

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH
KHOA CÔNG NGHỆ THÔNG TIN**



ĐỒ ÁN MÔN HỌC

DỰ ĐOÁN GIÁ VÀNG

Giảng viên hướng dẫn: HỒ KHÔI
Sinh viên thực hiện: CHU DOÃN ĐỨC
MSSV: 2000003917
Chuyên ngành: Khoa học dữ liệu
Môn học: Deep Learning
Khóa: 2020

Tp.HCM, 28 tháng 12 năm 2022

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH
KHOA CÔNG NGHỆ THÔNG TIN**



ĐỒ ÁN MÔN HỌC

DỰ ĐOÁN GIÁ VÀNG

Giảng viên hướng dẫn: HỒ KHÔI
Sinh viên thực hiện: CHU DOÃN ĐỨC
MSSV: 2000003917
Chuyên ngành: Khoa học dữ liệu
Môn học: Deep Learning
Khóa: 2020

Tp.HCM, 28 tháng 12 năm 2022

[illegible]

.....

.....

.....

Giáo viên hướng dẫn

MỤC LỤC

| | |
|---|----|
| LỜI CẢM ƠN..... | 1 |
| CHƯƠNG 1: GIỚI THIỆU..... | 2 |
| 1.1 – GIỚI THIỆU ĐỀ TÀI:..... | 2 |
| 1.2 – LÝ DO CHỌN ĐỀ TÀI:..... | 2 |
| 1.3 – MỤC TIÊU CỦA ĐỀ TÀI: | 3 |
| 1.4 – PHƯƠNG PHÁP ĐỀ TÀI: | 3 |
| 1.5 – ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU: | 4 |
| CHƯƠNG 2: ỨNG DỤNG THUẬT TOÁN..... | 5 |
| 2.1 - MÔ TẢ BÀI TOÁN: | 5 |
| 2.1.1 – Hồi quy tuyến tính: | 5 |
| 2.1.2 – Bài toán hồi quy trong Máy học: | 5 |
| 2.1.3 – Bài toán dự đoán xu hướng giá Vàng: | 6 |
| 2.1.4 – Hồi quy đa thức:..... | 7 |
| 2.1.5 – Quy trình phân tích hồi quy tuyến tính đa biến: | 8 |
| 2.2 - XÂY DỰNG BỘ DỮ LIỆU:..... | 11 |
| 2.3 - ÁP DỤNG THUẬT TOÁN VÀO BÀI TOÁN:..... | 12 |
| 2.3.1 – Đặt ra bài toán:..... | 12 |
| 2.3.2 – Áp dụng thuật toán Hồi quy tuyến tính: | 12 |
| 2.4 - THỰC NGHIỆM VỚI THƯ VIỆN PYTHON: | 14 |
| 2.4.1 – Import thư viện: | 14 |
| 2.4.2 – Đọc dữ liệu: | 14 |
| 2.4.3 – Kiểm tra và phân tích dữ liệu: | 15 |
| CHƯƠNG 3: XÂY DỰNG ỨNG DỤNG BẰNG NGÔN NGỮ PYTHON..... | 19 |
| 3.1 – XÂY DỰNG ỨNG DỤNG VÀ GIẢI THÍCH:..... | 19 |
| 3.1.1 – Xây dựng mô hình dự đoán, tạo model và huấn luyện:..... | 19 |
| 3.1.2 – Dự đoán và đánh giá mô hình:..... | 20 |
| 3.2 – KẾT LUẬN: | 25 |
| 3.2.1 – KẾT QUẢ ĐẠT ĐƯỢC:..... | 25 |
| 3.2.2 – HẠN CHẾ VÀ HƯỚNG PHÁT TRIỂN:..... | 26 |
| TÀI LIỆU THAM KHẢO | 28 |

DANH MỤC HÌNH

| | |
|---|----|
| Hình 2. 1: mô hình tổng quát bài toán dự đoán xu hướng giá vàng | 6 |
| Hình 2. 2: biểu đồ hồi quy đa thức (sự tương quan của cá qua từng độ tuổi)..... | 8 |
| Hình 2. 3: các bước của kiểm định Durbin – Watson..... | 10 |
| Hình 2. 4: dữ liệu giá của vàng và các tài nguyên liên quan | 12 |
| Hình 2. 5: thư viện cần dùng để xử lý, phân tích và trực quan dữ liệu | 14 |
| Hình 2. 6: tải dữ liệu từ google lên colab | 14 |
| Hình 2. 7: thông tin tổng quát về dữ liệu..... | 15 |
| Hình 2. 8: biểu đồ Seaborn của dữ liệu..... | 16 |
| Hình 2. 9: biểu đồ phân phối giá Vàng..... | 17 |
| Hình 2. 10: bản đồ nhiệt của dữ liệu..... | 18 |
| | |
| Hình 3. 1: phân tách dữ liệu..... | 19 |
| Hình 3. 2: tạo biến chứa các dữ liệu | 19 |
| Hình 3. 3: in ra các dữ liệu vừa tạo mới | 20 |
| Hình 3. 4: hàm huấn luyện hồi quy tuyến tính | 20 |
| Hình 3. 5: kết quả dự đoán..... | 21 |
| Hình 3. 6: biểu đồ phân tán của kết quả dự đoán | 21 |
| Hình 3. 7: biểu đồ cufflinks của kết quả dự đoán..... | 22 |
| Hình 3. 8: biểu đồ displot thể hiện sự chênh lệch..... | 22 |
| Hình 3. 9: chỉ số đánh giá mô hình dự đoán | 23 |
| Hình 3. 10: kết quả dự đoán mô hình dựa trên phương sai | 24 |
| Hình 3. 11: đánh giá mô hình bằng hệ số coeff..... | 24 |
| Hình 3. 12: kết quả khi tăng một đơn vị trong cột..... | 25 |

LỜI CẢM ƠN

Đầu tiên, em xin gửi lời cảm ơn chân thành đến Trường Đại học Nguyễn Tất Thành đã đưa môn học “Deep Learning” vào trương trình giảng dạy. Đặc biệt, em xin gửi lời cảm ơn sâu sắc đến giảng viên bộ môn – Thầy Hồ Khôi trực tiếp hướng dẫn, dạy dỗ, truyền đạt những kiến thức quý báu cho em trong suốt thời gian học tập vừa qua. Trong thời gian tham gia lớp học của thầy, em đã có thêm cho mình nhiều kiến thức bổ ích, tinh thần học tập hiệu quả, nghiêm túc và đã cho em chắc chắn được hoạch định trong tương lai của mình.

“Deep Learning” là môn học thú vị, vô cùng bổ ích và có tính thực tế cao. Đảm bảo cung cấp đủ kiến thức, gắn liền với nhu cầu thực tiễn của sinh viên nói chung và riêng bản thân em nói riêng. Tuy nhiên, do vốn kiến thức còn nhiều hạn chế và khả năng tiếp thu thực tế còn nhiều bỡ ngỡ và hạn hẹp. Mặc dù em đã cố gắng hết sức nhưng chắc chắn bài báo cáo của em khó có thể tránh khỏi những thiếu sót và nhiều chỗ còn chưa chính xác, kính mong các thầy/cô chấm bài xem xét và góp ý để bài tiểu luận của em được hoàn thiện hơn.

Kính chúc thầy có nhiều sức khỏe, hạnh phúc, thành công trên con đường giảng dạy.

Em xin chân thành cảm ơn!

CHƯƠNG 1: GIỚI THIỆU

1.1 – GIỚI THIỆU ĐỀ TÀI:

Vàng luôn đóng góp giá trị cho nền kinh tế thế giới trong nhiều thập kỷ. Nó đã phục vụ như một công thông tin an toàn trong thời kỳ lạm phát vô thời hạn cũng như một vật trang trí. Vàng đóng vai trò trụ cột trong hầu hết tất cả các danh mục đầu tư vì nhiều lý do, có thể kể đến một số ít - nó đã vượt trội so với các khoản đầu tư khác về lợi nhuận khi thị trường toàn cầu cho các sản phẩm khác đi xuống, mang lại lợi nhuận rất cao là 13,66% trong 15 năm qua, chỉ thấp hơn một chút so với lợi nhuận của Sensex là 13,97% trong cùng kỳ khoảng thời gian. Một điểm thu hút lớn khác với vàng là nó được cho là thiếu tương quan với các tài sản khác, do đó làm giảm rủi ro bằng cách đa dạng hóa danh mục đầu tư. Điều đáng kinh ngạc là, nhu cầu về kim loại màu vàng đã tăng hơn 120 phần trăm trên cơ sở hàng năm. Nhu cầu vàng chủ yếu là để đầu tư và ngành kim hoàn theo sau là nhu cầu công nghiệp. Trong khi người tiêu dùng trân trọng vàng như một tài sản, nó hoạt động như nguyên nhân chính dẫn đến thâm hụt tài khoản vãng lai, đây là một thách thức đối với các điều kiện tài chính của quốc gia. Vàng hoạt động như một tài sản nhân rồi làm tăng thêm tai ương của sự bất ổn tài chính. Lập mô hình giá do đó vàng mang lại lợi ích to lớn cho người tiêu dùng. Nghiên cứu hiện tại cố gắng phát triển một hồi quy mô hình dự đoán giá vàng sử dụng 7 biến tài chính: SPX, GLD, USO, SLV, EUR/USD. Các biến này đã được kiểm tra tính đa cộng tuyến và mô hình hồi quy được phát triển bằng cách sử dụng phân tích thành phần chính (PCA) để tránh hiện tượng đa cộng tuyến. Phân tích hồi quy của Giá sử dụng thời gian khác nhau của các yếu tố dự đoán biến thể vàng như nhu cầu vàng, giá dầu và chỉ số chứng khoán (SPX) theo một mô hình hồi quy hệ số thay đổi sẽ giúp ước tính các biến thể tương đối trong hồi quy biến.

