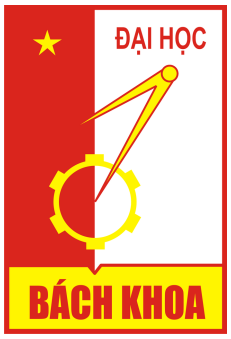
ĐẠI HỌC BÁCH KHOA HÀ NỘI
KHOA TOÁN - TIN



XÂY DỰNG MÔ HÌNH ĐẦU TƯ TRỰC TUYẾN TRÊN THỊ TRƯỜNG TIỀN MÃ HÓA

ĐỒ ÁN TỐT NGHIỆP

Chuyên ngành: TOÁN ỨNG DỤNG

Chuyên sâu: Các phương pháp tối ưu

Giáo viên hướng dẫn: PGS.TS. NGUYỄN THỊ THU THỦY
Sinh viên thực hiện: ĐOÀN MINH BẢO
Mã sinh viên: 20190111

HÀ NỘI – 2024

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

1. Mục tiêu và nội dung của đồ án

(a) Mục tiêu: Đề tài đồ án nghiên cứu về bài toán tối ưu hóa lợi nhuận cho nhà đầu tư Crypto.

(b) Nội dung: Giới thiệu bài toán tối ưu hóa lợi nhuận cho nhà đầu tư Crypto. Xây dựng danh mục đầu tư trên cơ sở lý thuyết chuỗi thời gian. Xây dựng các chiến lược đầu tư sử dụng các thuật toán và mô hình học máy nhằm tối ưu hóa lợi nhuận.

2. Kết quả đạt được

(a) Xây dựng danh mục đầu tư tối ưu.

(b) Xây dựng các chiến lược đầu tư nhằm tối ưu hóa lợi nhuận.

(c) Kiểm định các mô hình đầu tư.

(d) Quan sát sự phụ thuộc về giá của các đồng coin khác vào BTC.

3. Ý thức làm việc của sinh viên:

(a) Có ý thức, trách nhiệm cao trong quá trình học tập và làm đồ án.

(b) Chịu khó học hỏi và tìm hiểu những kiến thức chuyên sâu liên quan đến đồ án.

(c) Có khả năng dịch thuật, nghiên cứu tài liệu chuyên ngành. Hoàn thành tốt Đồ án theo đúng yêu cầu của giáo viên hướng dẫn.

Hà Nội, ngày 30 tháng 05 năm 2024
Giảng viên hướng dẫn

PGS.TS. Nguyễn Thị Thu Thủy

Lời cảm ơn

Trước hết, em xin gửi lời cảm ơn chân thành đến PGS.TS. Nguyễn Thị Thu Thủy, Khoa Toán - Tin, Đại học Bách khoa Hà Nội. Cô đã tận tình hướng dẫn, góp ý, chỉ bảo em trong suốt quá trình em thực hiện Đồ án tốt nghiệp. Không chỉ giới thiệu, định hướng một cách tổng quát về lý thuyết chuỗi thời gian nói riêng và lý thuyết toán thống kê nói chung, cô còn dành thời gian để hướng dẫn em cách trình bày, các hành văn sao cho đúng, đồng thời chỉ ra những chỗ chưa được để em có thể kịp thời khắc phục. Được cô hướng dẫn, em không chỉ học được những kiến thức chuyên môn vô cùng sâu sắc mà đó còn là cả phong cách làm việc chuẩn mực, chuyên nghiệp, thái độ làm việc tập trung, nghiêm túc. Bên cạnh đó, cô còn truyền lại cho em những kinh nghiệm, những trải nghiệm quý báu mà cô đúc kết lại được sau nhiều năm nghiên cứu. Đó là niềm vinh dự và động lực rất lớn để em có thể vững chân trên con đường mình chọn. Một lần nữa, em xin gửi lời cảm ơn chân thành và sâu sắc nhất tới cô.

Tiếp theo, em xin gửi lời cảm ơn đến các anh chị và các bạn trong nhóm seminar đồ án do cô Thủy hướng dẫn trong học kỳ này. Tham gia các buổi seminar, em đã học được nhiều kiến thức và kinh nghiệm của mọi người để có thể hoàn thành đồ án này.

Lời cuối cùng, em cũng xin cảm ơn các Thầy Cô Khoa Toán - Tin, Đại học Bách khoa Hà Nội đã giảng dạy, hướng dẫn em, giúp em có nền tảng kiến thức đủ tốt để có thể hoàn thiện đồ án này.

Hà Nội, ngày 30 tháng 05 năm 2024

Tác giả đồ án

Đoàn Minh Bảo

Mục lục

Mở đầu

Đặt vấn đề

Hiện nay, khi đời sống người Việt được nâng cao, nhu cầu đầu tư vào thị trường tài chính bằng tiền nhàn rỗi ngày càng nhiều. Một trong những kênh đầu tư mới nhưng độ phủ sóng tương đối cao đó là tiền mã hóa.

Việt Nam hiện có hơn 16,6 triệu người sở hữu tiền mã hóa. Trong số này, có khoảng 31% sở hữu Bitcoin. Trong một cuộc khảo sát 389.345 người trên 26 quốc gia, Việt Nam đứng thứ 3 về chấp nhận Crypto (tiền mã hóa), sau Ấn Độ và Nigeria. Khoảng 23% dân số Việt Nam cho biết có sở hữu tài sản Crypto. Crypto là một thị trường mới, tuy rủi ro cao nhưng đồng thời mức lợi nhuận cũng rất hấp dẫn, vậy nên nó đã thu hút nhiều nhà đầu tư. Ở thị trường chứng khoán, các nhà đầu tư có khá nhiều phương thức để phân tích như sử dụng các báo cáo tài chính, sử dụng thông tin của các quỹ đầu tư lớn, đo lường giá trị nội tại ... Tuy nhiên đối với Crypto, đây là thị trường hình thành chủ yếu trên giá trị niềm tin, lại chưa được sự bảo trợ của nhà nước nên các hình thức phân tích đầu tư chưa nhiều.

Nhìn chung, có hai cách phân tích để đầu tư vào thị trường: phân tích cơ bản (sử dụng tin tức chính trị, tin tức thị trường, vốn hóa ...) và phân tích kỹ thuật (dự báo hướng của giá cả thông qua việc nghiên cứu các dữ liệu thị trường quá khứ). Song những cách phân tích này vẫn tương đối mơ hồ và thiếu chính xác.

Phân tích cơ bản thường gặp phải trường hợp tin đã cũ hoặc tin hợp lý đường giá (hay tin giả được tạo ra bởi các nhà đầu tư thâu tóm thị trường). Các lý thuyết phân tích kỹ thuật thì thay đổi liên tục theo thời gian do hành vi giá, tâm lý thị trường biến động không ngừng. Một điểm yếu lớn của 2 phương thức đề cập ở trên là mọi phân tích đều được thực hiện thủ công, số Crypto được phân tích khá ít, các cơ hội đầu tư được phát hiện chậm. Vì thế, tạo ra

một công cụ đầu tư tự động dựa trên cơ sở toán học nhằm khắc phục những điểm yếu đã đề cập ở trên là cần thiết ở thời điểm hiện tại.

Mục tiêu của đồ án

Với đồ án này, tác giả mong muốn xây dựng một mô hình đầu tư theo ngày hình thành trên cơ sở toán học mà người đọc với những kiến thức cơ bản về toán học và kinh tế có thể ứng dụng vào đầu tư. Mô hình nhằm tối ưu hóa lợi nhuận của các nhà đầu tư Crypto và quy trình sử dụng mô hình được thực hiện hoàn toàn tự động sử dụng máy tính.

Nội dung nghiên cứu

Trong khuôn khổ đồ án tốt nghiệp, em trình bày cơ sở lý thuyết của các nội dung:

1. Lựa chọn các Crypto tốt, ổn định dựa vào dữ liệu trong quá khứ.
2. Xây dựng mô hình đầu tư trực tuyến sử dụng thuật toán K-means Clustering.
3. Xây dựng mô hình đầu tư trực tuyến sử dụng mô hình LSTM.

Ngoài phần Mở đầu, Kết luận và Danh mục tài liệu tham khảo, nội dung đồ án được trình bày trong 3 chương.

Chương 1 Thuật toán lựa chọn danh mục đầu tư

Chương 2 Mô hình đầu tư trực tuyến sử dụng thuật toán K-means Clustering

Chương 3 Mô hình đầu tư trực tuyến sử dụng mô hình LSTM

Bảng ký hiệu và chữ viết tắt

Crypto	Cryptocurrency hay tiền mã hóa
Coin	Đồng tiền mã hóa
BTC	Bitcoin
$\mathbb{R}$	Tập các số thực
$\mathbb{R}^n$	Không gian Euclide n chiều
P_X	Xác suất của đại lượng X thỏa mãn điều kiện xác định
$E(X)$	Kỳ vọng của đại lượng X
σ_X	Độ lệch chuẩn của đại lượng X
$\in$	Thuộc
$\notin$	Không thuộc
$\text{mean}(R)$	Giá trị trung bình của đại lượng R
$\text{sd}(R)$	Độ lệch chuẩn của đại lượng R
$\text{cov}(R_i, R_j)$	Hiệp phương sai của đại lượng R_i và R_j
$\text{Var}(R)$	Phương sai của đại lượng R
CC	Cumulative Capital
LSTM	Long Short Term Memory

Danh sách hình vẽ

Chương 1

Thuật toán lựa chọn danh mục đầu tư

Để đầu tư tài chính nói chung hay Crypto nói riêng, nhìn chung có hai vấn đề chính cần quan tâm là đầu vào danh mục nào và đầu tư vào thời điểm nào. Một số nhà đầu tư ngắn hạn chỉ quan tâm đến thời điểm vào và thời điểm chốt lời, bất chấp danh mục đầu tư. Giao dịch như vậy có thể có lời trong ngắn hạn nhưng không bền vững cho đầu tư lâu dài, rủi ro cao. Vì vậy, có một thuật toán lựa chọn danh mục đầu tư với rủi ro thấp là vô cùng quan trọng. Trong quá khứ, đã có một số bài báo viết về vấn đề này ví dụ như bài báo của Lorenzo và Arroyo [2] hay bài báo của Majid Khedmati và Pejman Azin [3]. Dù vậy những thuật toán tương đối phức tạp và khó sử dụng. Chương này mong muốn trình bày một thuật toán dễ tiếp cận với người đọc hơn, dựa trên thuật toán phân bổ vốn đơn giản theo tỷ lệ. Chương sẽ bao gồm cơ sở lý thuyết toán học và thống kê, sau đó là nội dung thuật toán lựa chọn danh mục đầu tư.

1.1. Ý tưởng thuật toán

Trong Chương 1, ta cần quan tâm đến một đại lượng đó là tỷ suất lợi nhuận của Crypto thứ i tại thời điểm t (ngày):

$$R_{t,i} = \frac{P_{t,i} - P_{t-1,i}}{P_{t-1,i}}; t = 1, \ldots, n; i = 1, \ldots, m_0 \tag{1.1}$$

trong đó,

- $P_{t,i}$ là giá của Crypto thứ i tại ngày t (Giá của Crypto lấy theo phân tích ở Mục 2.1);
- n là số ngày ta sử dụng để chạy thuật toán;
- m_0 là số Crypto dùng cho thuật toán để từ đó lựa chọn danh mục đầu tư.

Dễ thấy:

- $R > 0$ ngụ ý đầu tư có lời sau một ngày;
- $R < 0$ ngụ ý đầu tư thua lỗ;
- $R = 0$ cho thấy giá không thay đổi, tức nhà đầu tư không lời không lỗ.

Nhìn chung, R càng lớn thì càng tốt cho các nhà đầu tư. Ta sẽ đánh giá R để quyết định có lựa chọn Crypto này vào danh mục đầu tư không.

1.2. Cơ sở lý thuyết

1.2.1. Chuỗi thời gian

Để tìm hiểu về thuật toán lựa chọn danh mục đầu, cần tìm hiểu về khái niệm chuỗi thời gian. Mục tiêu của việc phân tích kinh tế là chỉ ra cơ chế kinh tế và đưa ra các quyết sách. Vì vậy, đòi hỏi có một số lượng lớn các quan sát cho các đại lượng thích hợp để nghiên cứu mối quan hệ giữa các đại lượng này. Các quan sát này có thể được tiến hành đều đặn qua từng thời kỳ, chẳng hạn, theo từng tháng, từng quý hoặc hàng năm hoặc chỉ trong những thời điểm đặt biệt như các thời kỳ xảy ra khủng hoảng kinh tế. Dãy các quan sát này được gọi là chuỗi thời gian.

1.2.2. Chuỗi dừng

Trong phân tích hồi quy với dữ liệu chuỗi thời gian, một giả định rất quan trọng là chuỗi thời gian đang xem xét là chuỗi dừng (stationary). Nói chung, một chuỗi thời gian dừng nếu trung bình (mean) và phương sai (variance) của nó không đổi qua thời gian và giá trị hiệp phương sai (covariance) giữa hai giai đoạn chỉ phụ thuộc vào khoảng cách giữa hai giai đoạn ấy chứ không phụ thuộc vào thời gian thực sự tại đó hiệp phương sai được tính.

Vậy tại sao chúng ta lo lắng chuyện một chuỗi thời gian có dừng hay không? Nếu một chuỗi không dừng, chúng ta chỉ có thể nghiên cứu hành vi của nó cho riêng giai đoạn đang xem xét. Vì thế, mỗi chuỗi thời gian là một giai đoạn riêng biệt. Cho nên, chúng ta không thể khái quát hóa kết quả phân tích cho các giai đoạn khác. Đối với các mục đích dự báo, chuỗi không dừng sẽ không có giá trị ứng dụng thực tiễn.

1.2.3. Kiểm định tính dừng

Có nhiều cách để kiểm định tính dừng của một chuỗi thời gian. Trong đồ án này, em đề xuất sử dụng kiểm định gia tăng Dickey-Fuller (ADF). Đây là một học thuyết kiểm định có tính chính xác cao đồng thời được hỗ trợ tính toán tự động bởi nhiều công cụ. Vì vậy nên em không đề cập sâu vào lý thuyết toán (tiêu chuẩn kiểm định).

