nhược điểm:**

+ Phương pháp tính toán phức tạp

+ Độ chính xác không cao

**\* So sánh giữa học có giám sát và học không giám sát:**



Trong kỷ nguyên số hóa, công nghệ là yếu tố thiết yếu giúp doanh nghiệp thực hiện những cải tiến, nâng cao vị thế cạnh tranh trên thị trường. Nhu cầu nguồn nhân lực chất lượng cao thông hiểu công nghệ  nổi bật liên quan đến AI như Big Data, Data Science, Machine Learning ngày càng nâng cao. Thị trường nhân lực ngành Công nghệ đã phát triển mạnh trong nửa đầu năm nay với số lượng đăng tuyển và ứng tuyển đều tăng cao. Các công ty sẽ đối mặt với thách thức giữ chân nhân tài trong lĩnh vực này bằng cách đề xuất mức lương, thưởng chế độ thăng tiến tương xứng.

## **1.2  Giới thiệu về thuật toán Linear Regression**

Project của nhóm em lần này sẽ sử dụng thuật toán Linear Regression (Hồi Quy Tuyến Tính) thuộc nhóm Supervised learning ( Học có giám sát ). Hồi quy tuyến tính (Linear Regression) là một phương pháp thống kê để thiết lập mối quan hệ giữa một biến phụ thuộc và một nhóm tập hợp các biến độc lập rất đơn giản nhưng đã được chứng minh được tính hữu ích cho một số lượng lớn các tình huống:

Như ta đã biết, vàng chịu tác động của nền kinh tế toàn cầu nên đều có xu hướng biến động và tăng hoặc giảm theo từng ngày. Vì vậy nhóm em sẽ ứng dụng thuật toán Hồi quy tuyến tính nhằm biểu diễn giá vàng trung bình sau mỗi ngày, ngoài biểu diễn dữ liệu, mô hình còn đưa ra dự đoán giá vàng cho thời gian sắp tới nhờ vào việc liên tục cập nhật dữ liệu thông tin giá vàng qua các ngày trong năm.

## **1.3 Trình bày bài toán**

**Input**: Bảng dữ liệu về thông tin ảnh hưởng đến giá

Open : Giá mở

Hight : Giá thời điểm cao nhất

Low : Giá thời điểm thấp nhất

**Ouput**: Biểu diễn và Dự đoán giá vàng của những ngày sau đó

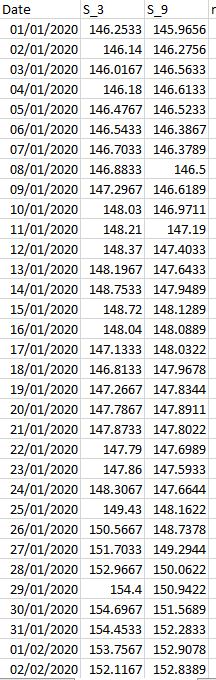
Close : Giá đóng

# 2.Thực nghiệm

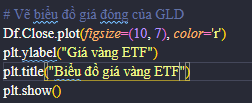
Tập ví dụ huấn luyện :