1.2 – LÝ DO CHỌN ĐỀ TÀI:

Trong những năm gần đây, với sự phát triển bùng nổ của khoa học công nghệ, các nhà nghiên cứu đã nỗ lực để trong việc tìm ra những giải pháp, những phương pháp đánh giá, tối ưu hóa để nâng cao độ chính xác của mô hình. Chính vì lẽ đó mà các bài toán khó

tưởng chừng không thể giải quyết đã ngày một được giải đáp thúc đẩy cho nền kinh tế phát triển từ đó các rào cản về công nghệ cũng được tháo gỡ.

Hiện nay việc dự đoán giá cả trên thị trường đã trở nên cần thiết đối với nhiều ngành nghề, lĩnh vực khác nhau. Việc xây dựng một mô hình dự đoán giá có thể đáp ứng được nhu cầu sử dụng vẫn đang là vấn đề cần thiết mặc dù đã tồn tại rất nhiều mô hình dự đoán giá cả. Chính vì lẽ đó dự đoán thị trường Vàng cũng là một nhu cầu cấp thiết và có ý nghĩa thực tiễn. Chủ đề này đã được nhiều nhà nghiên cứu trong và ngoài nước quan tâm và đưa ra nhiều giải pháp. Mỗi giải pháp có ưu nhược điểm khác nhau, tuy nhiên việc sử dụng máy học là giải pháp mang lại hiệu quả tốt. Vì các lý do trên chúng em đã lựa chọn đề tài “Dự đoán giá Vàng” là chủ đề cho môn học Deep Learning.

1.3– MỤC TIÊU CỦA ĐỀ TÀI:

Từ việc sử dụng dữ liệu giá vàng qua các năm gần đây và áp dụng các thư viện của Python huấn luyện và dự đoán để phát triển và xây dựng mô hình có tính chính xác cao để đưa ra dự đoán giá vàng.

1.4– PHƯƠNG PHÁP ĐỀ TÀI:

Mô hình hồi quy tuyến tính bội là một trong những kỹ thuật thống kê hiệu quả nhất sử dụng nguyên tắc bình phương nhỏ nhất để ước lượng tham số. Mô hình giả thuyết là:

- $$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k + \varepsilon.$$

Giá trị của các tham số b_0, b_1, \dots, b_k sẽ được ước lượng theo nguyên tắc bình phương bé nhất. Theo phương pháp này, các ước tính tốt nhất của các tham số là những tham số sẽ có tổng bình phương giá trị phần dư nhỏ nhất. Các ước lượng sẽ là các ước lượng không chệch tuyến tính tốt nhất chỉ trong trường hợp không có đa cộng tuyến của các biến giải thích và không có tự tương quan của các phần dư ngoài đối với tất cả các điều kiện cơ bản khác. Có thể tránh được sự hiện diện của đa cộng tuyến bằng cách áp dụng phân tích thành phần chính sẽ đưa ra các biến trực giao là các phép biến đổi tuyến tính của các biến giải thích.

Đề tài này sử dụng Linear Regression (Hồi quy tuyến tính) thuộc nhóm Supervised learning (Học có giám sát) để phân tích dữ liệu dự và đoán giá trị của Vàng.

1.5 – ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU:

Dự đoán thị trường Vàng thông qua: giá thị trường chứng khoán bao gồm 500 cổ phiếu của các công ty lớn được giao dịch trên thị trường chứng khoán Hoa Kỳ (SPX), giá quỹ dầu mỏ Hoa Kỳ (USO), và sự chênh lệch tỷ giá giữa Euro và Đô la Mỹ (EUR/USD).

CHƯƠNG 2: ỨNG DỤNG THUẬT TOÁN

2.1 - MÔ TẢ BÀI TOÁN:

2.1.1 – Hồi quy tuyến tính:

- **Hồi quy (regression):**

là phương pháp thống kê toán học để ước lượng và kiểm định các quan hệ giữa các biến ngẫu nhiên, và có thể từ đó đưa ra các dự báo. Các quan hệ ở đây được viết dưới dạng các hàm số hay phương trình.

- Khẳng định mối liên hệ giữa hai biến số.
- Dự đoán hoặc ước lượng giá trị của một biến số từ các giá trị của một hay nhiều biến số khác.

Ví dụ: dự đoán huyết áp dựa trên tuổi, cân nặng,

- **Tương quan (correlation):** Đo lường độ lớn của mối quan hệ giữa các biến số với nhau.
- **Mô hình hồi quy:**

Cần đưa ra một dự đoán hoặc ước lượng giá trị của một biến số từ các giá trị của một hay nhiều biến số.

Người nghiên cứu đưa ra được một mô hình toán học hoặc áp dụng được các mô hình để phân tích các quần thể.

Mô hình có, hoặc ít nhất là một xấp xỉ đại diện cho quần thể đó không - mô hình đó là một đại diện tốt nhất cho quần thể cần quan tâm.

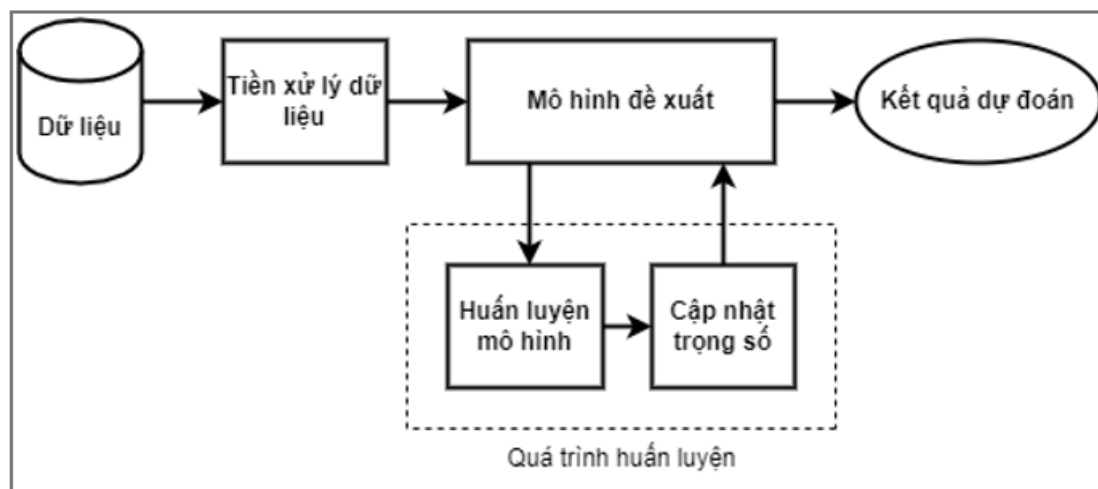
2.1.2 – Bài toán hồi quy trong Máy học:

Bài toán hồi quy là một trong những bài toán lớn trong lĩnh vực học máy, liên quan đến việc dự đoán một giá trị số thực bằng cách sử dụng các biến số học được thông qua dữ

liệu để tìm mối quan hệ giữa các biến số đó. Có nhiều mô hình hồi quy có thể kể đến như: Hồi quy đơn biến, hồi quy đa biến, hồi quy tuyến tính hoặc phi tuyến.

2.1.3 – Bài toán dự đoán xu hướng giá Vàng:

Lần lượt các nhà tiên phong về phương pháp dự đoán chuỗi thời gian đã xuất hiện, có thể kể đến mô hình AutoRegressive được nhà thống kê Udney Yule và các đồng nghiệp phát minh vào những năm 1920. Mô hình này là nền tảng cho một số mô hình thống kê và kinh tế lượng sau này như ARMA, ARIMA. Điều đó cho thấy từ trước những năm phát triển vượt bậc của lĩnh vực trí tuệ nhân tạo, sự quan tâm của các nhà khoa học đối với bài toán này là không hề nhỏ. Trong những năm gần đây, các mô hình máy học đã được ứng dụng vào bài toán này để hỗ trợ các nhà đầu tư tạo ra lợi nhuận, tuy nhiên với những mô hình máy học truyền thống thì độ chính xác vẫn còn những hạn chế nhất định. Tuy nhiên với sự phát triển của những mô hình học sâu, việc nhận dạng được những mẫu phi tuyến tính trong chuỗi thời gian của chứng khoán đã trở nên dễ tiếp cận hơn bao giờ hết. Một hướng tiếp cận khá phổ biến và hiệu quả trong những năm gần đây cho bài toán dự đoán chuỗi thời gian là sử dụng mô hình LSTM đây là mô hình học sâu thu hút được nhiều sự quan tâm của các nhà nghiên cứu trong và ngoài nước. LSTM được sử dụng rất nhiều cho các bài toán có dữ liệu thời gian hay tuần tự như dịch máy, nhận diện giọng nói, dự báo thời tiết với độ chính xác cao.