Bài toán kiểm định có cặp giả thuyết là:

- Giả thuyết H_0: Chuỗi không dừng
- Đối thuyết H_1: Chuỗi dừng với mức ý nghĩa α

Sau khi đưa dữ liệu của các Crypto vào công cụ, kết quả được đưa ra là p-giá trị. Nếu p-giá trị $< \alpha$, bác bỏ giả thuyết H_0, chấp nhận đối thuyết H_1. Ngược lại, ta chưa đủ cơ sở để bác bỏ H_0.

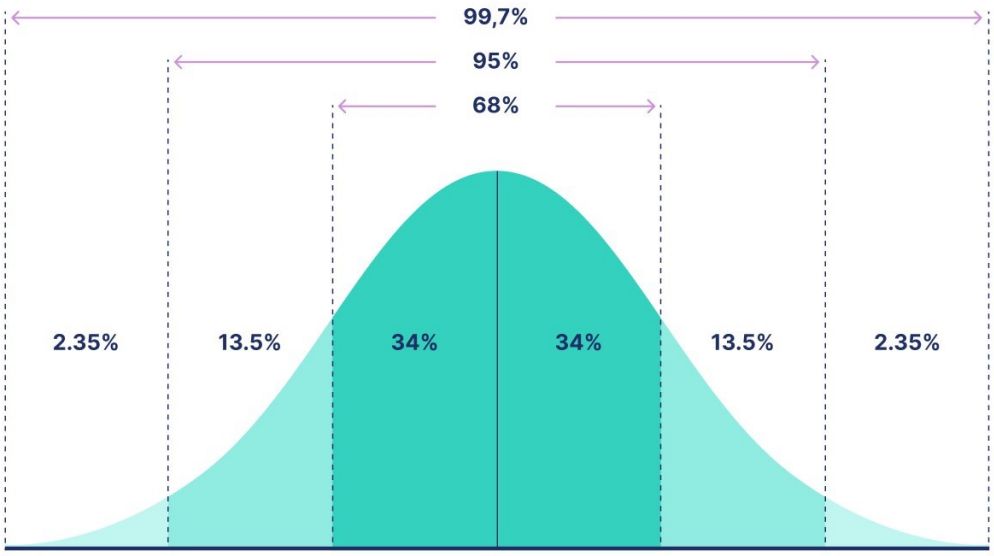
1.2.4. Giới thiệu phân phối nặng đuôi

Một phân phối khá quen thuộc đó là phân phối chuẩn. Giả sử ta quan tâm đến phân phối của đại lượng X; X có phân phối chuẩn. Khi đó,

$$(P_X|X \in [E(X) - 3\sigma_X; E(X) + 3\sigma_X]) \approx 99.7\%.$$

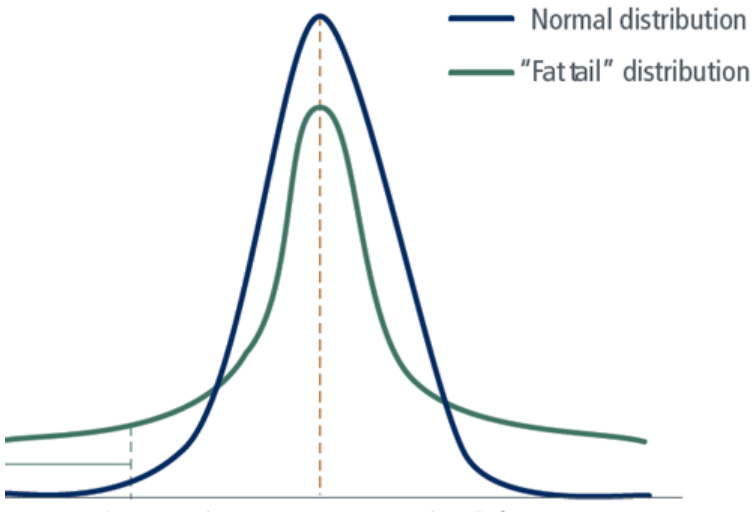
Ngược lại,

$$(P_X|X \notin [E(X) - 3\sigma_X; E(X) + 3\sigma_X]) \approx 0.3\%.$$



Hình 1.1: Phân phối chuẩn [7]

Nhìn chung, các giá trị của đại lượng ta quan tâm tập trung càng nhiều quanh giá trị trung bình của nó (mean) thì tức là đại lượng đó càng ổn định, càng dễ dự đoán. Ngược lại, phân phối nặng đuôi (hay phân phối đuôi béo) có các giá trị tập trung ở xa giá trị trung bình hơn.



Hình 1.2: Phân phối nặng đuôi [8]

Ví dụ 1.1. Giả sử:

- $\text{mean}(R) = 0.2$
- $\text{sd}(R) = 0.1$

Giả sử R tuân theo phân phối chuẩn, $(P_R, R \notin [-0.1, 0.5]) \approx 0.3\%$. Nếu R có phân phối nặng đuôi thì tỷ lệ này sẽ lớn hơn. Điều này ngụ ý rủi ro sẽ cao do tỷ lệ $R < -0.1$ tăng, thuật ngữ tài chính gọi là rủi ro đuôi béo (nặng đuôi). Tất nhiên, ngược lại, tỷ lệ $R > 0.5$ tăng sẽ mang ý nghĩa tích cực, ngụ ý tỷ suất lợi nhuận cao. Tuy nhiên, vì các nhà đầu tư thường quan tâm đến những khoản lỗ bất ngờ hơn là lãi, nên một cuộc tranh luận về rủi ro phần đuôi tập trung vào phần đuôi bên trái. Các nhà quản lý tài sản thận trọng thường thận trọng với phần đuôi liên quan đến các khoản lỗ có thể làm hỏng hoặc hủy hoại danh mục đầu tư, chứ không phải phần đuôi có lợi của các khoản lãi quá lớn.

Kiểm định phân phối nặng đuôi

Có khá nhiều định nghĩa cho phân phối nặng đuôi. Một trong những định nghĩa đơn giản nhất của phân phối nặng đuôi là phân phối có phần đuôi nặng hơn phân phối chuẩn. Giả sử đại lượng ta xét là R. Nếu,

$$(P_R | R \notin [E(R) - 3\sigma_R; E(R) + 3\sigma_R]) > 0.3\%$$

ta sẽ coi X có phân phối nặng đuôi. Song ta chỉ quan tâm đến đuôi bên trái (rủi ro), nên nếu Crypto có tỷ lệ R thỏa mãn:

$$(P_R|R \in (-\infty; E(R) - 3\sigma_R)) > p_0 = 0.15\%$$

ta sẽ loại Crypto.

1.3. Thuật toán lựa chọn danh mục đầu tư

Đầu vào: Ma trận giá của các Crypto (xem chi tiết ở Mục 2.1)
Đầu ra: Danh mục m các Crypto có thể sử dụng để đầu tư

- **Bước 1:** Lấy dữ liệu đầu vào dưới dạng ma trận giá theo ngày của các Crypto (**Ma trận** A).

  Trong chương này em đề xuất lấy dữ liệu trong khoảng một năm tính từ thời điểm hiện tại, tương ứng với $n = 365$ ngày giao dịch. Không nên lấy dữ liệu quá xa vì như đã đề cập ở phần đặt vấn đề, dữ liệu ở quá xa không còn phản ánh đúng hành vi giá ở thời điểm hiện tại.

  Tính đến thời điểm hiện tại, đã có 17,414 Crypto trên Coinbase (một trang web hỗ trợ theo dõi thông tin về thị trường tiền ảo). Theo em không nên đánh giá toàn bộ các Crypto này vì việc này làm khối lượng tính toán trở nên quá lớn. Giá trị tiền mã hóa chủ yếu là giá trị niềm tin, vì vậy vốn hóa tỷ lệ nghịch với mức độ rủi ro của Crypto. Em đề xuất sử dụng $m_0 = 32$ Crypto có vốn hóa lớn nhất (lớn hơn 1 tỷ Dollar) để đưa vào đánh giá.

  Lấy số cột là số Crypto, số hàng là số ngày lấy dữ liệu (sắp xếp theo thứ tự ngày xa nhất ở đầu). Vậy ma trận A có kích thước 365×32.

- **Bước 2:** Làm sạch dữ liệu, loại bỏ những hàng có giá trị không xác định (NaN) hoặc vô cùng (Inf). Những hàng bị lặp lại cũng sẽ bị loại bỏ.

- **Bước 3:** Từ ma trận đầu vào A, tính ra ma trận tỷ suất lợi nhuận theo ngày (**ma trận** X) theo công thức (1.1). Do không có dữ liệu để tính toán hàng đầu tiên trong ma trận X nên lấy toàn bộ giá trị hàng đầu tiên của X là 0.

- **Bước 4:** Kiểm tra tính dừng cho từng cột (Crypto) của ma trận X sử dụng hàm adfuller() trong thư viện statsmodels.tsa.stattools của Python. Lựa chọn mức ý nghĩa α phù hợp (có thể lấy $\alpha = 1\%; 5\%; 10\%$; α lấy càng nhỏ thì số lượng Crypto còn lại sau khi đánh giá càng ít). Loại bỏ những Crypto không thỏa mãn tính dừng.

- **Bước 5:** Kiểm tra rủi ro nặng đuôi trái của các Crypto. Dựa vào đó ta tính toán trung bình mean(R) và độ lệch chuẩn sd(R) của mỗi Crypto. Từ đó thống kê tỷ lệ p các giá trị R nhỏ hơn $\text{mean}(R) - 3\text{sd}(R)$. Nếu p lớn hơn tỷ lệ loại bỏ p_0, ta loại bỏ Crypto. p_0 (xem Mục 1.2.5) là 0.15%. Trường hợp không có hoặc có quá ít các Crypto còn lại sau loại bỏ, ta có thể tăng p_0 nhằm giảm số Crypto bị loại. Ngược lại nếu còn quá nhiều Crypto so với nhu cầu, ta có thể giảm p_0.

Kết thúc thuật toán, chúng ta đã có có một danh mục gồm m Crypto còn lại có thể sử dụng để đầu tư ở Chương 2 và Chương 3.

Chương 2

Mô hình đầu tư trực tuyến sử dụng thuật toán K-means Clustering

Sau khi đã có danh mục đầu tư từ Chương 1, em mong muốn xây dựng một mô hình đầu tư trực tuyến theo ngày. Trong quá khứ, Majid Khedmati và Pejman Azin đã xây dựng một mô hình có độ hiệu quả tương đối cao trong bài báo "An online portfolio selection algorithm using clustering approaches and considering transaction costs" [3]. Tuy nhiên, mô hình có độ phức tạp lớn khó tiếp cận với đa số người đọc. Ở chương này, em xây dựng mô hình đầu tư trực tuyến theo ngày có tham khảo mô hình của Majid Khedmati và Pejman Azin với độ phức tạp thấp hơn nhưng vẫn mong muốn đạt hiệu quả cao.

2.1. Xây dựng mô hình đầu tư trực tuyến

2.1.1. Một số giả thiết

- **Tính thanh khoản của thị trường:** Thị trường Crypto là một thị trường lớn, đồng thời ở Chương 1 em đã chọn 32 Crypto có vốn hóa lớn nhất, dẫn đến giả thiết tính thanh khoản dễ dàng được thỏa mãn.
- **Ảnh hưởng tới thị trường:** Giả thiết giao dịch của chúng ta không ảnh hưởng đến giá của thị trường. Tương tự như giả thiết trên, đây là một thị trường lớn và giao dịch những Crypto có vốn hóa cao nên giả thiết cũng dễ dàng thỏa mãn. Phần đông nhà đầu tư là nhà đầu tư nhỏ lẻ, khó có thể ảnh hưởng đến đường giá như các nhà đầu tư lớn (cá mập).

- **Giá Crypto:** Giá của Crypto thay đổi liên tục theo thời gian, hiện tại trên sàn Binance đã có cả đồ thị giá theo đơn vị giây, trong khi thuật toán của chúng ta lấy giá theo ngày. Vậy em lấy giá của Crypto là giá đóng cửa mỗi ngày trên sàn Binance.

- **Chi phí giao dịch:** Giả sử phí giao dịch cho mỗi lần mua hoặc bán là 0,1% khối lượng giao dịch. Đây là mức phí cho nhà đầu tư tiêu chuẩn của sàn giao dịch Binance.

- **Giá của các Crypto có tương quan thuận:** Giả thuyết tương đối dễ hiểu. Thậm chí, đôi khi chúng ta có thể dự đoán giá của các Crypto có vốn hóa bé dựa vào các Crypto có vốn hóa lớn.

2.1.2. Ý tưởng thuật toán

Trong Chương 2 này, ta cần quan tâm đến một đại lượng đó là tỷ số giữa giá Crypto ngày thứ t với giá Crypto ngày ngay trước đó:

$$x_{t,i} = \frac{P_{t,i}}{P_{t-1,i}}; t = 1, \ldots, n; i = 1, \ldots, m \tag{2.1}$$

trong đó,

- $x_{t,i}$ là giá của Crypto thứ i tại ngày t;

- n là số ngày ta sử dụng đẻ chạy thuật toán;

- m là số Crypto trong danh mục đầu tư.

Dễ thấy $x_{t,i} = R_{t,i} - 1$, nên nếu $R_{t,i}$ đã là chuỗi dừng thì $x_{t,i}$ cũng là chuỗi dừng.

Em sẽ đánh giá x để thực hiện thuật toán đầu tư theo ngày. Ý tưởng của em là sử dụng nguyên tắc **Đối sánh mẫu** (*parttern matching*), nghĩa là dữ liệu lịch sử được sử dụng để đưa ra quyết định về danh mục đầu tư của giai đoạn hiện tại. Trên thực tế, nó đang tìm kiếm các mẫu lịch sử tương tự như mẫu hiện tại.