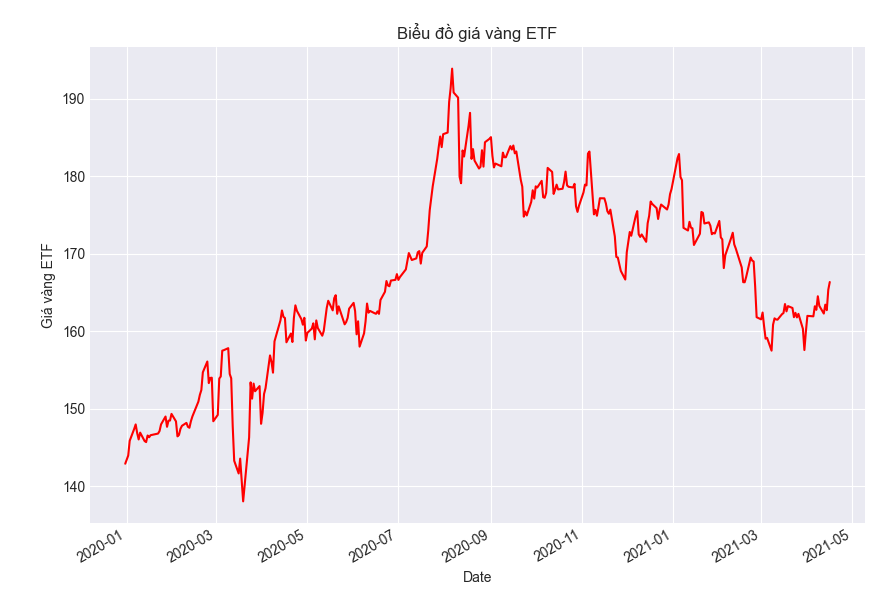


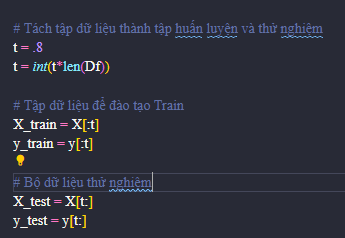
Tập test:



**Giá đóng vàng GLD**

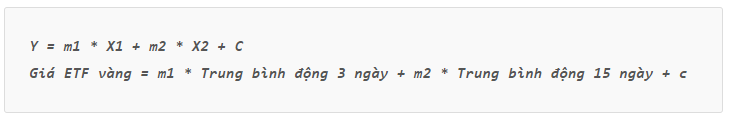
****

****

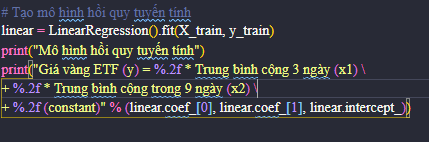


**Phương trình hồi quy tuyến tính**

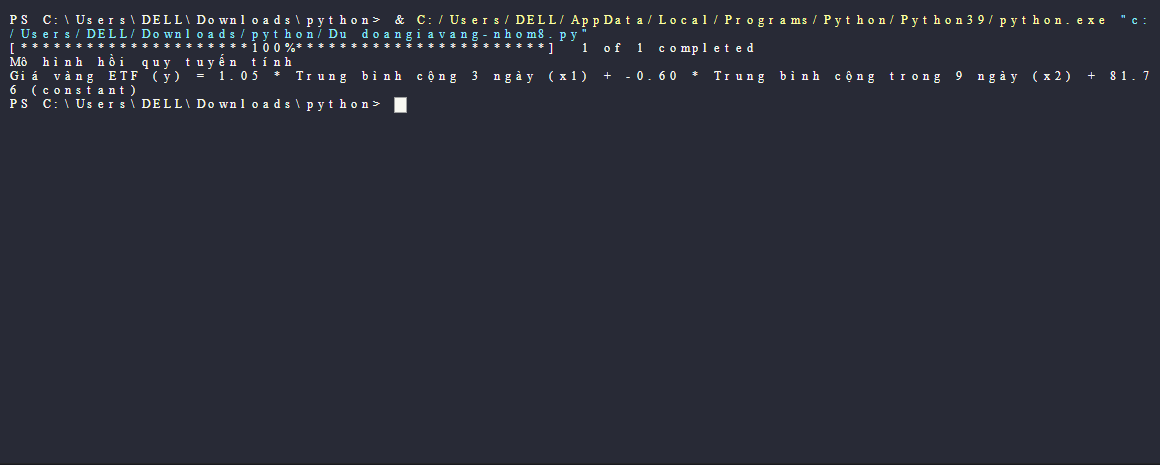
Biến phụ thuộc ‘Y’ là biến mà bạn muốn dự đoán, Các biến độc lập ‘X’ là các biến giải thích để sử dụng dự đoán biến phụ thuộc.