Hình 2. 1: mô hình tổng quát bài toán dự đoán xu hướng giá vàng

2.1.4 – Hồi quy đa thức:

Mô hình hồi qui logistic đa thức (Multinomial logistic regression) tương tự như mô hình hồi qui logistic nhị thức nhưng biến phụ thuộc là biến định tính có nhiều hơn 2 trạng thái (hoặc mức). Ví dụ (khỏi bệnh, khỏi với dư chứng, tử vong) hoặc (tốt, trung bình, xấu).

Mô hình hồi qui logistic đa thức phát biểu:

$$\text{Log}(p_i/p_j) = \alpha_{ij} + \beta_{ij}x_1 + \beta_{ij}x_2 + \dots + \varepsilon_{ij}$$

Gọi: p_0 là xác suất khỏi bệnh p_1 là xác suất khỏi với dư chứng p_2 xác suất tử vong. Ta có 3 phương trình sau:

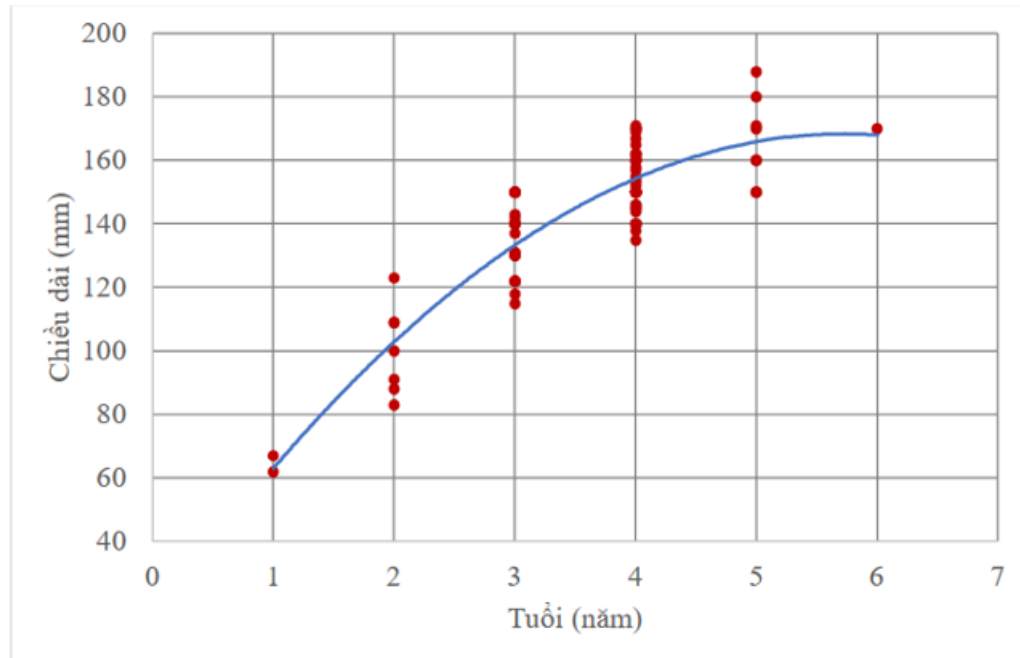
- $\text{Log}(p_1/p_0) = \alpha_{10} + \beta_{10}x_1 + \beta_{10}x_2 + \dots$ (1)
- $\text{Log}(p_2/p_0) = \alpha_{20} + \beta_{20}x_1 + \beta_{20}x_2 + \dots$ (2)
- $\text{Log}(p_2/p_1) = \alpha_{21} + \beta_{21}x_1 + \beta_{21}x_2 + \dots$ (3)

Như đã đề cập ở phần trên, đa hồi quy là một dạng của hồi quy, trong đó có nhiều hơn một biến độc lập. Đa hồi quy bao gồm một kỹ thuật gọi là hồi quy đa thức. Trong hồi quy đa thức, biến phụ thuộc hồi quy vào lũy thừa của các biến độc lập.

Ví dụ: Vào năm 1981, $n = 78$ con cá *Thái Dương xanh* (*Bluegrill*) được lấy mẫu ngẫu nhiên ở Lake Mary, tiểu bang Minnesota, Mỹ. Nhà nghiên cứu đã đo lường và ghi lại các dữ liệu sau:

- Chiều dài (y) của con cá, đơn vị milimet.
- Độ tuổi (x) của con cá đó, đơn vị năm.

Kết quả thu thập được được thể hiện qua biểu đồ sau:



Hình 2. 2: biểu đồ hồi quy đa thức (sự tương quan của cá qua từng độ tuổi)

Mặc dù chiều dài của con cá tăng lên qua từng độ tuổi, nhưng lại không hoàn toàn theo tuyến tính (Hình 9). Để mô hình hoá dữ liệu này, người ta xây dựng một mô hình đa thức bậc 2, hay còn gọi là hàm số bậc 2 như sau: $\hat{y}^{(i)} = \theta_0 + \theta_1 x^{(i)} + \theta_2 (x^{(i)})^2$

Trong đó:

- $\hat{y}^{(i)}$ là chiều dài của con cá Thái dương xanh thứ i (mm)
- $x^{(i)}$ là tuổi của con cá Thái dương xanh thứ i (năm)

Bên cạnh đa thức bậc 2, hồi quy đa thức còn có các dạng khác từ bậc 3 đến n. Công thức tổng quát:

$$\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots + \theta_i x^i, \forall i = 1, 2, \dots, n$$

Để tìm các tham số ta sử dụng phương pháp bình phương tối thiểu.

2.1.5 – Quy trình phân tích hồi quy tuyến tính đa biến:

- Kiểm định ý nghĩa của hệ số hồi quy:

Kiểm định này xem xét biến độc lập tương quan có ý nghĩa với biến phụ thuộc hay không.

Theo Green (1991), sử dụng kiểm định t. Khi mức ý nghĩa sig (significance) của hệ số hồi quy ≤ 0.10 hoặc độ tin cậy từ 90% trở xuống thì kết luận biến X tương quan tuyến tính với biến Y.

- Mức độ giải thích của mô hình:

Kiểm định này xem xét mức độ giải thích của mô hình lựa chọn. Theo Green (1991), sử dụng thước đo R² hiệu chỉnh (Adjusted R square). R² hiệu chỉnh cho biết % thay đổi của biến phụ thuộc được giải thích bởi biến độc lập của mô hình. Thước đo này càng tiến về 100% càng tốt, cho thấy mô hình có mức độ giải thích cao.

- Mức độ phù hợp của mô hình:

Kiểm định này xem xét mức độ phù hợp của mô hình lựa chọn. Nói cách khác, mô hình hồi quy tuyến tính có phù hợp với dữ liệu thực tiễn không. Về tổng thể, biến độc lập có tương quan tuyến tính với phụ thuộc.

Theo Green (1991), sử dụng phân tích phương sai (Analysis of variance, ANOVA), với kiểm định F, mức ý nghĩa (sig.) ≤ 0.05 hoặc độ tin cậy 95%.

- Kiểm định hiện tượng đa cộng tuyến:

Kiểm tra hiện tượng các biến độc lập tương quan tuyến tính với nhau. Thước đo mức độ phóng đại phương sai (Variance Inflation Factor, VIF) đòi hỏi phải nhỏ hơn 10.

Tương ứng với mỗi biến độc lập, $VIF < 10$, không có hiện tượng đa cộng tuyến.

- Kiểm định hiện tượng tương quan:

Theo Durbin – Watson (1971), khi giá trị các phần dư (Residuals) tương quan với nhau, kết quả ước lượng OLS không còn tin cậy. Do đó, sử dụng kiểm định Durbin – Watson để tra hiện tượng này.

Các bước của kiểm định Durbin – Watson:

Bước 1: Xác định trị số thống kê Durbin – Watson (d) của mô hình (Trong bảng summary).

Bước 2: Dựa vào bảng số thống kê Durbin – Watson, căn cứ vào số quan sát (n), số tham số (k-1) của mô hình hồi quy, mức ý nghĩa 95%, xác định giá trị thống kê trên (dU, upper d) và giá trị thống kê dưới (dL, lower d).

| | | | | |
|---------------------|----------------|------------------------|----------------|------------------|
| Tự tương quan dương | Không kết luận | Không có tự tương quan | Không kết luận | Tự tương quan âm |
| 0 | dL | dU | 4-dU | 4-dL |

Hình 2. 3: các bước của kiểm định Durbin – Watson

Trong bảng tra, khi d của mô hình lớn hơn dU và nhỏ hơn (4-dU), không có hiện tượng tự tương quan. Khi d lớn hơn dL và nhỏ hơn dU hoặc d lớn hơn (4-dU) và nhỏ hơn (4-dL), không kết luận có hoặc không có hiện tượng tự tương quan. Khi d lớn hơn 0 và nhỏ hơn dL, có hiện tượng tự tương quan dương. Khi d lớn hơn (4-dL), có hiện tượng tự tương quan âm.

Theo Fomby, Hill và Johnson (1984), trong trường hợp d rơi vào vùng không kết luận, ta sử dụng kiểm định Durbin – Watson cải tiến.