Ví dụ 2.1. Ta mong muốn dự đoán giá của Crypto i. Hiện tại là ngày 29/2. Có $x_{28,i} = 0.2; x_{27,i} = 0.3$. Tìm kiếm dữ liệu trong quá khứ em thấy
$x_{4,i} = 0.1; x_{3,i} = 0.2; x_{2,i} = 0.3$. Vậy em kỳ vọng $x_{29,i} = 0.1$

Chú ý 2.1. Đây là một ví dụ rất đơn giản và chỉ cho 1 Crypto. Theo giả thuyết đã nêu ở Mục 2.1, các Crypto có tương quan thuận. Vì vậy em không xét riêng $x_{i,t}$ của từng Crypto mà em cần quan tâm đến x của tất cả m Crypto. Nghĩa là chuỗi x của cả m Crypto phải tương đối giống nhau ở quá khứ và hiện tại thì em mới coi nó là một cơ hội đầu tư. Qua đó, em quan tâm đến đại lượng véctơ giá tương đối $x_t = (x_{1,t}; x_{2,t}; \ldots; x_{m,t})$.

2.1.3. Cơ sở lý thuyết

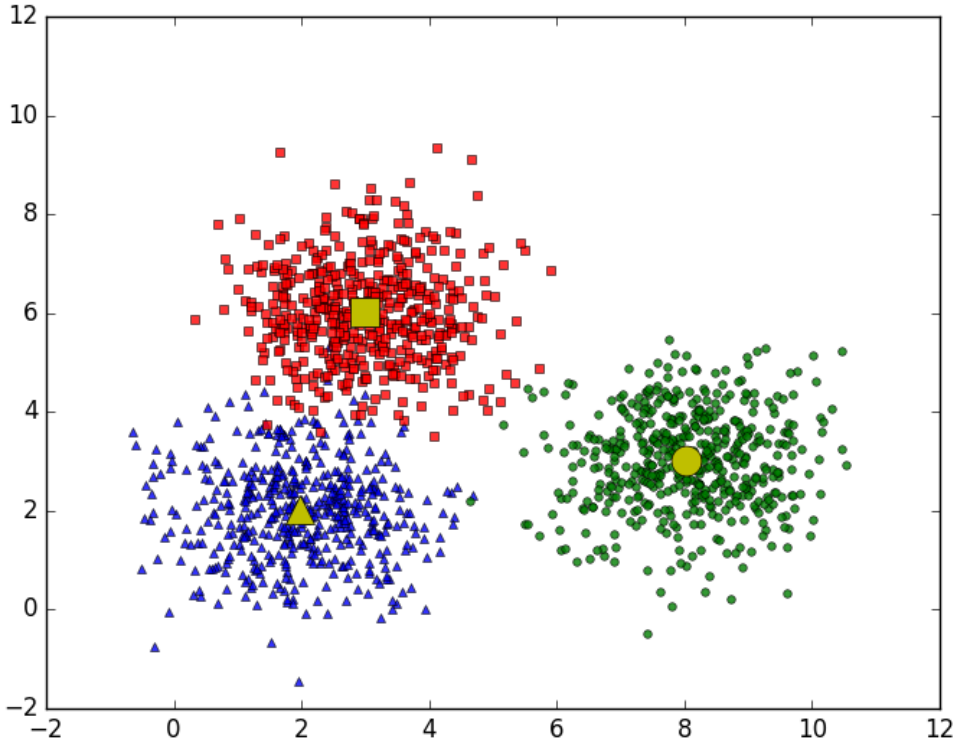
Phân cụm K-means

Thuật toán phân cụm K-means là một thuật toán Machine Learning, mục đích là làm thể nào để phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau nhất.

Ví dụ 2.2. Một công ty muốn tạo ra những chính sách ưu đãi cho những nhóm khách hàng khác nhau dựa trên sự tương tác giữa mỗi khách hàng với công ty đó (số năm là khách hàng; số tiền khách hàng đã chi trả cho công ty; độ tuổi; giới tính; thành phố; nghề nghiệp; ...). Giả sử công ty đó có rất nhiều dữ liệu của rất nhiều khách hàng nhưng chưa có cách nào chia toàn bộ khách hàng đó thành một số nhóm/cụm khác nhau. Nếu một người biết Machine Learning được đặt câu hỏi này, phương pháp đầu tiên họ nghĩ đến sẽ là K-means Clustering.

Ý tưởng đơn giản nhất về cụm (cluster) là tập hợp các điểm ở gần nhau trong một không gian nào đó (không gian này có thể có rất nhiều chiều trong trường hợp thông tin về một điểm dữ liệu là rất lớn).

Ví dụ 2.3. Hình bên dưới là một ví dụ về 3 cluster của các điểm dữ liệu 2 chiều.



Hình 2.1: 3 cluster 2 chiều [9]

Giả sử mỗi cluster có một điểm đại diện (center) màu vàng. Một cách đơn giản nhất, xét một điểm bất kỳ, em xét xem điểm đó gần với center nào nhất thì nó thuộc về cùng nhóm với center đó.

Thuật toán phân cụm K-means:

Đầu vào: X điểm dữ liệu và K là số lượng cluster cần tìm.

Đầu ra: Các center M và các điểm dữ liệu thuộc các cluster của các center đó.

- **Bước 1:** Chọn K điểm bất kỳ làm các center ban đầu. Chú ý, chúng ta có thể tùy chọn K (số lượng cụm) theo nhu cầu phân cụm.
- **Bước 2:** Phân mỗi điểm dữ liệu vào cluster có center gần nó nhất.
- **Bước 3:** Nếu việc gán dữ liệu vào từng cluster ở Bước 2 không thay đổi so với vòng lặp trước nó thì ta dừng thuật toán.
- **Bước 4:** Cập nhật center cho từng cluster bằng cách lấy trung bình cộng của tất các các điểm dữ liệu đã được gán vào cluster đó sau Bước 2.
- **Bước 5:** Quay lại Bước 2.

Chú ý 2.2. Trong phạm vi đồ án tốt nghiệp, em chỉ trình bày thuật toán mà không phân tích chuyên sâu về cơ sở toán học cũng như các nhược điểm của thuật toán phân cụm K-means.

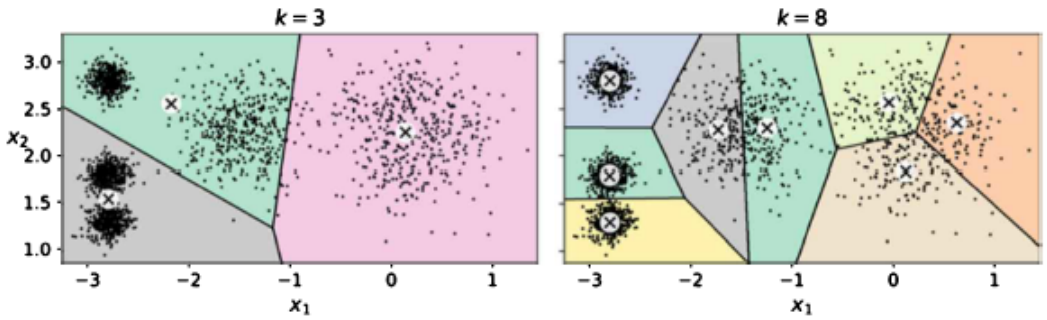
Lựa chọn K sử dụng chỉ số Silhouette

Vấn đề của K-means là chúng ta phải xác định trước giá trị tham số K, ở một số trường hợp này dataset của chúng ta có thể visualize được và phân bố cũng khá rõ ràng thì chúng ta có thể dễ dàng xác định được K. Vậy trong trường hợp không thể rõ ràng xác định trước giá trị K thì chúng ta phải làm như thế nào?

Ví dụ 2.4. Hình 2.2 cho ta thấy một ví dụ minh họa việc lựa chọn sai giá trị K có thể ảnh hưởng tới hiệu quả phân cụm của thuật toán. Rõ ràng ta thấy số cụm bằng 5 sẽ là lý tưởng, $K = 3$ (quá nhỏ) hoặc $K = 8$ (quá lớn) đều dẫn tới kết quả phân cụm không được tốt.

Một trong những cách thông dụng nhất để lựa chọn K là sử dụng chỉ số Silhouette của mỗi cách chọn K: $S(K)$. Gọi số điểm dữ liệu là X.

$$S(K) = \frac{\sum_{n=1}^{X} S(i, K)}{X} \tag{2.2}$$



Hình 2.2: Lựa chọn K [9]

$$S(i, K) = (b - a)/\max(a, b)$$

trong đó:

- a là khoảng cách trung bình từ điểm dữ liệu i tới các điểm dữ liệu khác ở trong cùng một cụm.
- b là khoảng cách trung bình từ điểm dữ liệu i tới các điểm dữ liệu trong cụm gần nhất.

Dễ thấy:

- $S(i, K)$ nằm trong khoảng từ -1 đến 1.
- $S(i, K)$ nằm gần 1 nghĩa là điểm dữ liệu đang được phân cụm chính xác, xa các cụm khác.
- $S(i, K)$ gần 0 nghĩa là điểm dữ liệu đang nằm gần đường bao của cụm.
- $S(i, K)$ gần -1 nghĩa là điểm dữ liệu đang bị phân sai cụm.

Dễ dàng tính được $S(K)$ theo công thức (2.2). $S(K)$ càng lớn, càng gần 1 thì chứng tỏ phân cụm càng hiệu quả. Như vậy, ta chọn K cho ta $S(K)$ lớn nhất.

2.1.4. Thuật toán đầu tư trực tuyến

Các tham số trong thuật toán

Các tham số và biến số của lựa chọn danh mục đầu tư trực tuyến được trình bày dưới đây:

- $\boldsymbol{m \geqslant 2}$: Số lượng Crypto trong danh mục đầu tư dã được hình thành từ Chương 1.
- $\boldsymbol{n \geqslant 1}$: Thời gian ta lấy dữ liệu để chạy thuật toán. Theo đề xuất từ Chương 1, $\boldsymbol{n = 485}$.
- $\boldsymbol{p_t}$: Ở ngày thứ t (t = 1, 2, ..., n), giá của các Crypto được biểu diễn bằng một véctơ $p_t = (p_{t,1}; p_{t,2}; \ldots; p_{t,m}) \in \mathbb{R}_+^m$ với $p_{t,i}$ là giá của Crypto i ngày thứ t.
- $\boldsymbol{x_t}$: Như đã được đề cập ở cơ sở lý thuyết phần Chú ý 2.1, x_t là sự thay đổi về giá của các Crypto được biểu diễn bằng véctơ giá tương đối $x_t = (x_{t,1}; x_{t,2}; \ldots; x_{t,m}) \in \mathbb{R}_+^m$ với $x_{t,i}$ là tỷ số giữa giá Crypto ngày thứ t với giá ngày ngay trước đó. $x_{t,i} = \frac{P_{t,i}}{P_{t-1,i}}$ (xem Mục 2.2).
- $\boldsymbol{x^n}$: $x^n = (x_1; x_2; \ldots; x_n)$ là ma trận giá tương đối của m Crypto trong thời gian lấy dữ liệu xét (n ngày). Ma trận có kích thước $n \times m$.
- $\boldsymbol{W}$: Ma trận trọng số của các ma trận con.
- $\boldsymbol{b}$: là véctơ biểu thị cho tỷ lệ đầu tư m Crypto. $b = (b_1; b_2; \ldots; b_m) \in \mathbb{R}_+^m$ với b_i là tỷ lệ vốn được chia cho Crypto i. $1 \geqslant b_i \geqslant 0$. Đây chính là đầu ra chúng ta mong muốn: phân bổ vốn đầu tư theo ngày.

  Có $\sum_{i=1}^{m} b_i = 1$.
- $\boldsymbol{TW}$: là viết tắt của Time Window. Ý nghĩa của nó là mô hình của chúng ta sẽ sử dụng chuỗi bao nhiêu ngày để dự đoán giá ở thời điểm hiện tại. Trong Ví dụ 2.1, $TW = 3$. Có thể lựa chọn TW trong khoảng 2-10 ngày.
- $\boldsymbol{I}$: là viết tắt của Iteration. Ý nghĩa của nó số lần ta chạy thuật toán phân cụm K-means. Lý do là vì thuật toán K-means có những nhược điểm của nó. Nghiệm cuối cùng phụ thuộc vào các center được khởi tạo ban đầu. Thuật toán có thể có tốc độ hội tụ rất chậm hoặc thậm chí cho

chúng ta nghiệm không chính xác (chỉ là local minimum - điểm cực tiểu mà không phải giá trị nhỏ nhất). Vậy nên chạy càng nhiều lần sau đó phân tích kết quả thì ta có kết quả cuối cùng càng chính xác. Người dùng có thể tùy chọn I theo nhu cầu và nguồn lực.

Các bước thực hiện

Đầu vào: Ma trận giá của các Crypto (Ma trận A ở Chương 1).
Đầu ra: Chiến lược phân bổ vốn đầu tư $\boldsymbol{b}$.

- **Bước 1:** Lấy ma trận $\boldsymbol{A}$ đã được làm sạch từ Bước 2 Mục 1.3 Chương 1.
- **Bước 2:** Từ ma trận $\boldsymbol{A}$, tính toán ma trận $\boldsymbol{x^n}$ theo công thức (2.1).
- **Bước 3:** Sử dụng thuật toán phân cụm K-means, dự đoán giá tương lai của Crypto.

  Bước 3 tương đối phức tạp nên em sẽ chia thành nhiều bước nhỏ hơn.

  + B1: Chia ma trận x^n lần lượt thành $n - TW + 1$ ma trận con kích thước $TW \times m$. Các ma trận con này sử dụng để chạy thuật toán phân cụm K-means.

  + B2: Đưa các ma trận con về dạng các véctơ $\in R_+^{m \times TW}$ vì thuật toán phân cụm K-means tính toán khoảng cách giữa các véctơ. Các bước bên dưới em vẫn sử dụng thuật ngữ ma trận con nhưng thực chất đó là các véctơ.

  + B3: Quy ước ma trận con cuối cùng, bao gồm véctơ x_t với t là ngày ở thời điểm hiện tại, gọi là ma trận hiện tại. Ở bước này ta chạy thuật toán phân cụm K-means với các K khác nhau, mỗi K chạy I lần để tìm các ma trận con cùng cụm với ma trận hiện tại. Khoảng giá trị của K: $\sqrt{n - TW + 1} \geqslant K \geqslant 2$.