**Tạo mô hình hồi quy tuyến tính**

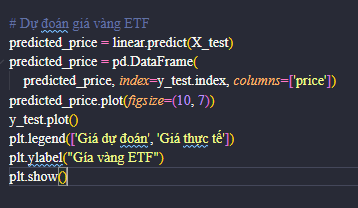


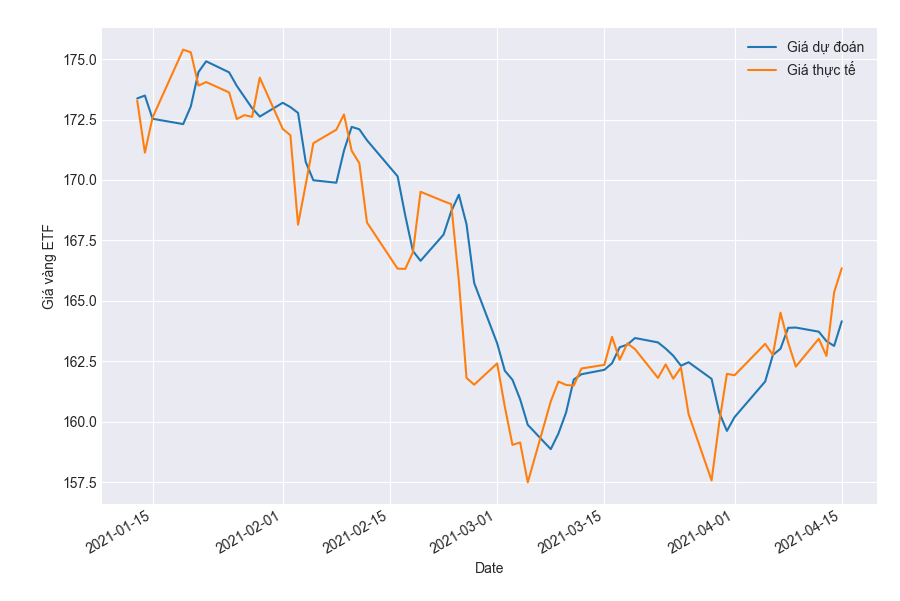
**Output**



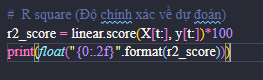
**Dự đoán giá vàng ETF**

Chúng em dự đoán giá vàng ETF bằng cách sử dụng mô hình hồi quy tuyến tính được tạo ra bằng cách sử dụng tập huấn luyện đào tạo .Phương pháp dự đoán tìm giá vàng ETF(Y) cho biến giải thích (X) đã cho.



****

**Độ chính xác**

****

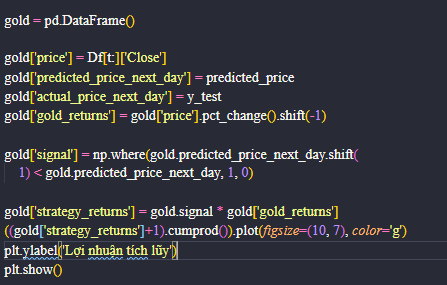
**Kết quả**

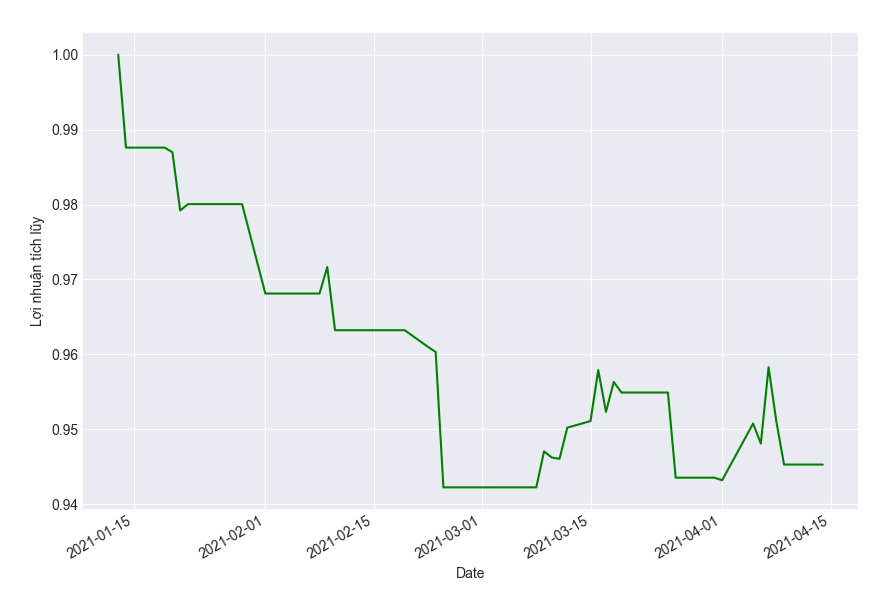
**Lập kế hoạch lợi nhuận tích lũy**

Để tính toán lợi nhuận tích lũy của chiến lược này để phân tích hiệu suất của vàng