- + Nếu $1 < d < 3$: không có tự tương quan
- + Nếu $0 < d < 1$: có tự tương quan dương
- + Nếu $3 < d < 4$: có tự tương quan âm
- Kiểm định hiện tượng phương sai phần dư thay đổi:

Phương sai thay đổi (Heteroskedasticity) gây ra nhiều hậu quả với mô hình ước lượng bằng phương pháp OLS. Nó làm cho các ước lượng của các hệ số hồi quy không chệch không hiệu quả, ước lượng của phương sai bị chệch làm kiểm định của các giả thuyết mất hiệu lực, dễ đánh giá nhầm về chất lượng của mô hình hồi quy tuyến tính. Có nhiều kiểm định như White, Glejser... Và có 1 kiểm định cũng đơn giản đó là kiểm định tương quan hạng Spearman.

2.2 - XÂY DỰNG BỘ DỮ LIỆU:

Dữ liệu giá vàng 10 năm (2008 – 2018) và 5 biến giải thích được thu thập từ nhiều nguồn khác nhau. Dữ liệu liên quan đến giá vàng được thu thập từ trang web chính thức của Kaggle.com. Các biến giải thích là Date, SPX, GLD, USO, SLV, EUR/USD.

Bộ dữ liệu bao gồm:

- Date: Ngày đưa ra giá.
- SPX: giá thị trường chứng khoán bao gồm 500 cổ phiếu của các công ty lớn được giao dịch trên thị trường chứng khoán Hoa Kỳ.
- GLD: giá vàng:

Đồng USD cũng là một trong các yếu tố ảnh hưởng đến giá vàng Việt Nam. Theo các chuyên gia, giá vàng và tiền tệ có mối quan hệ tỷ lệ nghịch với nhau. Cụ thể, khi đồng USD tăng, giá vàng sẽ giảm xuống và ngược lại.

Sự suy giảm tiền tệ ảnh hưởng đến nền kinh tế, kéo theo đó là sự suy giảm niềm tin của các nhà đầu tư. Họ sẽ dần chuyển sang dòng tiền tệ khác hoặc dùng vàng làm công cụ trao đổi hàng hóa.

- USO: giá quỹ dầu mỏ hoa kỳ.

Đối với bất ổn về kinh tế và địa chính trị nhu cầu mua vàng tăng cao khiến giá vàng tăng. Trong trường hợp này, giá dầu tăng khiến giá vàng tăng.

Mặt khác, cả giá dầu thô và giá vàng đều được định giá bằng đồng dollar Mỹ và chúng biến thiên ngược chiều với giá đồng USD.

- SLV: giá Bạc.
- EUR/USD: chênh lệch tỷ giá giữa Euro và Đô la Mỹ.

| Date | SPX | GLD | USO | SLV | EUR/USD |
|-----------|-------------|-----------|-----------|-----------|----------|
| 1/2/2008 | 1447.160034 | 84.860001 | 78.470001 | 15.18 | 1.471692 |
| 1/3/2008 | 1447.160034 | 85.57 | 78.370003 | 15.285 | 1.474491 |
| 1/4/2008 | 1411.630005 | 85.129997 | 77.309998 | 15.167 | 1.475492 |
| 1/7/2008 | 1416.180054 | 84.769997 | 75.5 | 15.053 | 1.468299 |
| 1/8/2008 | 1390.189941 | 86.779999 | 76.059998 | 15.59 | 1.557099 |
| 1/9/2008 | 1409.130005 | 86.550003 | 75.25 | 15.52 | 1.466405 |
| 1/10/2008 | 1420.329956 | 88.25 | 74.019997 | 16.061001 | 1.4801 |
| 1/11/2008 | 1401.02002 | 88.580002 | 73.089996 | 16.077 | 1.479006 |
| 1/14/2008 | 1416.25 | 89.540001 | 74.25 | 16.280001 | 1.4869 |

Hình 2. 4: dữ liệu giá của vàng và các tài nguyên liên quan

Dữ liệu trên đã được phân loại và làm sạch.

2.3 - ÁP DỤNG THUẬT TOÁN VÀO BÀI TOÁN:

2.3.1 – Đặt ra bài toán:

Chúng ta cùng đi đến một bài toán về giá Vàng sau:

Bạn đang có nhu cầu cần mua Vàng nhưng bạn chưa biết nên mua với giá bao nhiêu và khi nào cho hợp lý. Nhưng bạn lại có một vài thông tin như: giá cổ phiếu SPX là 2725.78 USD, giá quỹ dầu mỏ của Hoa Kỳ là 122.54 USD, giá của bạc là 14,4 USD và tỷ giá của EUR/USD là 1.182 (giả sử rằng giá tiền của Vàng thuộc phần lớn vào yếu tố giá dầu mỏ của Hoa Kỳ, chênh lệch tỷ giá giữa EUR/USD, giá cổ phiếu SPX, giá bạc)... Vậy làm sao để biết vàng khi có x_1 giá dầu mỏ, x_2 chênh lệch tỷ giá EUR/USD, x_3 giá cổ phiếu SPX, x_4 giá bạc có giá như thế nào? Vấn đề này có thể giải quyết bằng thuật toán Linear Regression.

2.3.2 – Áp dụng thuật toán Hồi quy tuyến tính:

Để giải quyết bài toán này, đầu tiên chúng ta cần tìm ra giá vàng bị ảnh hưởng bởi các yếu tố nào. Và nhận ra giá trị vàng phụ thuộc rất nhiều vào giá dầu mỏ của Hoa Kỳ, chênh lệch tỷ giá giữa EUR/USD, giá cổ phiếu SPX, giá bạc. Từ đó chúng ta sẽ sử dụng toán học để mô hình hóa vấn đề, các yếu tố sẽ được biểu diễn thông qua các biến.

Chúng ta có thể thấy rằng:

- Giá cổ phiếu SPX càng lớn thì giá nhà càng cao
- Giá bạc càng lớn giá nhà càng cao
- Giá dầu mỏ tăng thì giá vàng cũng tăng theo

Ta sẽ biểu diễn chúng thông qua biểu thức sau:

$$y \approx f(\mathbf{x}) = \hat{y}$$

$$f(\mathbf{x}) = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_0$$

- với y : là một số vô hướng (scalar) biểu diễn output (tức giá của vàng), $\mathbf{x} = [x_1, x_2, x_3, x_4]$: một vector hàng chứa thông tin input và $\mathbf{w} = [w_0, w_1, w_2, w_3, w_4]$ là vector bao gồm các hệ số tối ưu.
- \hat{y} là giá trị mà thuật toán dự đoán được.

Vấn đề đặt ra: Liệu đường thẳng ta vẽ có thể đi qua toàn bộ các điểm cho trước hay không? Câu trả lời là có, nhưng trong thực tế rất khó xảy ra trường hợp này. Vì vậy, khi có một hay nhiều điểm không cùng thuộc một đường thẳng, ta cần tìm một đường thẳng sao cho nó gần với các điểm nhất có thể, hay độ lệch của điểm đó so với đường thẳng là nhỏ nhất. Từ đó ta có công thức tính sai số dự đoán:

$$y \approx f(\mathbf{x}) = \hat{y}$$

$$f(\mathbf{x}) = w_1x_1 + w_2x_2 + w_3x_3 + w_0$$

Việc sử dụng e^2 vì có thể $y - \hat{y}$ có thể âm và thuận tiện cho việc đạo hàm ở dưới.

Và để độ lệch hay sai số của mọi điểm so với đường thẳng là nhỏ nhất, ta sẽ có tổng sai số là nhỏ nhất, từ đó ta có hàm là tổng các sai, hàm này có tên là hàm mất mát.

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N (y_i - \bar{\mathbf{x}}_i \mathbf{w})^2$$

Chúng ta luôn mong muốn rằng sự mất mát (sai số) là nhỏ nhất, đồng nghĩa với việc tìm các hệ số \mathbf{w} sao cho giá trị của hàm mất mát này càng nhỏ càng tốt. Ta có thể thực hiện

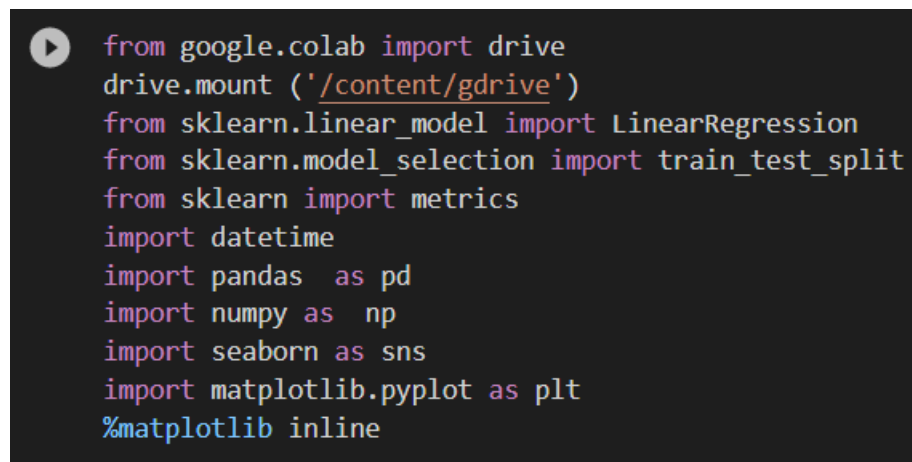
bằng cách đạo hàm, rồi từ đó tìm ra các điểm cực tiểu cục bộ, và khi tìm được toàn bộ điểm cực bộ rồi so sánh với nhau, ta sẽ tìm ra được điểm cực tiểu toàn phần. Từ đó giải quyết được bài toán. Sau khi đạo hàm ta được w thỏa mãn yêu cầu:

$$\mathbf{w} = \mathbf{A}^\dagger \mathbf{b} = (\bar{\mathbf{X}}^T \bar{\mathbf{X}})^\dagger \bar{\mathbf{X}}^T \mathbf{y}$$

2.4 - THỰC NGHIỆM VỚI THƯ VIỆN PYTHON:

2.4.1 – Import thư viện:

Trước khi thực nghiệm bài toán với thư viện phải chắc chắn chúng ta đã import đầy đủ các thư viện của python.