  + B4: Tính toán chỉ số Silhouette ứng với mỗi K. Do mỗi K ta chạy I lần nên ta lấy chỉ số Silhouette ứng với mỗi K là chỉ số Silhouette trung bình của I lần chạy.

$$S(K) = \frac{\sum_{i=1}^{I} S(i,K)}{I}$$

Chọn K có chỉ số Silhouette lớn nhất.

+ B5: Với K đã chọn ở **B4**, tìm n^* các ma trận con 80% số lần chạy I được phân vào cùng cụm với ma trận hiện tại.

+ B6: Tính giá trị tương quan giữa n^* các ma trận con được chọn ở B5 với ma trận hiện tại. Tức là tính độ tương tự cosin giữ các véctơ này:

$$\cos(\theta) = \frac{A.B}{||A||.||B||}$$

+ B7: Dựa vào các giá trị tương quan này, tính trọng số của mỗi ma trận con rồi đưa vào ma trận trọng số W $\in R_+^{n^*}$ theo công thức:

$$w_i = \frac{\cos(\theta_i)}{\sum_{i=1}^{n^*} \cos(\theta_i)}$$

$$w_i = \frac{\cos(\theta_i)}{\sum_{i=1}^{n^*} \cos(\theta_i)}$$

$\cos(\theta_1) = 0.8$

$\cos(\theta_2) = 0.9$

$w_1 = \frac{\cos(\theta_1)}{\cos(\theta_1)+\cos(\theta_2)} = \frac{0.8}{0.8+0.9} = 0.47$

$w_2 = \frac{\cos(\theta_2)}{\cos(\theta_1)+\cos(\theta_2)} = \frac{0.9}{0.8+0.9} = 0.53$

Ví dụ 2.5. Giả sử $n^* = 2$, tức chọn được 2 ma trận con tương tự với ma trận hiện tại. Hệ số tương quan của 2 ma trận này với ma trận hiện tại lần lượt là $\cos(\theta_1) = 0.8$ và $\cos(\theta_2) = 0.9$. Ta tính được $W = (0.47; 0.53)$.

+ B8: Lấy ma trận C bao gồm dãy các véctơ x_t với t là ngày ngay sau ngày kết thúc của n^* ma trận con. Ma trận có kích thước $n^* \times m$.

+ B9: Tính $P = C \times W$ ta được véctơ giá tương đối kỳ vọng ngày hôm sau. Như vậy ta áp dụng đối sánh mẫu dựa trên nhiều ma trận con của quá khứ để dự đoán giá tương đối trong tương lai.

Chú ý 2.3. Trong thuật toán vừa rồi, em chưa xác định rõ TW bằng bao nhiêu. Tại mỗi thời điểm, TW tối ưu có thể khác nhau. Em đề xuất người dùng có thể sử dụng dữ liệu quá khứ để test thử với các TW khác nhau, tìm ra TW tối ưu nhất trong tương lai gần.

- **Bước 4:** Phân bổ vốn cho m Crypto.

  Trọng số phân bổ vốn của Crypto phụ thuộc vào giá tương đối ngày hôm sau P.

  + **B1:** Crypto có giá tương đối dự đoán nhỏ hơn 1 sẽ bị loại. Còn lại m^* Crypto sẽ được phân bố vốn.

  + **B2:** Crypto có giá tương đối dự đoán dương sẽ có trọng số theo công thức:

$$w_i = \frac{P_i - 1}{\sum_{i=1}^{m^*} P_i - 1}$$

2.2. Xây dựng chương trình

2.2.1. Lựa chọn các Crypto phù hợp

Kiểm định tính dừng

	Close day	BTC	ETH	BNB	XRP	ADA	DOGE	SOL	TRX	DOT	...	ETC	HBAR	FIL	LDO	ICP	VET	MKR	QNT	OP	NI
0	1656633599999	19942.21	1071.01	219.6	0.3321	0.4599	0.06644	33.76	0.06481	7.06	...	15.01	0.0635	5.40	0.453	5.35	0.02277	909.0	53.5	0.547	3.
1	1656719999999	19279.80	1059.73	216.9	0.3139	0.4487	0.06644	32.84	0.06512	6.75	...	14.64	0.0617	5.35	0.453	5.24	0.02226	892.0	52.5	0.530	3.
2	1656806399999	19252.81	1067.01	218.3	0.3156	0.4559	0.06667	33.36	0.06475	6.83	...	14.77	0.0633	5.34	0.467	5.24	0.02235	908.0	55.8	0.525	3.
3	1656892799999	19315.83	1074.26	219.2	0.3215	0.4561	0.06721	33.39	0.06628	6.85	...	14.81	0.0624	5.30	0.506	5.24	0.02257	904.0	55.7	0.522	3.
4	1656979199999	20236.71	1151.00	231.5	0.3287	0.4695	0.06938	36.71	0.06733	7.17	...	15.40	0.0636	5.55	0.540	5.59	0.02332	941.0	60.7	0.560	3.

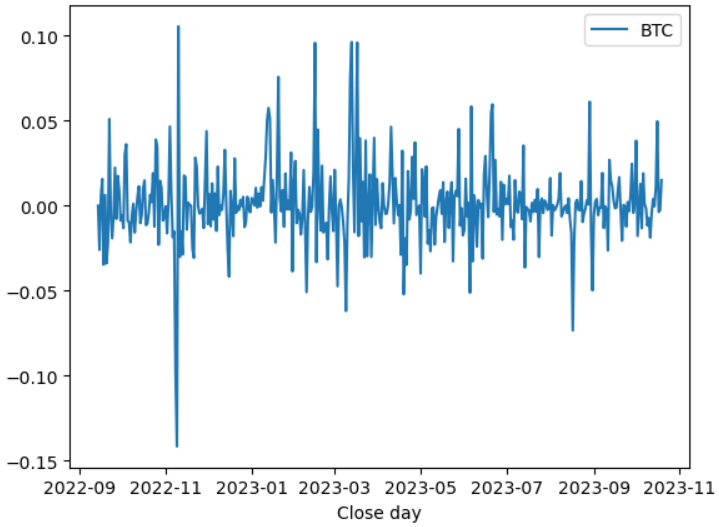
5 rows × 30 columns

Hình 2.3: Dữ liệu sau khi lấy về từ mô hình và loại các Crypto tồn tại ít hơn 3 năm

```
DOGE : 2.6236894956041864e-07
SOL : 0.0
TRX : 3.256870622739122e-25
DOT : 0.0
MATIC : 4.0332363888230025e-22
LTC : 0.0
SHIB : 8.990614023546178e-21
BCH : 2.2938835974808245e-06
XLM : 5.947314081280 51e-09
AVAX : 0.0
LINK : 1.4081479304746628e-08
XMR : 1.3623363691789235e-10
UNI : 5.428750036496 07e-29
ATOM : 1.9705859339834772e-29
ETC : 0.0
HBAR : 0.0
FIL : 0.0
LDO : 0.0
ICP : 0.0
VET : 0.0
MKR : 3.678315421450109e-29
QNT : 2.5204217505377383e-10
OP : 2.948328389480034e-28
NEAR : 1.7001117638375514e-05

Vậy các đồng Coin đều thoả mãn tính dừng
```

Hình 2.4: Kiểm định tính dừng



Hình 2.5: Đồ thị giá tương đối của BTC

Kiểm định nặng đuôi

Sau khi kiểm thử nặng đuôi trái, em nhận được kết quả như sau:

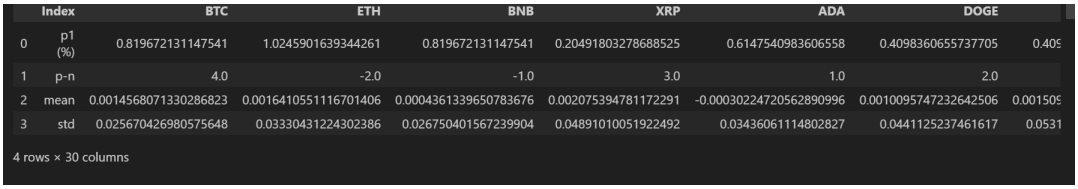
Python

Index	BTC	ETH	BNB	XRP	ADA	DOGE	
p1 (%)	0.8849557522123894	0.7964601769911505	0.6194690265486725	0.35398230088495575	0.4424778761061947	0.0	0.44247787

Hình 2.6: Tỷ lệ nặng đuôi trái

Do tính chất biến động lớn của Crypto, không Crypto nào thỏa mãn $(\boldsymbol{P_R|R \in (-\infty; E(R) - 3\sigma_R)) < p_0 = 0,15\%}$ (theo Mục 1.2.5). Có thể tăng $\boldsymbol{p_0}$ để chọn các Crypton như đã đề xuất, song em nhận thấy tính phân hóa nặng đuôi trái giữa các Crypto không cao. Vì vậy em quyết định đánh giá thêm một số chỉ số khác:

- $\boldsymbol{p_1}$: Tỷ lệ nặng đuôi trái;
- $\boldsymbol{p - n}$: Positive - Negative = Hiệu của tổng số nặng đuôi trái và phải;
- $\boldsymbol{mean}$: Giá tương đối trung bình;
- $\boldsymbol{std}$: Độ lệch chuẩn giá tương đối.



	Index	BTC	ETH	BNB	XRP	ADA	DOGE	
0	p1 (%)	0.819672131147541	1.0245901639344261	0.819672131147541	0.20491803278688525	0.6147540983606558	0.4098360655737705	0.40
1	p-n	4.0	-2.0	-1.0	3.0	1.0	2.0	
2	mean	0.0014568071330286823	0.0016410551116701406	0.0004361339650783676	0.002075394781172291	-0.0003022472056289099 6	0.00100957472326425 06	0.00150
3	std	0.025670426980575648	0.03330431224302386	0.026750401567239904	0.04891010051922492	0.03436061114802827	0.0441125237461617	0.0531

4 rows × 30 columns

Hình 2.7: Kết quả các chỉ số

Sau khi có kết quả, em sử dụng khoảng tứ phân vị để gán điểm cho mỗi chỉ số. Điểm nằm trong khoảng $[\mathbf{1, 4}]$. Các chỉ số $\boldsymbol{p - n}$ và $\boldsymbol{mean}$ càng cao thì điểm càng cao, ngược lại $\boldsymbol{p_1}$ và $\boldsymbol{std}$ càng cao thì điểm càng thấp.

Sau khi có điểm mỗi chỉ số, em tính tổng điểm

$\boldsymbol{Totalmark} = \sum_{i=1}^{4} \boldsymbol{mark_i} \times \boldsymbol{w_i}$ với trọng số như sau:

- $\boldsymbol{p_1 : w_1 = 0.4}$
- $\boldsymbol{p - n : w_2 = 0.2}$
- $\boldsymbol{mean : w_3 = 0.2}$
- $\boldsymbol{std : w_4 = 0.2}$

Khi có tổng điểm, em sử dụng khoảng tứ phân vị để đánh giá tổng điểm, loại bỏ khoảng tứ phân vị đầu tiên, tức là 25% Crypto có tổng điểm thấp nhất.

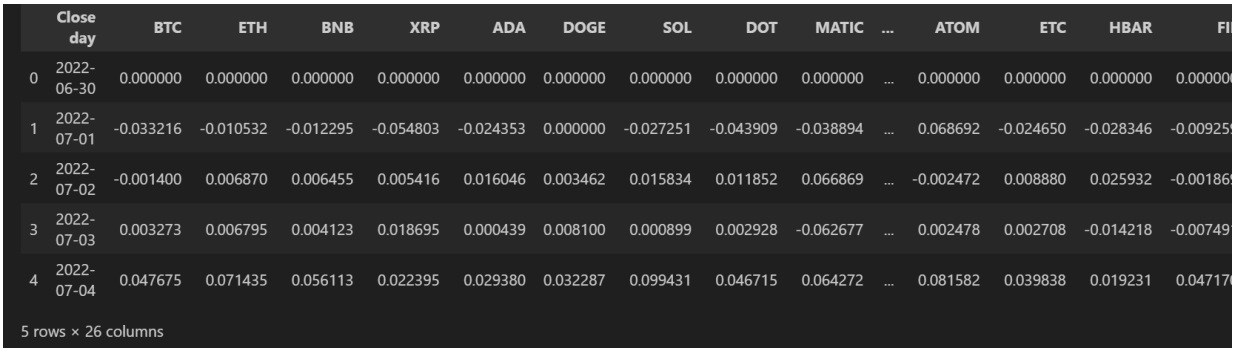
	Index	BTC	ETH	BNB	XRP	ADA	DOGE	SOL	TRX	DOT	...	ETC	HBAR	FIL	LDO	ICP	VET	MKR	QNT	OP	NEAR
0	p1 (%)	4	1	4	4	4	4	4	1	4	...	4	4	4	4	4	4	4	4	4	4
1	p-n	4	1	1	4	4	4	4	1	4	...	4	4	4	4	1	4	4	4	4	4
2	mean	3	3	2	4	1	2	3	4	1	...	3	4	2	4	1	1	4	4	4	1
3	std	4	4	4	1	4	2	1	4	4	...	1	3	1	1	4	4	2	3	1	4
4	Total mark	3	2	3	3	3	3	3	2	3	...	3	3	3	3	2	3	3	3	3	3

5 rows × 30 columns

Hình 2.8: Tổng điểm

```
Loại đồng coin: LTC
Loại đồng coin: LINK
Loại đồng coin: ATOM
```