-Các bước tính ra lợi nhuận tích lũy như sau :

* Tạo ra phần trăm thay đổi hằng ngày của giá vàng
* Tạo tín hiệu giao dịch mua được biểu thị bằng “1”khi giá dự đoán của ngày hôm sau cao hơn giá dự đoán của ngày hiện tại .Không có vị trí nào khác được thực hiện
* Tính toán lợi nhuận của chiến lược bằng cách nhân tỷ lệ phần trăm thay đổi hàng ngày với tín hiệu giao dịch
* Cuối cùng vẽ biểu đồ lợi nhuận tích lũy

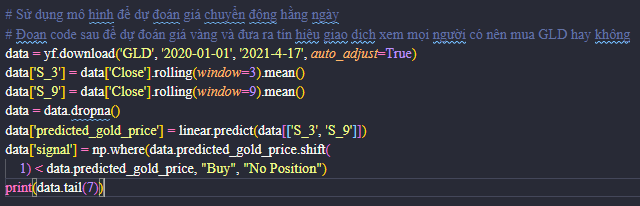
****

****

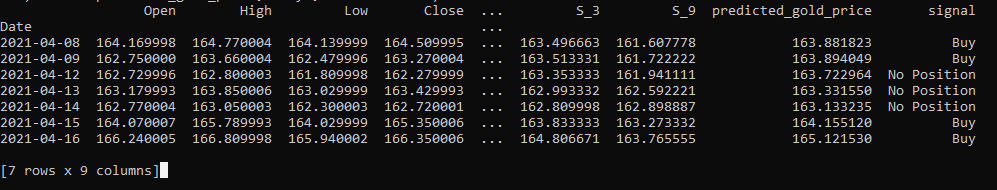
# 

# 

**Sử dụng mô hình hồi quy để dự đoán giá chuyển động hằng ngày**



**Output:**

****

# 3.Kết luận

## **3.1 Ứng dụng của phương pháp hồi quy tuyến tính (Linear Regression)**

Sử dụng phương pháp hồi quy tuyến tính giúp ta dễ dàng dự đoán được các giá trị tuyến tính dựa vào tập huấn luyện cho trước. Ưu điểm của hồi quy tuyến tính là tính đơn giản của hàm, khi giá trị x tăng hoặc giảm thì giá trị y cũng tăng hoặc giảm, tuyến tính theo f(x)

## **3.2 Các bài toán có thể giải bằng hồi quy tuyến tính (Linear Regression)**

Từ phương pháp hồi quy tuyến tính và mô hình thuật toán được xây dựng dựa vào phương pháp trên, ta có thể thấy hàm số  là một hàm tuyến tính theo cả **w** và **x**. Trên thực tế, phương pháp hồi quy tuyến tính (Linear Regression) có thể áp dụng cho các mô hình chỉ cần tuyến tính theo **w**

## **3.3 Hạn chế của phương pháp hồi quy tuyến tính** **(Linear Regression)**

Hạn chế đầu tiên của Linear Regression là nó rấtnhạy cảm với nhiễu. Trong ví dụ về mối quan hệ giữa diện tích và giá căn phòng bên trên, nếu có chỉ một cặp dữ liệu nhiễu thì kết quả sẽ sai khác đi rất nhiều. Vì vậy, trước khi thực hiện Linear Regression, các nhiễu (*outlier*) cần phải được loại bỏ. Bước này được gọi là tiền xử lý

Hạn chế thứ hai của Linear Regression là nó không biểu diễn được các mô hình phức tạp.

# 

# 4.Một số tài liệu tham khảo

[**https://machinelearningcoban.com/2016/12/28/linearregression/**](https://machinelearningcoban.com/2016/12/28/linearregression/)

[**https://blog.quantinsti.com/gold-price-prediction-using-machine-learning-python/**](https://blog.quantinsti.com/gold-price-prediction-using-machine-learning-python/)

**https://www.etf.com/GLD**