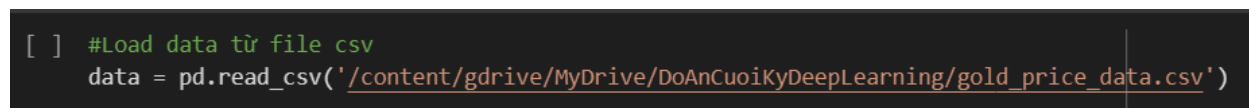


```
from google.colab import drive
drive.mount('/content/gdrive')
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn import metrics
import datetime
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

Hình 2. 5: thư viện cần dùng để xử lý, phân tích và trực quan dữ liệu

2.4.2 – Đọc dữ liệu:

Load tập dữ liệu đã lưu ở google drive, kết nối google vào google colab để truy cập dữ liệu.



```
[ ] #Load data từ file csv
data = pd.read_csv('/content/gdrive/MyDrive/DoAnCuoikyDeepLearning/gold_price_data.csv')
```

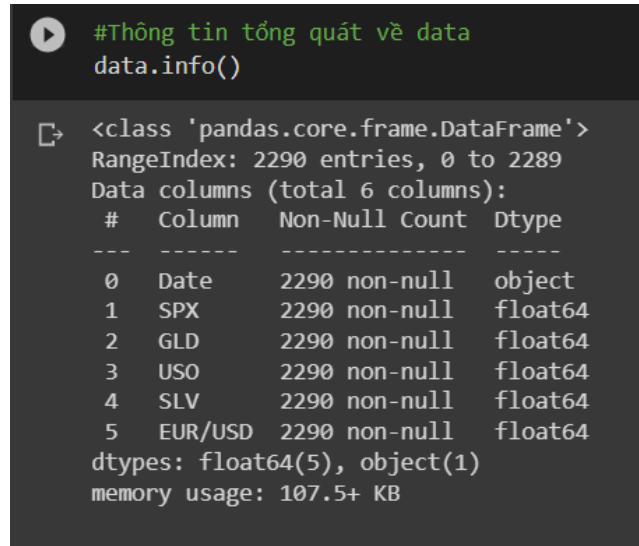
Hình 2. 6: tải dữ liệu từ google lên colab

Trước khi bắt đầu phân tích một tập dữ liệu, ta phải tìm hiểu thông tin về dữ liệu đó, ý nghĩa nội dung các cột, dòng. Thường là các thông tin này được đính kèm bổ sung khi ta

được cung cấp dữ liệu từ 1 công ty, doanh nghiệp hoặc các nguồn cung cấp khác như kaggle.

2.4.3 – Kiểm tra và phân tích dữ liệu:

Xem các thông tin tổng quát về dữ liệu.



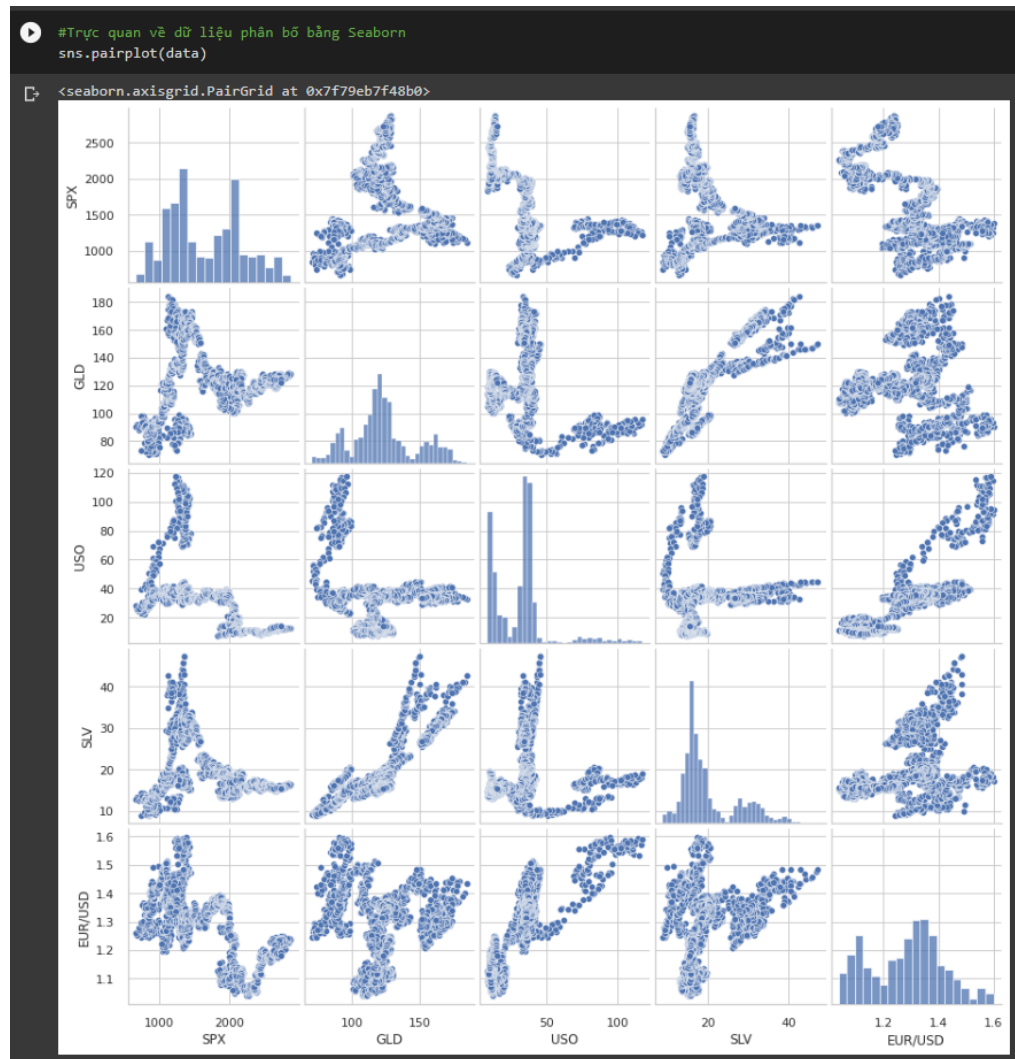
```
#Thông tin tổng quát về data
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2290 entries, 0 to 2289
Data columns (total 6 columns):
#   Column      Non-Null Count  Dtype
---  -
0    Date        2290 non-null   object
1    SPX          2290 non-null   float64
2    GLD          2290 non-null   float64
3    USO          2290 non-null   float64
4    SLV          2290 non-null   float64
5    EUR/USD      2290 non-null   float64
dtypes: float64(5), object(1)
memory usage: 107.5+ KB
```

Hình 2. 7: thông tin tổng quát về dữ liệu

Dữ liệu trên ta có 2290 dòng tương ứng với dữ liệu 2290 ngày mà giá vàng chốt cuối ngày, Với dữ liệu này ta đủ để xây dựng một mô hình máy học dự đoán giá nhà nằm trong các khu vực trên.

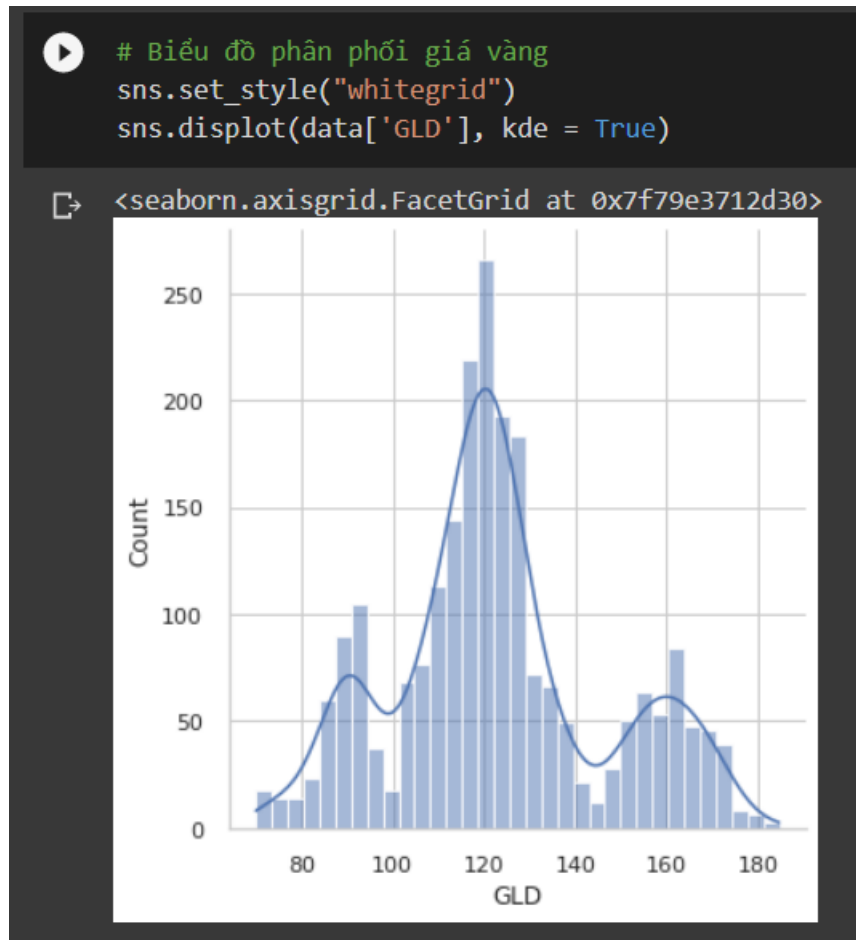
Giờ ta sẽ xem trực quan về dữ liệu được phân bố bằng Seaborn.



Hình 2. 8: biểu đồ Seaborn của dữ liệu

Về tương quan giữa các cột, ta thấy Cột GLD có kiểu phân tán theo mô hình tuyến tính, dựa trên thông tin này, ta xây dựng mô hình máy học hồi quy tuyến tính để dự đoán nó dựa trên giá trị các cột khác, trừ cột ngày (Date).

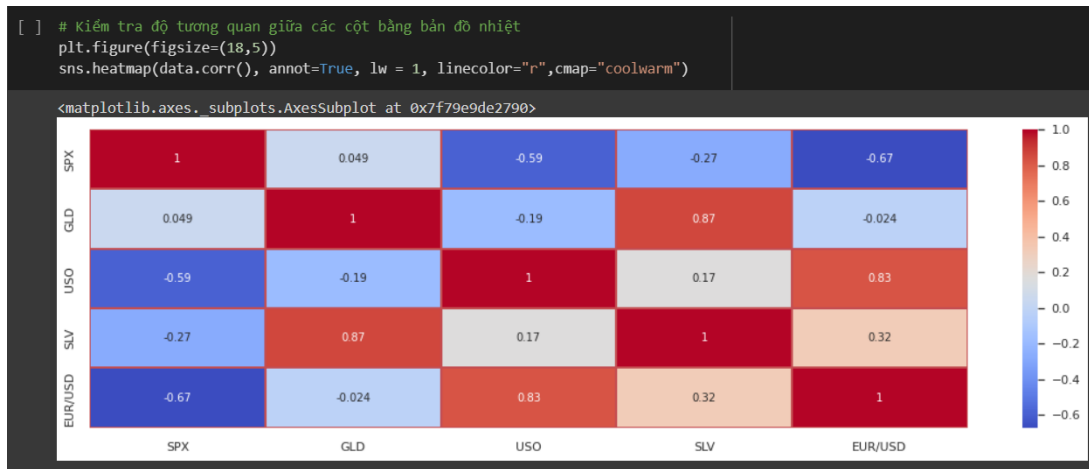
Tiếp theo ta sẽ xem biểu đồ phân phối giá Vàng



Hình 2. 9: biểu đồ phân phối giá Vàng

Từ biểu đồ ta có thể thấy giá vàng đã bán từ năm 2008 đến năm 2018 thường tập trung ở mức giá 80 đến 180USD, và nhiều nhất là 120 USD trên 1 Ounce.

Kiểm tra mức độ tương quan giữa các cột bằng bản đồ nhiệt



Hình 2. 10: bản đồ nhiệt của dữ liệu

Qua đó, ta phân tích được các cột có giá trị tương quan như thế nào với nhau. Sau khi đã phân tích sơ qua về dữ liệu chúng ta sẽ bắt đầu xây dựng mô hình dự đoán giá vàng.

CHƯƠNG 3: XÂY DỰNG ỨNG DỤNG BẰNG NGÔN NGỮ PYTHON

3.1 – XÂY DỰNG ỨNG DỤNG VÀ GIẢI THÍCH:

3.1.1 – Xây dựng mô hình dự đoán, tạo model và huấn luyện:

- Phân tách dữ liệu thành train và test:

Bây giờ chúng ta hãy bắt đầu đào tạo mô hình hồi quy. Trước tiên, chúng ta sẽ cần tách dữ liệu của mình thành một mảng X chứa các tính năng cần đào tạo (các biến độc lập) và một mảng y với biến mục tiêu (biến phụ thuộc), trong trường hợp này là cột Giá. Chúng ta sẽ loại bỏ cột “Date” vì nó chỉ có thông tin văn bản mà mô hình hồi quy tuyến tính không thể sử dụng.