Hình 2.9: Các Crypto bị loại

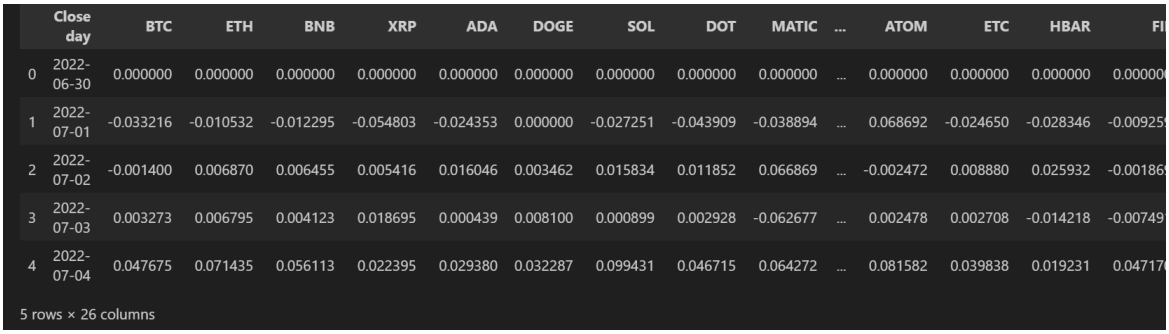


	Close day	BTC	ETH	BNB	XRP	ADA	DOGE	SOL	DOT	MATIC	...	ATOM	ETC	HBAR	FI
0	2022-06-30	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.00000
1	2022-07-01	-0.033216	-0.010532	-0.012295	-0.054803	-0.024353	0.000000	-0.027251	-0.043909	-0.038894	...	0.068692	-0.024650	-0.028346	-0.00925
2	2022-07-02	-0.001400	0.006870	0.006455	0.005416	0.016046	0.003462	0.015834	0.011852	0.066869	...	-0.002472	0.008880	0.025932	-0.00186
3	2022-07-03	0.003273	0.006795	0.004123	0.018695	0.000439	0.008100	0.000899	0.002928	-0.062677	...	0.002478	0.002708	-0.014218	-0.00749
4	2022-07-04	0.047675	0.071435	0.056113	0.022395	0.029380	0.032287	0.099431	0.046715	0.064272	...	0.081582	0.039838	0.019231	0.04717

5 rows × 26 columns

Hình 2.10: Dữ liệu xây dựng danh mục đầu tư

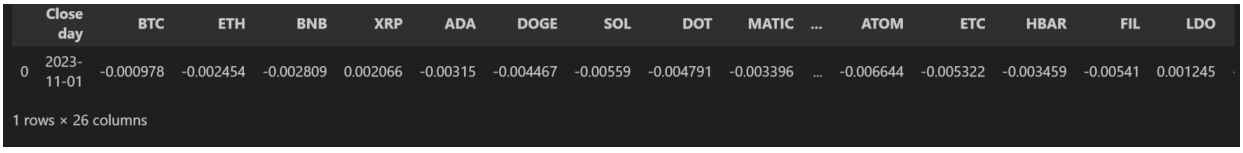
2.2.2. Phân bổ danh mục đầu tư



	Close day	BTC	ETH	BNB	XRP	ADA	DOGE	SOL	DOT	MATIC	...	ATOM	ETC	HBAR	FI
0	2022-06-30	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.00000
1	2022-07-01	-0.033216	-0.010532	-0.012295	-0.054803	-0.024353	0.000000	-0.027251	-0.043909	-0.038894	...	0.068692	-0.024650	-0.028346	-0.00925
2	2022-07-02	-0.001400	0.006870	0.006455	0.005416	0.016046	0.003462	0.015834	0.011852	0.066869	...	-0.002472	0.008880	0.025932	-0.00186
3	2022-07-03	0.003273	0.006795	0.004123	0.018695	0.000439	0.008100	0.000899	0.002928	-0.062677	...	0.002478	0.002708	-0.014218	-0.00749
4	2022-07-04	0.047675	0.071435	0.056113	0.022395	0.029380	0.032287	0.099431	0.046715	0.064272	...	0.081582	0.039838	0.019231	0.04717

5 rows × 26 columns

Hình 2.11: Ma trận giá tương đối với hàng đầu bằng 0



	Close day	BTC	ETH	BNB	XRP	ADA	DOGE	SOL	DOT	MATIC	...	ATOM	ETC	HBAR	FIL	LDO
0	2023-11-01	-0.000978	-0.002454	-0.002809	0.002066	-0.00315	-0.004467	-0.00559	-0.004791	-0.003396	...	-0.006644	-0.005322	-0.003459	-0.00541	0.001245

1 rows × 26 columns

Hình 2.12: Véctơ dự đoán biến đổi giá tương đối

	Close day	XRP	XLM	LDO	MKR
0	2023-11-01	0.002066	0.001713	0.001245	0.00099

Hình 2.13: Các Crypto được phân bổ vốn

	Close day	XRP	XLM	LDO	MKR
0	2023-11-01	0.34355	0.284806	0.20701	0.164634

Hình 2.14: Phân bổ vốn cho các Crypto

```
Dự đoán tỷ suất lợi nhuận của ngày hôm nay: 0.162 %
```

Hình 2.15: Dự đoán tỷ suất lợi nhuận

2.3. Kiểm định mô hình đầu tư

2.3.1. Lựa chọn dữ liệu kiểm định

Để kiểm tra độ hiệu quả của mô hình, em sẽ kiểm định trên 3 giai đoạn có xu hướng khác nhau của thị trường là: Bull(tăng), Bear(giảm) và Sideway(đi ngang). Xu hướng của thị trường Crypto phụ thuộc nhiều vào xu hướng của Bitcoin (BTC), vậy em sẽ quan sát xu hướng của BTC và coi đó là xu hướng chính của toàn bộ thị trường.

Hình 2.16: Thị trường Bull bắt đầu từ ngày 16/07/2021 [10]

Hình 2.17: Thị trường Bear bắt đầu từ ngày 29/05/2022 [10]

Hình 2.18: Thị trường Sideway bắt đầu từ ngày 16/03/2023 [10]

2.3.2. Kiểm định mô hình với các TW khác nhau

Như đã đề cập ở Chương 1, dữ liệu em sử dụng để huấn luyện mô hình sẽ gồm 365 ngày trước mỗi ngày giao dịch. Để kiểm định mô hình, em sẽ tính toán 2 chỉ số:

- CC: Cumulative capital (Tăng trưởng vốn tích lũy sau 120 ngày giao dịch)
- Accuracy = Tổng số lần dự đoán đúng/Tổng số lần dự đoán

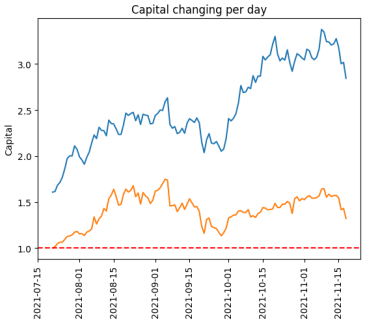
  Mô hình sử dụng phương thức đối sánh mẫu với mục đích dự đoán tăng trưởng giá tương đối của mỗi Crypto. Mô hình chỉ đầu tư vào các Crypto được dự đoán có tăng trưởng giá tương đối dương. Nếu ngày hôm sau giá tăng trưởng dương thì dự đoán được tính là đúng, ngược lại thì dự đoán là sai.

Trong các hình vẽ, đường màu cam thể hiện tăng trưởng vốn tích lũy sau mỗi ngày đầu tư, đường màu xanh thể hiện giá của BTC/20.

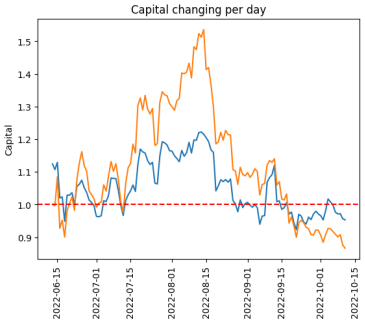
Tăng trưởng vốn tích lũy (CC) đã trừ đi chi phí giao dịch. Lấy phí giao dịch trung bình của sản Binance là 0,1% trên mỗi giao dịch. Như vậy sau mỗi giao dịch, tổng vốn sau giao dịch = tổng vốn sau giao dịch $\times$ **0, 999**. Mỗi

ngày giao dịch cần phân bổ lại danh mục đầu, có một số Crypto không thể trực tiếp trao đổi với các Crypto khác mà cần bán đi để lấy USDT sau đó lấy USDT mua các Crypto trong danh mục đầu tư mới. Vậy em lấy hệ số chi phí là 1,5. Nghĩa là tổng vốn sau 120 ngày giao dịch = tổng vốn tích lũy chưa tính chi phí giao dịch$\times \mathbf{0,999^{1.5\times 120}}$.

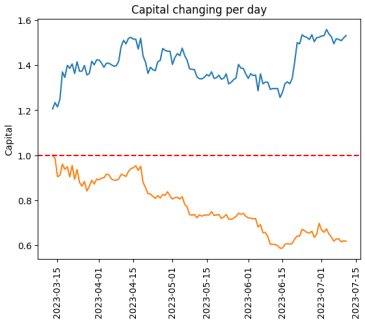
$TW = 3$



Hình 2.19: Bull market
CC = 32.04%
Accuracy = 55%



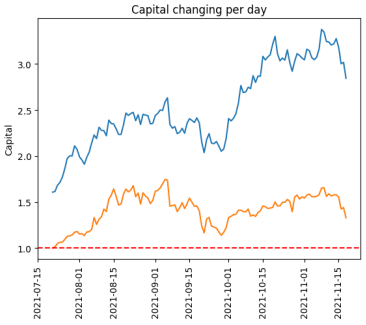
Hình 2.20: Bear market
CC = -13.37%
Accuracy = 46.32%



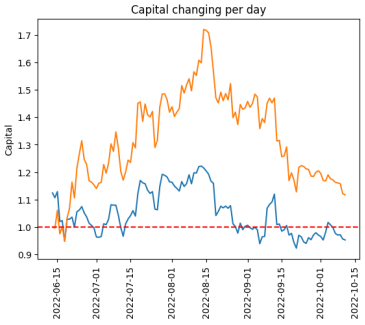
Hình 2.21: Sideway market
CC = -38.19%
Accuracy = 50.9%

- Average CC: -6.5%
- Average Accuracy: 50.74%

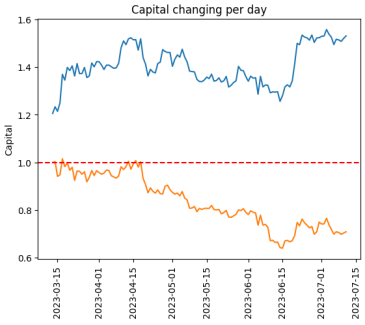
$TW = 4$



Hình 2.22: Bull market
CC = 32.79%
Accuracy = 54.7%



Hình 2.23: Bear market
CC = 11.73%
Accuracy = 47.55%

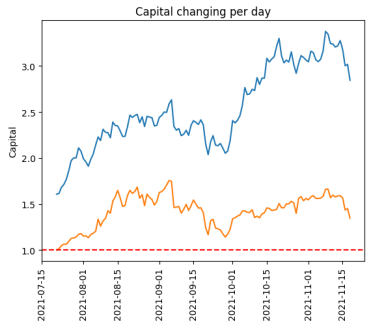


Hình 2.24: Sideway market
CC = -29.12%
Accuracy = 47.86%

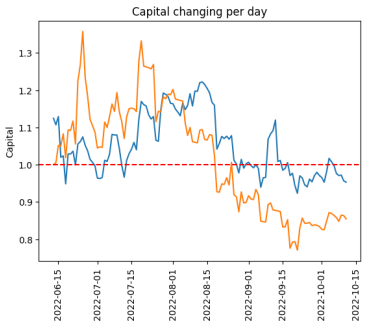
- Average CC: 5.13%

- Average Accuracy: 50.04%

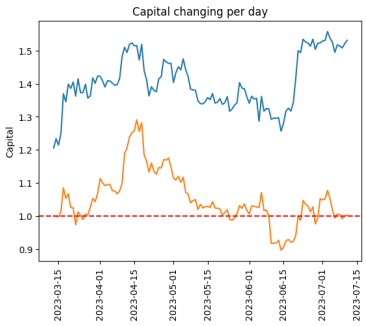
$TW = 5$



Hình 2.25: Bull market
CC = 34.59%
Accuracy = 55.08%



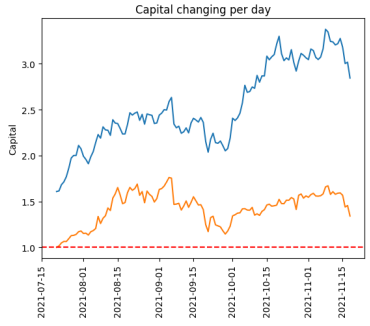
Hình 2.26: Bear market
CC = -14.56%
Accuracy = 47.52%



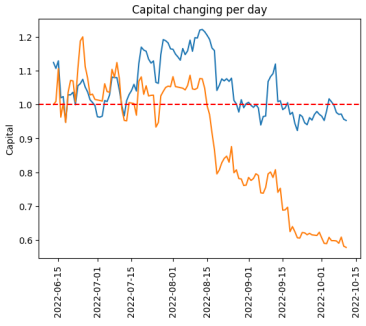
Hình 2.27: Sideway market
CC = -0.256%
Accuracy = 50.98%

- Average CC: 6.59%
- Average Accuracy: 51.19%

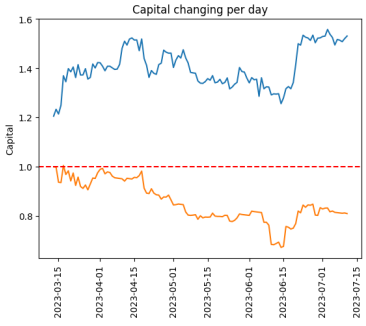
$TW = 6$



Hình 2.28: Bull market
CC = 34.04%
Accuracy = 55.32%



Hình 2.29: Bear market
CC = -42.25%
Accuracy = 46.12%



Hình 2.30: Sideway market
CC = -19.01%
Accuracy = 50.33%

- Average CC: -9.07%
- Average Accuracy: 50.59%

2.3.3. Nhận xét kết quả kiểm định mô hình

- Hiệu quả đạt được của mô hình chưa cao khi độ chính xác chỉ khoảng 50% (tỷ lệ ngang với đánh giá ngẫu nhiên) và lợi nhuận đạt được còn phụ thuộc nhiều vào xu hướng của thị trường. Có thể phương pháp đối sánh mẫu không phù hợp với thị trường Crypto cho việc đầu tư theo ngày. Nguyên nhân có thể là do đây là một thị trường dễ tham gia, còn phụ thuộc nhiều vào tâm lý giao dịch mỗi người nên giá Crypto phụ thuộc vào nhiều yếu tố, khó để áp dụng phân tích định lượng. Hành vi giá lặp lại yếu và phân cụm ít rõ ràng để có thể sử dụng cho việc xây dựng danh mục đầu tư. Với $\boldsymbol{TW}$ càng nhỏ thì hiệu suất phân cụm sẽ càng tốt, song với $\boldsymbol{TW_{min} = 3}$, ta đạt được $\boldsymbol{silhouette_{max} \approx 0,12}$, đây là một kết quả phân cụm yếu.

- Dựa vào kết quả kiểm định mô hình, có thể thấy sử dụng $\boldsymbol{TW = \{4, 5\}}$ cho ta hiệu quả mô hình cao nhất. TW trong 2 trường hợp này không quá lớn khiến cho phân cụm quá yếu, cũng không quá nhỏ khiến cho phương pháp đối sánh mẫu trở nên vô nghĩa.

- Dựa vào các đồ thị, có thể thấy một kết quả như sau: Tăng trưởng vốn tích lũy tương quan mạnh với giá của BTC khi thị trường có xu hướng tăng, song lại tương quan yếu khi thị trường có xu hướng giảm hoặc đi ngang. Điều này cho thấy tính chất dẫn dắt thị trường Crypto của BTC. Khi giá BTC tăng, giá của tất cả các Crypto khác sẽ tăng theo khiến đầu tư luôn có lời. Tuy nhiên khi giá BTC giảm hoặc đi ngang, thị trường biến động phức tạp hơn khiến tương quan yếu, song vẫn khiến mô hình đầu tư thua lỗ.

Có thể thấy khi giá BTC giảm, giá các đồng coin khác giảm mạnh hơn. Như vậy ta tìm hiểu về một chiến lược đầu tư bán khống phụ thuộc giá BTC ở Chương 3.

Chương 3

Mô hình đầu tư trực tuyến sử dụng mô hình LSTM

Để xây dựng một mô hình có thể dự đoán được giá BTC, em lựa chọn mô hình LSTM. Đây là một mô hình tương đối mạnh được sử dụng cho dữ liệu chuỗi thời gian ở thời điểm hiện tại. Sau khi tham khảo giáo trình "Deep Learning cơ bản" của tác giả Nguyễn Thanh Tuấn [12], và các website về LSTM, em áp dụng LSTM vào mô hình đầu tư của mình. Mô hình LSTM ở Chương 3 sử dụng thư viện con mạng nơ ron Keras trong thư viện Tensorflow do Google phát triển.

3.1. Xây dựng mô hình đầu tư trực tuyến

3.1.1. Ý tưởng đầu tư

Ý tưởng đầu tư lần này tập trung vào thị trường bán khống, bán trước mua sau, khi ta thấy giá các đồng alt-coin tụt mạnh hơn giá đồng Bitcoin. Vậy đầu tiên ta tìm các đồng coin nhạy cảm với sự giảm giá của bitcoin, sau đó ta tìm cách dự đoán giá bitcoin qua LSTM và tiến hành đầu tư.

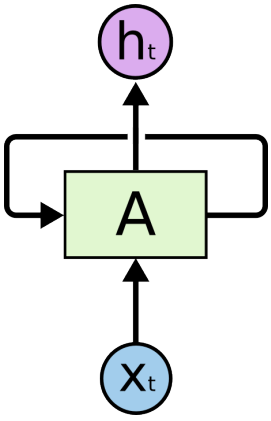
Tiến hành chọn $\boldsymbol{TW}$, tức là bán trước rồi sau bao nhiêu ngày ($\boldsymbol{TW}$) thì mua lại. Dựa vào biểu đồ giá của BTC có thể thấy các nhịp giảm mạnh rơi vào trong khoảng $\mathbf{3-7}$ ngày, vậy $\boldsymbol{TW} = [\mathbf{3, 7}]$.

Mỗi lần đầu tư, ta chỉ đầu tư với khối lượng bằng tổng số vốn $\boldsymbol{/TW}$, để đề phòng trường hợp có các cơ hội đầu tư khác ngay trong chính khoảng $\boldsymbol{TW}$ của lượt đầu tư phía trước.

3.1.2. Cơ sở lý thuyết

Mạng nơ ron truy hồi (RNN - Recurrent Neural Network)

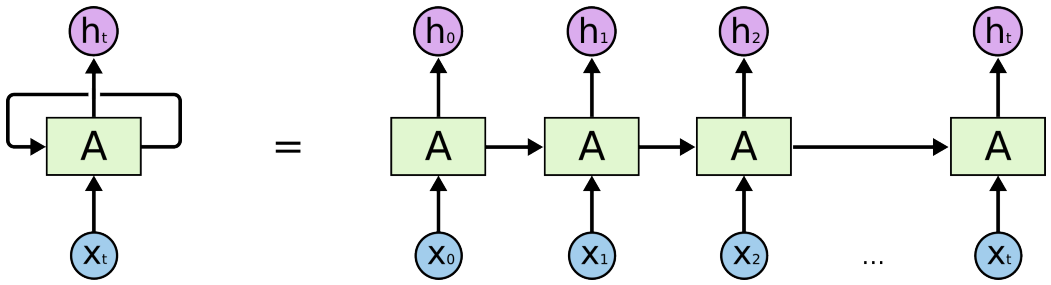
Trong lý thuyết về ngôn ngữ, ngữ nghĩa của một câu được tạo thành từ mối liên kết của những từ trong câu theo một cấu trúc ngữ pháp. Nếu xét từng từ một đứng riêng lẻ ta không thể hiểu được nội dụng của toàn bộ câu, nhưng dựa trên những từ xung quanh ta có thể hiểu được trọn vẹn một câu nói. Như vậy cần phải có một kiến trúc đặc biệt hơn cho các mạng nơ ron biểu diễn ngôn ngữ nhằm mục đích liên kết các từ liền trước với các từ ở hiện tại để tạo ra mối liên hệ xâu chuỗi. Mạng nơ ron truy hồi đã được thiết kế đặc biệt để giải quyết yêu cầu này.



Hình 3.1: Mạng nơ ron truy hồi với vòng lặp [11]

Hình trên biểu diễn kiến trúc của một mạng nơ ron truy hồi. Trong kiến trúc này mạng nơ ron sử dụng một đầu vào là một véctơ $\boldsymbol{x_t}$ và trả ra đầu ra là một giá trị ẩn $\boldsymbol{H_t}$. Đầu vào được đấu với một thân mạng nơ ron $\boldsymbol{A}$ có tính chất truy hồi và thân này được đấu tới đầu ra $\boldsymbol{h_t}$.

Vòng lặp $\boldsymbol{A}$ ở thân mạng nơ ron là điểm mấu chốt trong nguyên lý hoạt động của mạng nơ ron truy hồi. Đây là chuỗi sao chép nhiều lần của cùng một kiến trúc nhằm cho phép các thành phần có thể kết nối liền mạch với nhau theo mô hình chuỗi. Đầu ra của vòng lặp trước chính là đầu vào của vòng lặp sau. Nếu trải phẳng thân mạng nơ ron $\boldsymbol{A}$ ta sẽ thu được một mô hình dạng:



Hình 3.2: Cấu trúc trải phẳng của mạng nơ ron truy hồi [11]

Kiến trúc mạng nơ ron truy hồi này tỏ ra khá thành công trong các tác vụ của deep learning như: Nhận diện giọng nói (speech recognition), các mô hình ngôn ngữ, mô hình dịch, chú thích hình ảnh (image captioning),...

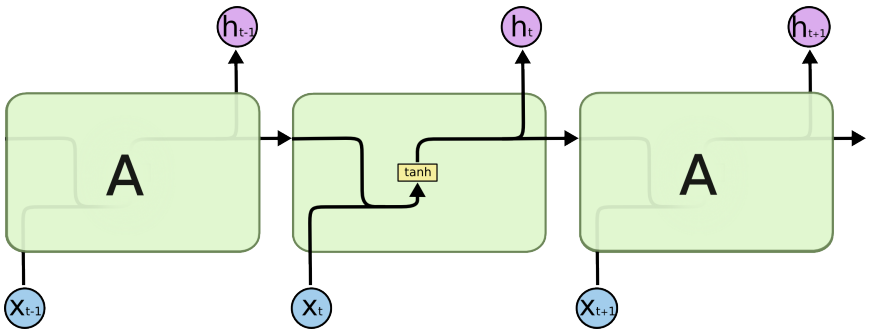
Hạn chế của mạng nơ ron truy hồi

Một trong những điểm đặc biệt của RNN đó là nó có khả năng kết nối các thông tin liền trước với nhiệm vụ hiện tại, chẳng hạn như trong câu văn: "Học sinh đang tới trường học". Dường như trong một ngữ cảnh ngắn hạn, từ trường học có thể được dự báo ngay tức thì mà không cần thêm các thông tin từ những câu văn khác gần đó. Tuy nhiên có những tình huống đòi hỏi phải có nhiều thông tin hơn chẳng hạn như: "Hôm qua Bảo đi học nhưng không mang áo mưa. Trên đường đi học trời mưa. Máy tính của Bảo bị ướt". Chúng ta cần phải học để tìm ra từ ướt ở một ngữ cảnh dài hơn so với chỉ một câu. Tức là cần phải biết các sự kiện trước đó như trời mưa, không mang áo mưa để suy ra sự kiện bị ướt. Những sự liên kết ngữ nghĩa dài như vậy được gọi là phụ thuộc dài hạn (long-term dependencies). Về mặt lý thuyết mạng RNN có thể giải quyết được những sự phụ thuộc trong dài hạn. Tuy nhiên trên thực tế RNN lại cho thấy khả năng học trong dài hạn kém hơn. Để hiểu thêm lý do tại sao mạng RNN lại không có khả năng học trong dài hạn cùng đọc bài Leanring Long - Term Dependencies with Gradient Descent is Difficult. Một trong những nguyên nhân chính được giải thích đó là sự triệt tiêu đạo hàm của hàm cost function sẽ diễn ra khi trải quả chuỗi dài các tính toán truy hồi. Một phiên bản mới của mạng RNN là mạng LSTM ra đời nhằm khắc phục hiện tường này nhờ một cơ chế đặc biệt.

Mạng trí nhớ ngắn hạn định hướng dài hạn (LSTM - Long short term memory)

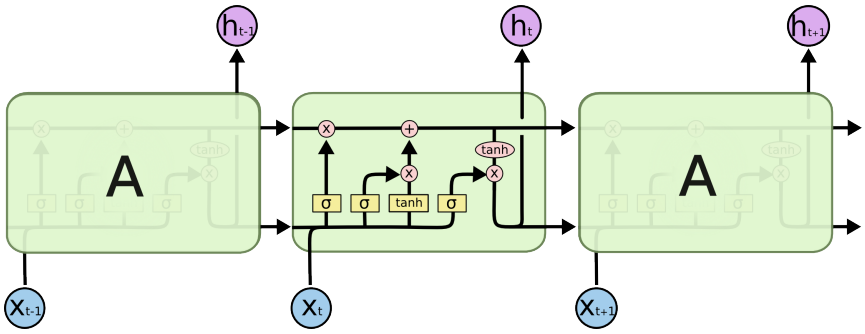
Mạng trí nhớ ngắn hạn định hướng dài hạn còn được viết tắt là LSTM làm một kiến trúc đặc biệt của RNN có khả năng học được sự phục thuộc trong dài hạn (long-term dependencies) được giới thiệu bởi Hochreiter & Schmidhuber (1997). Kiến trúc này đã được phổ biến và sử dụng rộng rãi cho tới ngày nay. LSTM đã tỏ ra khắc phục được rất nhiều những hạn chế của RNN trước đây về triệt tiêu đạo hàm. Tuy nhiên cấu trúc của chúng có phần phức tạp hơn mặc dù vẫn dữ được tư tưởng chính của RNN là sự sao chép các kiến trúc theo dạng chuỗi.

Một mạng RNN tiêu chuẩn sẽ có kiến trúc rất đơn giản chẳng hạn như đối với kiến trúc gồm một tầng ẩn là hàm tanh như hình:



Hình 3.3: Sự lặp lại kiến trúc module trong mạng RNN chứa một tầng ẩn [11]

LSTM cũng có một chuỗi dạng như thế nhưng phần kiến trúc lặp lại có cấu trúc khác biệt hơn. Thay vì chỉ có một tầng đơn, chúng có tới 4 tầng ẩn (3 sigmoid và 1 tanh) tương tác với nhau theo một cấu trúc đặc biệt. Xem hình:



Hình 3.4: Sự lặp lại kiến trúc module trong mạng LSTM chứa 4 tầng ẩn (3 sigmoid và 1 tanh) tương tác [11]

Trong sở đồ tính toán trên, mỗi một phép tính sẽ triển khai trên một véctơ. Trong đó hình tròn màu hồng biểu diễn một toán tử đối với véctơ như phép cộng véctơ, phép nhân vô hướng các véctơ. Màu vàng thể hiện hàm activation mà mạng nơ ron sử dụng để học trong tầng ẩn, thông thường là các hàm phi tuyến sigmoid và tanh. Ký hiệu 2 đường thẳng nhập vào thể hiện phép chập kết quả trong khi ký hiệu 2 đường thẳng rẽ nhánh thể hiện cho nội dung véctơ trước đó được sao chép để đi tới một phần khác của mạng nơ ron.

LSTM là một bước đột phá lớn mà ở đó chúng ta đã khắc phục được những hạn chế ở RNN đó là khả năng phụ thuộc dài hạn. Một số kỹ thuật học Attention gần đây được kết hợp với LSTM đã tạo ra những kết quả khá bất ngờ trong các tác vụ dịch máy cũng như phân loại nội dung, trích lọc thông tin,.... Các mô hình dịch máy của google đã ứng dụng kiểu kết hợp này trong các bài toán dịch thuật của mình và đã cải thiện được nội dung bản dịch một cách đáng kể.

3.1.3. Thuật toán đầu tư trực tuyến

- **Bước 1:** Phân bổ vốn cho các đồng coin khi thực hiện bán khống.

  Điều quan trọng nhất ở bước này là tìm ra các đồng coin giảm giá cùng với BTC, tức là tương quan cao với BTC. Vậy ta sẽ phân bổ khối lượng vào các đồng coin tỷ lệ thuận với độ tương quan về giá của chúng so với

BTC.