```
[ ] #Phân tách dữ liệu thành train và test
    x = data[['SPX', 'USO', 'SLV', 'EUR/USD']]#mảng X chứa các tính năng cần đào tạo (các biến độc lập)
    y = data['GLD']#mảng y chứa biến mục tiêu (biến phụ thuộc)
```

Hình 3. 1: phân tách dữ liệu

Giờ ta đã có hai biến X, y theo yêu cầu của mô hình, hai biến này dựa trên dữ liệu là ta có được để đào tạo mô hình.

Tiếp theo ta tách các biến trên thành giá trị train và test, hai giá trị này chúng ta sẽ luôn gặp và sử dụng trong quá trình xây dựng mô hình máy học

Sau đó ta tạo 4 biến, gồm X_train, y_train và X_test, y_test.

```
[ ] #tạo 4 biến, gồm X_train, y_train và X_test, y_test
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=101)
```

Hình 3. 2: tạo biến chứa các dữ liệu

Với đối số truyền vào là giá trị X, y ta đã lấy từ dữ liệu bên trên, test_size trả về cho ta phần trăm dữ liệu được chia, ví dụ 0.3 tương ứng với dữ liệu được chia thành 30% giá trị là test, còn lại là dữ liệu train. random_state bằng một số tương ứng nào đó để đảm bảo mỗi lần ta chạy lại mô hình, giá trị phân tách ngẫu nhiên nhận được là giống nhau, chúng ta có thể cho số nào bất kỳ.

Giờ ta có thể xem qua dữ liệu train và test vừa tạo.

```
[ ] #dữ liệu train và test ta vừa tạo
    print(X_train)

[ ] print(y_train)

[ ] print(X_test)
```

Hình 3. 3: in ra các dữ liệu vừa tạo mới

- **Tạo model và training Linear Regression:**

Tiến hành huấn luyện dữ liệu bằng phương thức fit mà hồi quy tuyến tính có sẵn trong thư viện. Sau khi huấn luyện kết quả trả về là một hàm `LinearRegression()` là hoàn thành.

```
[16] #Huấn luyện mô hình với hàm hồi quy tuyến tính
      lm = LinearRegression()
      lm.fit(X_train, y_train)