Công thức tính $\boldsymbol{w_i}$ trọng số phân bổ vốn vào các đồng coin như sau:

$$\boldsymbol{w_i} = \frac{\boldsymbol{corr_i}}{\sum_{i=1}^{m} \boldsymbol{corr_i}}$$

trong đó,

$\boldsymbol{w_i}$ là trọng số phân bổ vốn vào các đồng coin;
$\boldsymbol{corr_i}$ là tương quan về giá của các đồng coin với BTC;
$\boldsymbol{m}$ là số đồng coin sử dụng để đầu tư.

- **Bước 2:** Dự đoán giá của BTC trong $\boldsymbol{TW}$ ngày tiếp theo tính từ thời điểm bắt đầu đầu tư.

  Vì chiến lược đầu tư là bán khống, nên ta cần tìm ra những thời điểm mà BTC giảm giá trong $\boldsymbol{TW}$ ngày liên tiếp, tức là:

  $$\boldsymbol{P_n > P_{n+1} > P_{n+2} > \ldots > P_{n+TW-1}}$$

  với $\boldsymbol{P_n}$ là giá BTC được dự đoán ở ngày bắt đầu đầu tư.

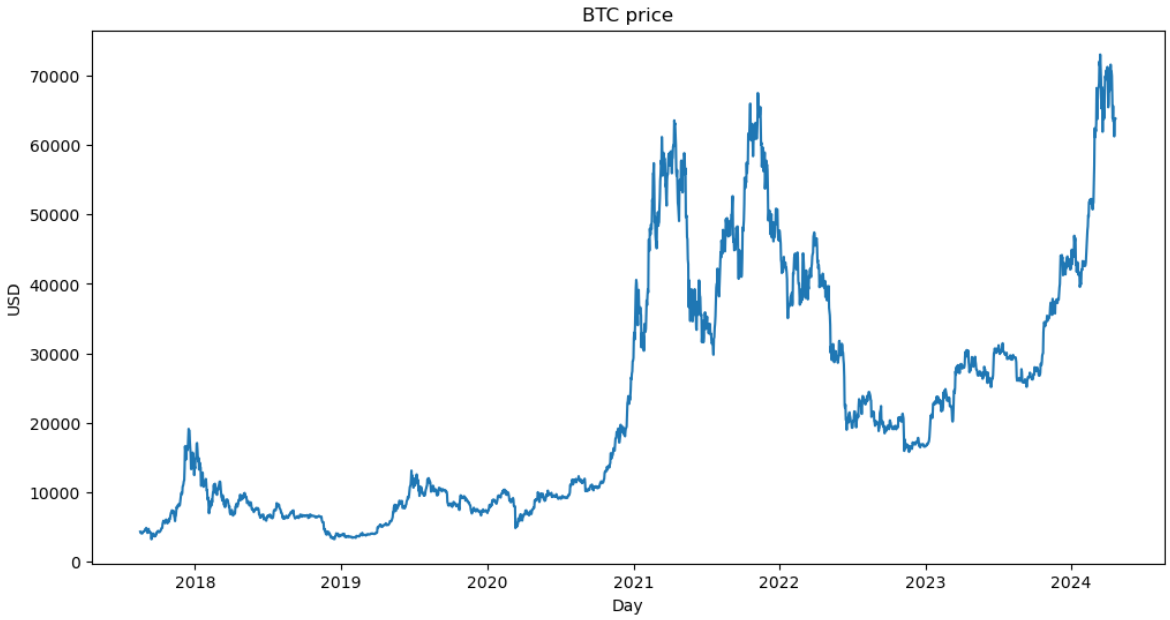
- **Bước 3:** Tính hành bán khống.
  Nếu giá BTC thỏa mãn điều kiện ở **Bước 2**, tiến hành bán khống với phân bổ vốn vào các đồng coin như đã đề cập ở **Bước 1**.
  Ngược lại, nếu không thỏa mãn các điều kiện ở **Bước 2**, quay lại **Bước 2** vào ngày tiếp theo.

Chú ý 3.1. Sau mỗi ngày ta sẽ có thêm dữ liệu thực tế về giá BTC ngày tiếp theo, vì vậy mỗi ngày mô hình sẽ được huấn luyện lại từ đầu. Chính vì lý do này mà thời gian huấn luyện mô hình sẽ vô cùng lớn nếu để epochs = **100**, vì vậy ở phần kiểm nghiệm tính hiệu quả của mô hình, ta sử dụng epochs = **10**. Độ chính xác sẽ giảm đáng kể song cũng giúp ta phần nào đánh giá được hiệu quả của mô hình.

3.2. Xây dựng chương trình

3.2.1. Xây dựng mô hình LSTM

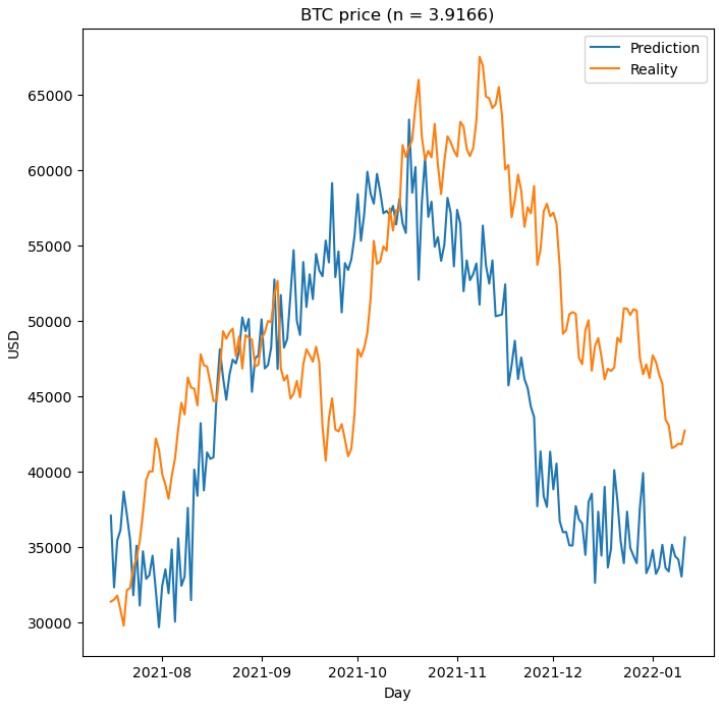
Sử dụng API Binance, có thể thu thập được dữ liệu đường giá của BTC từ ngày 17/08/2017 đến thời điểm hiện tại. Đường giá của BTC được thể hiện qua biểu đồ sau:



Hình 3.5: Đường giá của BTC qua các thời kỳ

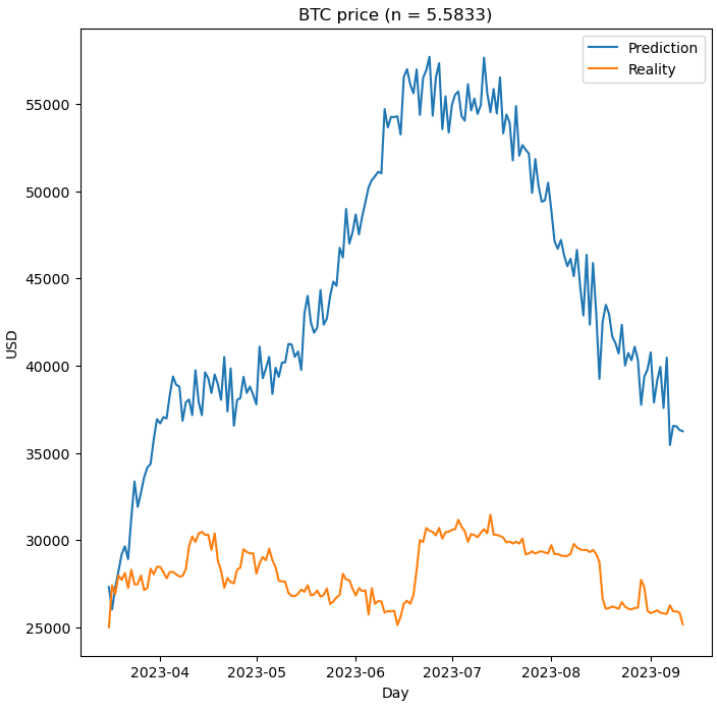
Sau khi có dữ liệu giá của BTC, sử dụng thư viện Keras của Python, có thể xây dựng mô hình LSTM dựa vào dữ liệu đã lấy về. Sử dụng mô hình mới dự đoán đường giá của BTC trong các thị trường đã đề cập ở Chương 4, có được kết quả dự đoán như các hình bên dưới.

Mô hình sử dụng tập huấn luyện là dữ liệu từ ngày đầu tiên thu thập dữ liệu (17/08/2017) đến ngày ngay trước bắt đầu bắt đầu dự đoán. Mô hình được huấn luyện với cài đặt mặc định. Thời gian mô hình dự đoán giá của BTC là 180 ngày (6 tháng).



Hình 3.6: Thị trường Bull bắt đầu từ ngày 16/07/2021

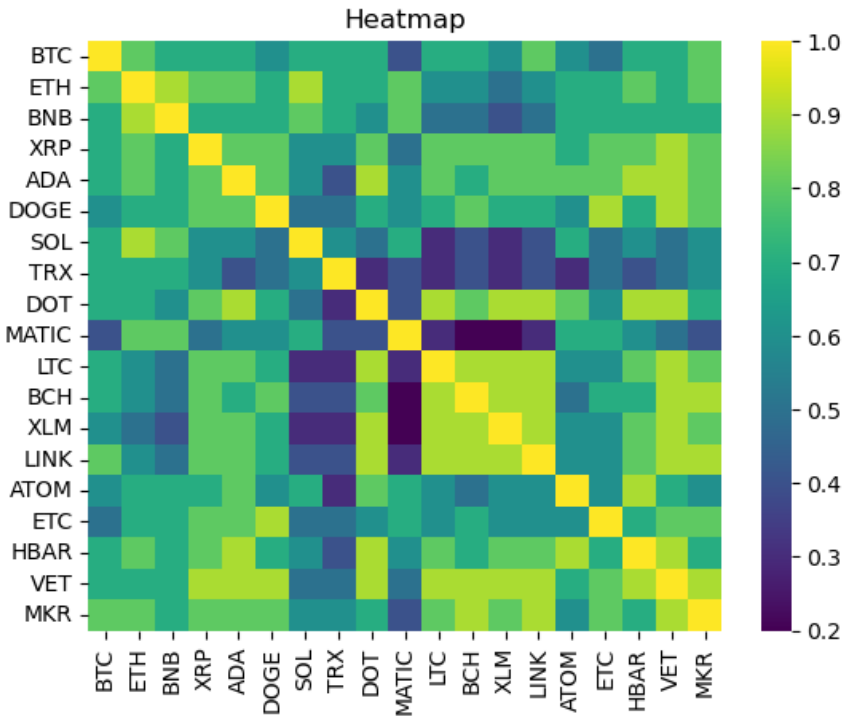
Hình 3.7: Thị trường Bear bắt đầu từ ngày 29/05/2022



Hình 3.8: Thị trường Sideway bắt đầu từ ngày 16/03/2023

Có thể thấy dự đoán không được tốt cho lắm ở các thị trường Bear và Sideway. Thị trường Bull có tập huấn luyện nhỏ nhất do diễn ra sớm nhất xong lại đạt được kết quả dự đoán tốt nhất với mô hình LSTM.

3.2.2. Phân bổ danh mục đầu tư



Hình 3.9: Tương quan giá giữa các Crypto

Phân bổ vốn đầu tư vào các đồng coin theo công thức Mục 3.1.3, ta đạt được như sau:

BTC: 7.66%

ETH: 6.4%

BNB: 5.55%

XRP: 5.2%

ADA: 5%

DOGE: 4.37%

SOL: 5.56%

TRX: 5.35%

DOT: 5.43%

MATIC: 3.1%

LTC: 5.11%

BCH: 5.51%

XLM: 4.89%

LINK: 5.85%

ATOM: 4.49%
ETC: 3.78%
HBAR: 5.45%
VET: 5.13%
MKR: 6.23%

Lựa chọn cài đặt phù hợp cho mô hình đối với việc dự đoán giá của BTC

Trong thư viện Keras, ta có thể sử dụng mô hình LSTM với nhiều cài đặt khác nhau:

- optimizer: "adam", "sgd", "RMSprop" - Các phương pháp huấn luyện mô hình đưa mô hình về các tham số đạt cực tiểu của hàm mất mát.
- epochs: **10, 50, 100, ...** - Số vòng lặp để huấn luyện mô hình, epochs càng lớn thì thời gian huấn luyện càng dài và mô hình có độ chính xác càng cao.
- batch_size: **32** hoặc **64** - Số lượng dữ liệu trong một mini-batch được sử dụng để tính đạo hàm trong quá trình huấn luyện mô hình (đọc thêm về kỹ thuật sử dụng mini-batch tại [12])
- train_year: **[1, 5]** - Số năm dữ liệu được dùng để huấn luyện mô hình. Thông thường càng nhiều dữ liệu sẽ cho ra dự đoán càng chính xác, song đối với dữ liệu chuỗi thời gian, đôi khi việc sử dụng dữ liệu quá cũ sẽ khiến mô hình giảm độ chính xác, vì vậy cần kiểm tra bộ dữ liệu huán luyện cho hiệu quả tốt nhất.
- n_lookback: **50, 100, 150, ...** - Số ngày trong quá khứ mô hình sử dụng để khớp các tham số, từ đó sử dụng các tham số để dự đoán giá của ngày tiếp theo.

Để tìm ra cài đặt hiệu quả nhất, ta sử dụng vòng lặp trong python để tìm ra tổ hợp cài đặt cho giá trị hàm mất mát đạt nhỏ nhất. Trong Keras, giá trị hàm mất mát đã được chia trung bình cho tổng số dữ liệu huấn luyện. Vì vậy, sử dụng train_year nhỏ hay lớn không ảnh hưởng đến việc đánh giá hiệu quả của các cài đặt.