      LinearRegression()
```

Hình 3. 4: hàm huấn luyện hồi quy tuyến tính

3.1.2 – Dự đoán và đánh giá mô hình:

- **Dự đoán:**

Để dự đoán và kiểm tra mô hình, ta sử dụng dữ liệu test bên trên mà ta đã tách ra. Trong đó, `X_test` là các tính năng mà mô hình chưa biết, `y_test` là kết quả biết trước để ta so sánh với kết quả dự đoán từ `X_test`.

Lấy kết quả dự đoán từ `X_test`, ta dùng phương thức `predict()` truyền đối số `X_test` vào.

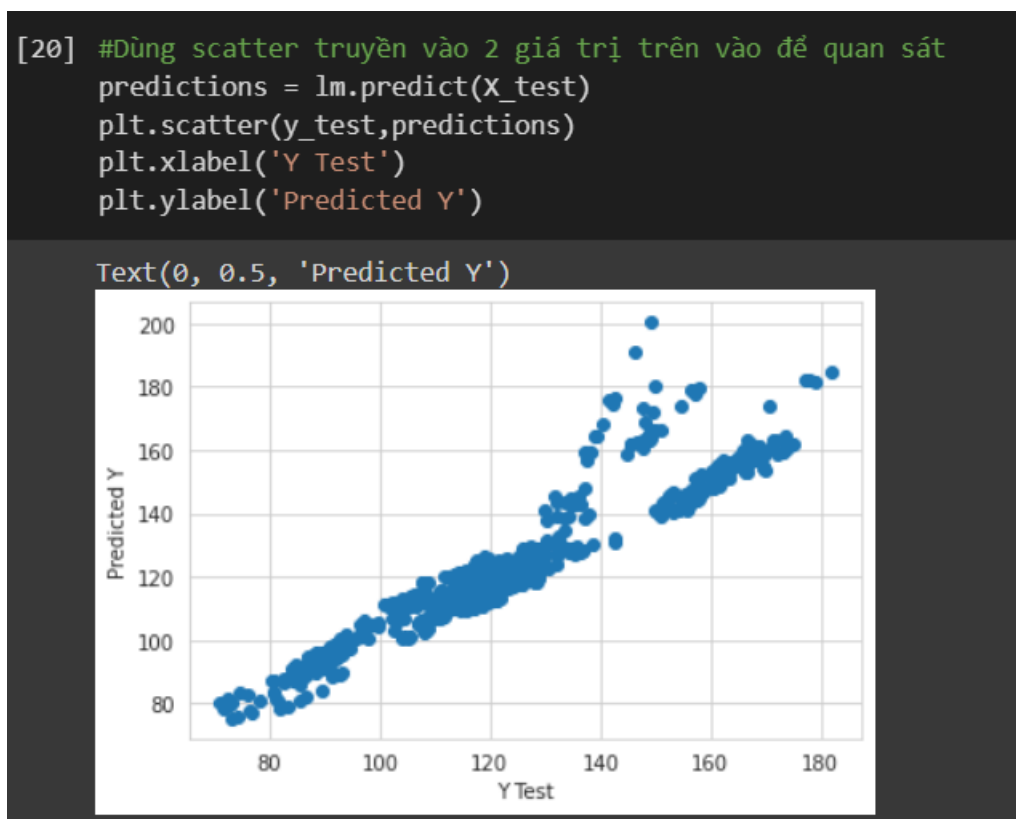
```
#Lấy kết quả dự đoán từ X_test, dùng phương thức predict() truyền đối số X_test vào
predictions = lm.predict(X_test)
print(predictions)
```

| | | | | |
|--------------|--------------|--------------|--------------|--------------|
| 122.6634602 | 107.02064719 | 121.66121847 | 120.09488694 | 116.4326306 |
| 124.91152566 | 142.68315001 | 121.3554871 | 104.16659794 | 120.72668271 |
| 116.39543443 | 114.57415388 | 145.09343279 | 114.50002329 | 114.43860046 |
| 96.0529044 | 159.39105084 | 119.64313568 | 113.40898538 | 163.06658376 |
| 140.31998636 | 110.52003642 | 131.6651 | 122.47306921 | 120.07886399 |

Hình 3. 5: kết quả dự đoán

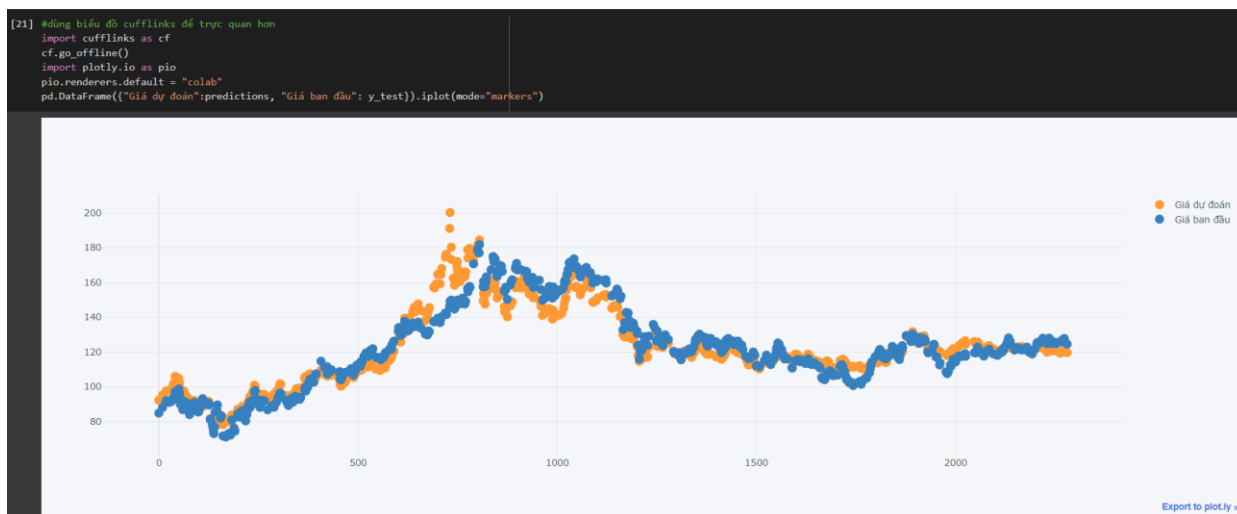
Kết quả dự đoán bên trên là một mảng trong numpy chứa kết quả dự đoán từ giá trị X_{test} , để kiểm tra kết quả dự đoán (predictions) và kết quả ban đầu (y_{test}) xem mô hình ta như thế nào. Ta có thể trực quan quan sát bằng biểu đồ.

Sử dụng biểu đồ phân tán (scatter) truyền vào 2 giá trị trên vào để quan sát:



Hình 3. 6: biểu đồ phân tán của kết quả dự đoán

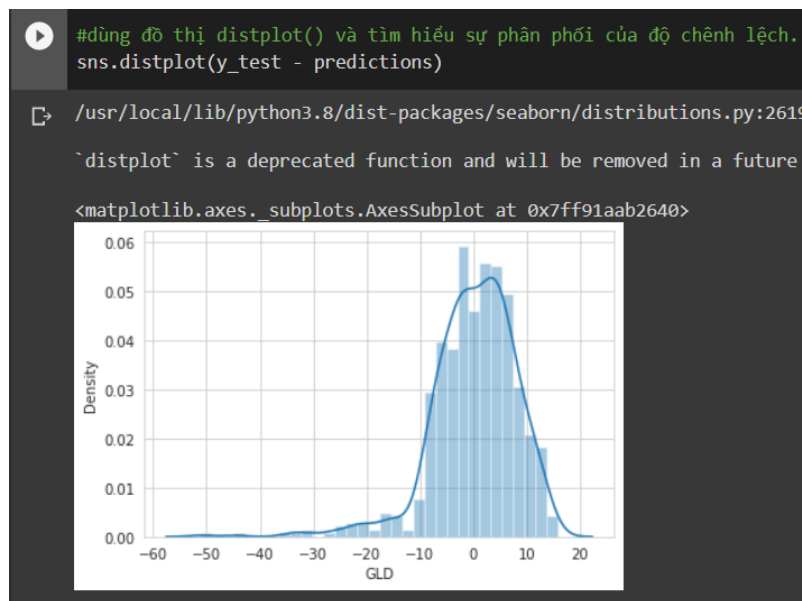
Sử dụng biểu đồ cufflinks để trực quan hơn



Hình 3. 7: biểu đồ cufflinks của kết quả dự đoán

Nhìn trực quan bên trên ta thấy giữa giá dự đoán (màu cam) và giá ban đầu (màu xanh dương) có 1 sự chênh lệch. Tùy theo độ chính xác mô hình, nếu độ chính xác mô hình càng cao thì độ chênh lệch các điểm trên biểu đồ càng ít lại.

Để trực quan độ chênh lệch bằng biểu đồ, ta dùng vẽ đồ thị `distplot()` trong Seaborn và tìm hiểu sự phân phối của độ chênh lệch này.

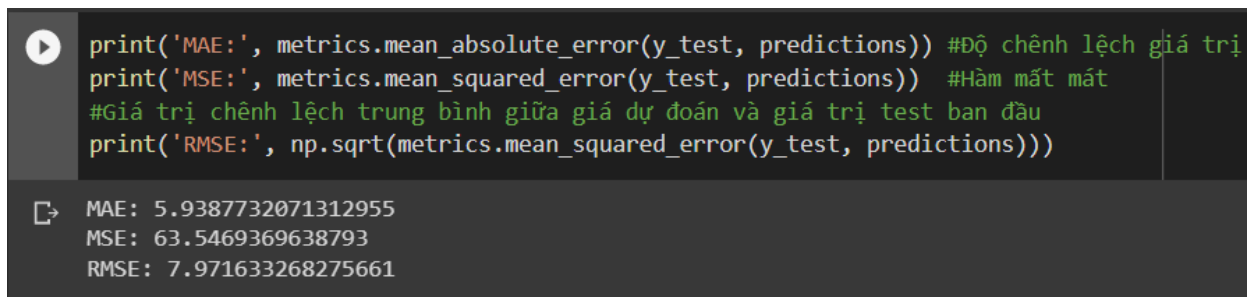


Hình 3. 8: biểu đồ distplot thể hiện sự chênh lệch

Nhìn vào biểu đồ trên, bạn thấy giá trị chênh lệch giữa giá vàng dự đoán (predictions) và giá vàng thực tế ban đầu (y_test), phân bố tập trung ở 0 và trên dưới 10 USD/ounce, chứng tỏ mô hình của chúng ta có độ chính xác tương đối cao và hợp lý khi kết quả dự đoán và kết quả ban đầu có sự chênh lệch thấp và phần lớn dao động trong khoảng + (-) 10%.

- **Đánh giá:**

Bài toán Regression – hồi quy tức là biến yy của chúng ta không phải là một giá trị rời rạc mà là một giá trị liên tục. Nó thường là số lượng, giá tiền, nhiệt độ, lượng mưa ... Do nó là giá trị liên tục nên chúng ta hoàn toàn không thể sử dụng độ chính xác để đo performance của mô hình được mà cần phải dùng một số loại độ đo khác.



```
print('MAE:', metrics.mean_absolute_error(y_test, predictions)) #Độ chênh lệch giá trị
print('MSE:', metrics.mean_squared_error(y_test, predictions)) #Hàm mất mát
#Giá trị chênh lệch trung bình giữa giá dự đoán và giá trị test ban đầu
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, predictions)))
```



```
MAE: 5.9387732071312955
MSE: 63.5469369638793
RMSE: 7.971633268275661
```

Hình 3. 9: chỉ số đánh giá mô hình dự đoán

Ta thấy, sử dụng chỉ số RMSE, cho thấy giá trị chênh lệch trung bình của giá vàng dự đoán từ mô hình và giá trị vàng thực tế là xấp xỉ 8 USD/ounce.

Giải thích các chỉ số trên:

- độ chênh lệch giá trị (MAE): là một phương pháp đo lường sự khác biệt (độ chênh lệch giá trị) giữa hai biến liên tục. Giả sử rằng X và Y là hai biến liên tục thể hiện kết quả dự đoán của mô hình và kết quả thực tế, đây là chỉ số dễ hiểu nhất, vì đó là giá trị chênh lệch trung bình.

- hàm mất mát (MSE): là giá trị trung bình của bình phương sai số (Hàm mất mát), là sự khác biệt giữa các giá trị được mô hình dự đoán và giá trị thực. MSE cũng được gọi là một hàm rủi ro, tương ứng với giá trị kỳ vọng của sự mất mát sai số bình phương hoặc mất mát bậc hai chỉ số này phổ biến hơn chỉ số MAE.

- căn bậc hai của giá trị trung bình của các sai số bình phương (RMSE): là căn bậc hai của giá trị trung bình của các sai số bình phương (MSE). Thông thường, ta thường dùng chỉ số này để xác định giá trị chênh lệch trung bình giữa giá dự đoán và giá trị test ban đầu.

Kiểm tra độ chính xác của mô hình dựa trên phương sai

```
[24] #kiểm tra độ chính xác mô hình dựa trên phương sai MSE
      metrics.explained_variance_score(y_test, predictions)

0.8864383929571654
```

Hình 3. 10: kết quả dự đoán mô hình dựa trên phương sai

Giá trị trả về tốt nhất cho mô hình là 1.0 vậy với mô hình trên, giá trị trả về đạt xấp xỉ 0.89, tương ứng với 89% hiệu quả của mô hình đào tạo.

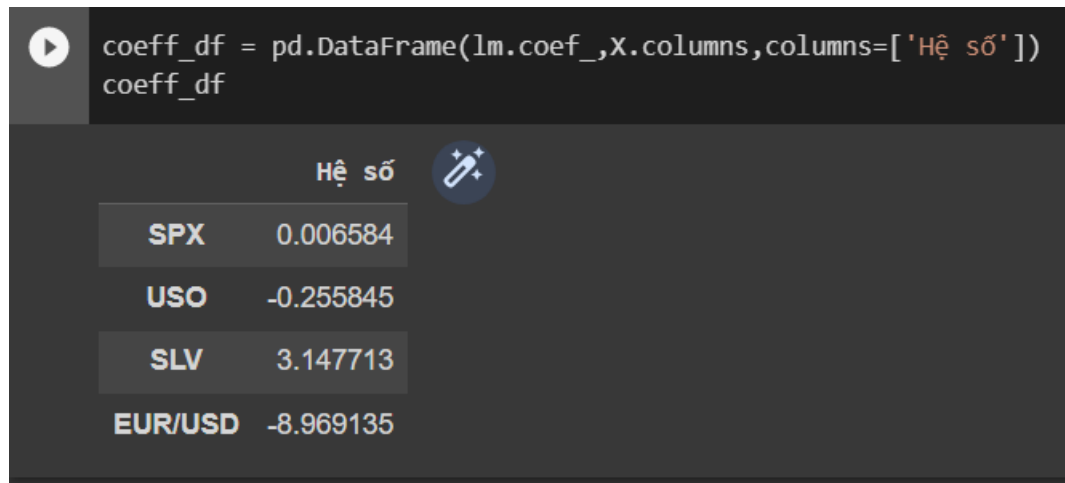
Hệ số coeff: để đánh giá sự tác động của các tính năng (các biến độc lập) lên kết quả đầu ra (biến phụ thuộc), ta sử dụng hệ số Coeff. Hệ số này cho ta biết khi giá trị biến độc lập thay đổi 1 đơn vị, thì giá trị đầu ra sẽ thay đổi như thế nào.

```
[25] #đánh giá sức tác động của các tính năng ( các biến độc lập) lên kết quả đầu ra (biến phụ thuộc)
      #sử dụng hệ số Coeff
      print(lm.coef_)

[ 6.58434327e-03 -2.55844584e-01  3.14771276e+00 -8.96913497e+00]
```

Hình 3. 11: đánh giá mô hình bằng hệ số coeff

Để hiểu được giá trị của hệ số coeff ta dùng DataFrame



Hình 3. 12: kết quả khi tăng một đơn vị trong cột

Diễn giải các hệ số trên:

- Giữ tất cả các tính năng khác không thay đổi, khi tăng 1 đơn vị trong Cột SPX thì sẽ tăng 0.006584USD trong giá Vàng.
- Giữ tất cả các tính năng khác không thay đổi, khi tăng 1 đơn vị trong Cột USO thì sẽ giảm 0.255845USD trong giá Vàng.
- Giữ tất cả các tính năng khác không thay đổi, khi tăng 1 đơn vị trong Cột SLV thì giá sẽ tăng 3.147713USD trong giá Vàng.
- Giữ tất cả các tính năng khác không thay đổi, khi tăng 1 đơn vị trong Cột EUR/USD sẽ giảm 8.969135USD trong giá Vàng.

3.2 – KẾT LUẬN:

3.2.1 – KẾT QUẢ ĐẠT ĐƯỢC:

Qua kết quả xây dựng mô hình, tạo model, dự đoán và đánh giá mô hình. Chúng ta được kết quả như sau:

- Giá trị chênh lệch trung bình ≈ 6 USD.
- Giá trị trung bình của bình phương sai số ≈ 64 USD.

- Giá trị chênh lệch trung bình giữa giá dự đoán và giá test ban đầu ≈ 8 USD.
- Hiệu quả của mô hình huấn luyện là 88.6%.

Hệ số coeff (biến động tương đối của những tập hợp dữ liệu chưa phân tổ có giá trị bình quân khác nhau):

- Giá SPX tăng khoảng 0.006584 USD/Ounce.
- Giá USO giảm khoảng 0.255845 USD/Ounce.
- Giá SLV tăng khoảng 3.147713 USD/Ounce.
- Tỷ giá của EUR/USD giảm 8.969135 USD/Ounce.

3.2.2 – HẠN CHẾ VÀ HƯỚNG PHÁT TRIỂN:

- **Hạn chế:**

Hạn chế đầu tiên của Linear Regression là nó rất nhạy cảm với nhiễu (sensitive to noise). Vì vậy, trước khi thực hiện Linear Regression, các nhiễu (outlier) cần phải được loại bỏ. Bước này được gọi là tiền xử lý (pre-processing).

Hạn chế thứ hai của mô hình là nó không biểu diễn được các mô hình phức tạp. Mặc dù trong phần trên, chúng ta thấy rằng phương pháp này có thể được áp dụng nếu quan hệ giữa outcome và input không nhất thiết phải là tuyến tính nhưng mỗi quan hệ này vẫn đơn giản nhiều so với các mô hình thực tế.

Bản chất của việc nâng bậc mô hình là chuyển vector thuộc tính X từ một vector có số chiều nhỏ thành một vector có số chiều lớn hơn rất nhiều theo quy luật số mũ. Nếu số lượng dữ liệu của chúng ta nhỏ mà số chiều của vector thuộc tính lại quá lớn dẫn đến việc giải quyết bài toán hồi quy với số chiều cao và hiển nhiên là độ chính xác sẽ rất thấp. Điều đó cho chúng ta thấy cần phải sàng lọc và lựa chọn thuộc tính thật tốt trước khi đưa vào mô hình.

- **Hướng phát triển:**

Qua việc nâng bậc của mô hình hồi quy có thể tìm ra được sự phụ thuộc kết hợp giữa các Feature đến giá của Vàng. Tuy nhiên nếu tập dữ liệu quá nhỏ so với số lượng Feature thì khi nâng bậc sẽ dẫn đến bài toán tối ưu với số chiều cao và làm giảm độ chính xác của mô hình. Điều đó cho thấy việc lựa chọn khéo léo các Feature là vô cùng cần thiết trước khi áp dụng bất kì mô hình nào.

Trong tương lai, với bài toán này, em sẽ cố gắng tiếp tục xây dựng lại model from scratch và không chỉ dừng lại ở mức dự đoán giá Vàng, em nghĩ sẽ hoàn toàn có thể tính được khối lượng giao dịch của Vàng để điều chỉnh lại giá cả dự đoán Vàng sao cho hợp lý. Em vẫn sẽ tiếp tục cải thiện bài toán này và tiếp tục làm những dự định kia, nhưng giờ em phải kiếm được dữ liệu đã.

TÀI LIỆU THAM KHẢO

1. <https://pythonnangcao.com/khoa-hoc/>
2. <https://www.kaggle.com/>
3. Thầy Hồ Khôi, slide bài giảng môn DeepLearning trong khoa học dữ liệu, khoa CNTT, trường đại học Nguyễn Tất Thành.

LINK SOURCE CODE:

https://drive.google.com/drive/folders/1eug9-tHjcUbfcN6XY9V3wErgmYt1Uj4?usp=share_link