Sau khi chạy vòng lặp, ta được các cài đặt tối ưu cho mô hình trong việc dự đoán giá của BTC như sau:

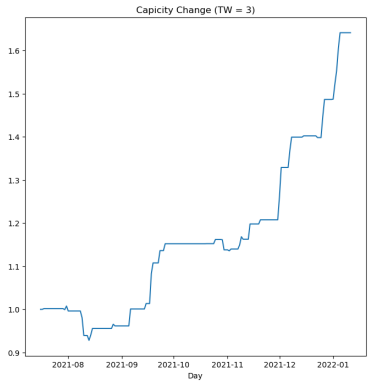
- optimizer: "adam"
- epochs = **100**
- batch_size = **32**
- train_year = **5**
- n_lookback = **150**

Có thể thấy mô hình hoạt động tôt nhất với số lượng dữ liệu huấn luyện đạt tốt đa, vậy ta sẽ sử dụng tối đã lượng dữ liệu có được để huấn luyện mô hình cho các mô hình đầu tư ở Mục 3.3.2.

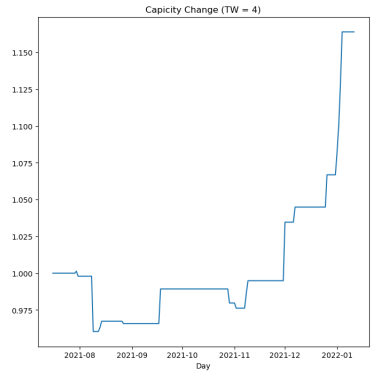
3.3. Kiểm định hiệu quả mô hình đầu tư

3.3.1. Kiểm định hiệu quả mô hình với các thị trường khác nhau sử dụng dữ liệu giá BTC thực tế

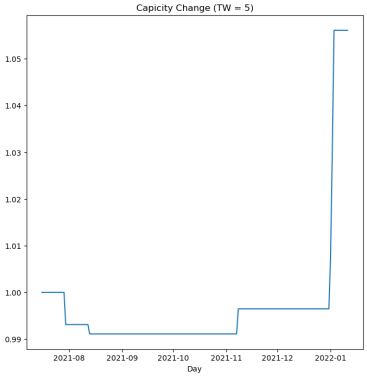
Bull market



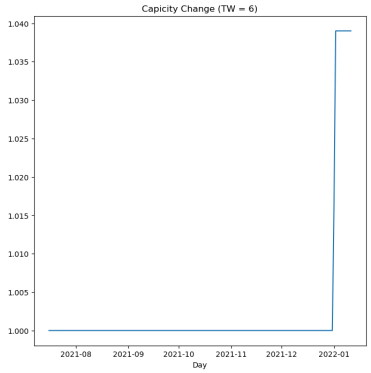
Hình 3.10: TW = 3
CC = 57.51%
Invest time = 41



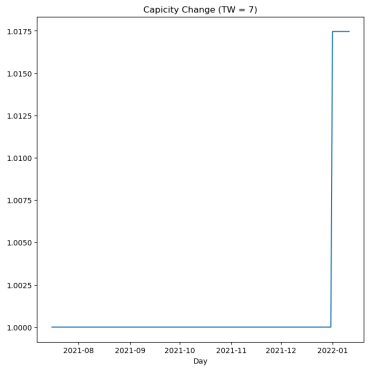
Hình 3.11: TW = 4
CC = 14.32%
Invest time = 18



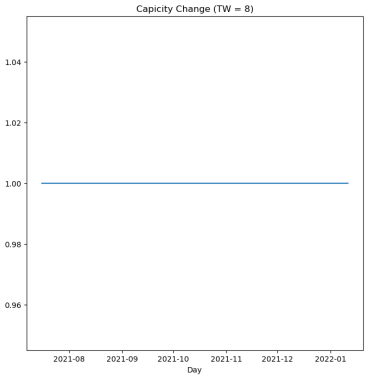
Hình 3.12: TW = 5
CC = 4.98%
Invest time = 6



Hình 3.13: TW = 6
CC = 3.69%
Invest time = 2

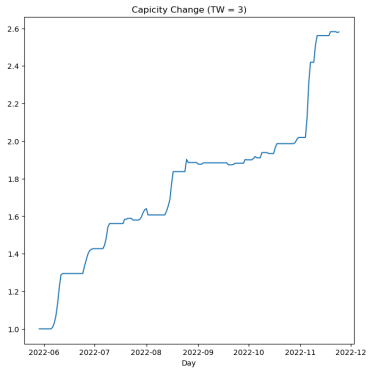


Hình 3.14: TW = 7
CC = 1.64%
Invest time = 1

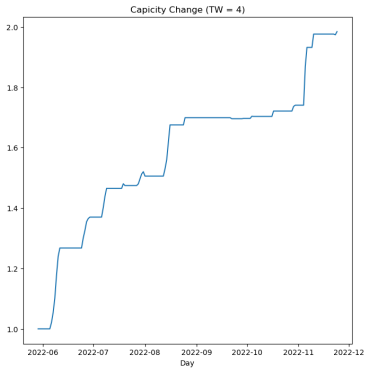


Hình 3.15: TW = 8
CC = 0%
Invest time = 0

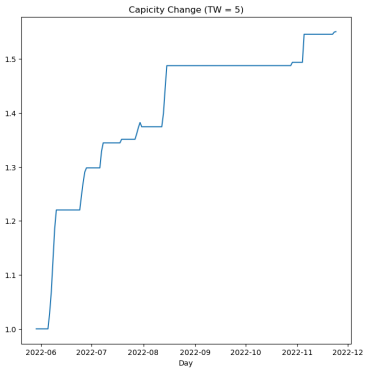
Bear market



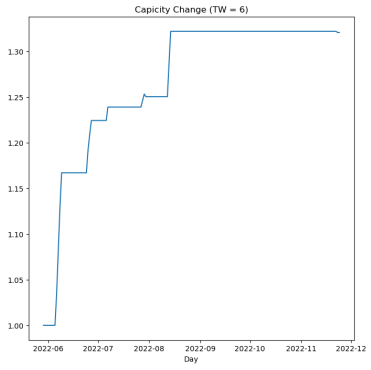
Hình 3.16: TW = 3
CC = 143.3%
Invest time = 59



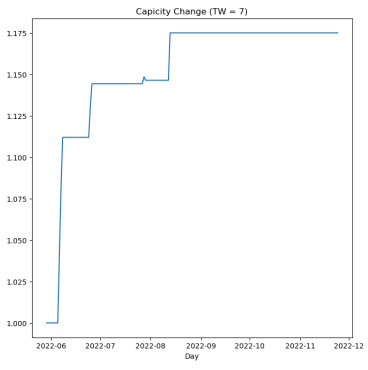
Hình 3.17: TW = 4
CC = 91.02%
Invest time = 38



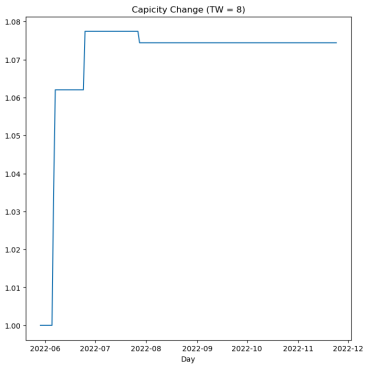
Hình 3.18: TW = 5
CC = 51.39%
Invest time = 24



Hình 3.19: TW = 6
CC = 30.25%
Invest time = 14

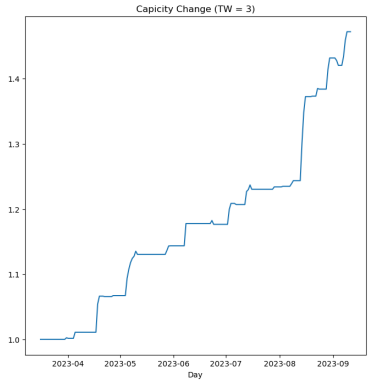


Hình 3.20: TW = 7
CC = 16.56%
Invest time = 8

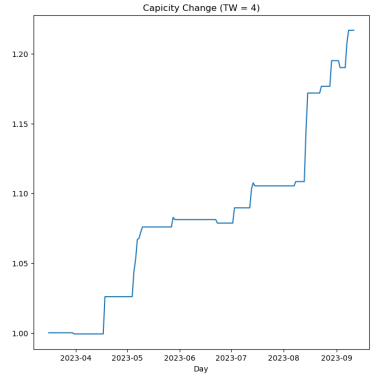


Hình 3.21: TW = 8
CC = 7.01%
Invest time = 4

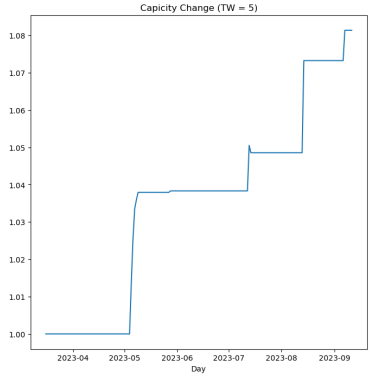
Sideway market



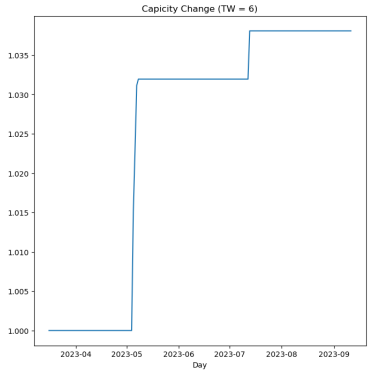
Hình 3.22: TW = 3
CC = 40.88%
Invest time = 44



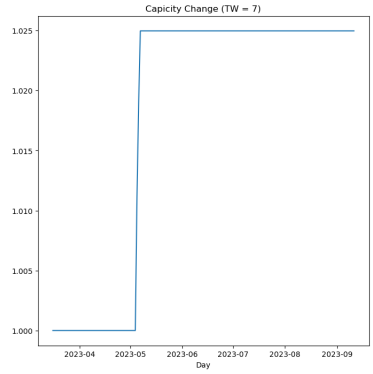
Hình 3.23: TW = 4
CC = 18.91%
Invest time = 23



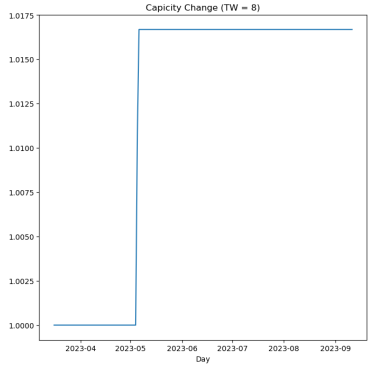
Hình 3.24: TW = 5
CC = 7.06%
Invest time = 10



Hình 3.25: TW = 6
CC = 3.29%
Invest time = 5



Hình 3.26: TW = 7
CC = 2.19%
Invest time = 3

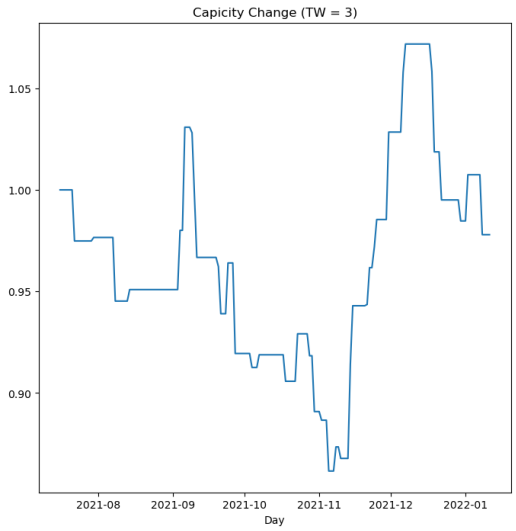


Hình 3.27: TW = 8
CC = 1.47%
Invest time = 2

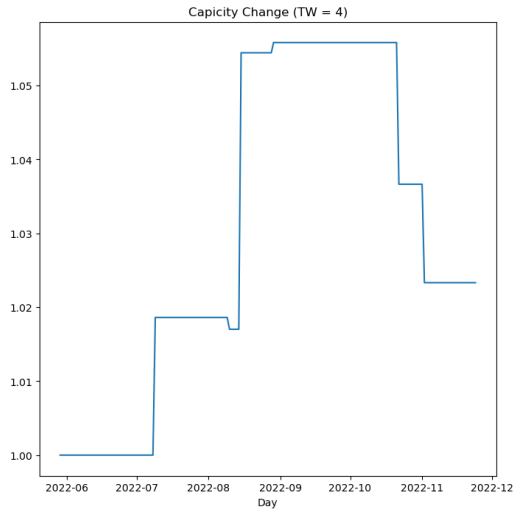
Mô hình đầu tư sử dụng dữ liệu giá BTC trong thực tế trong khoảng thời gian 6 tháng không sử dụng đòn bảy. Có thể thấy mô hình đầu tư hoạt động rất tốt ở cả 3 loại thị trường với các ***TW*** khác nhau. Nếu LSTM dự đoán tốt đường giá của BTC, ta sẽ có một công cụ đầu tư rất mạnh.

3.3.2. Kiểm định hiệu quả mô hình với các thị trường khác nhau sử dụng dữ liệu giá BTC dự đoán sử dụng mô hình LSTM

Bull market

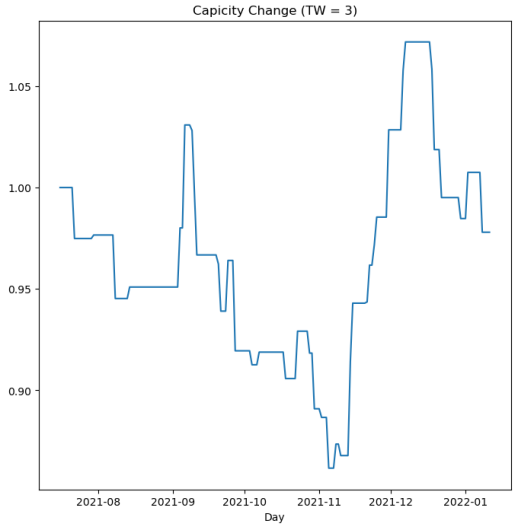


Hình 3.28: TW = 3
CC = -5.85%
Invest time = 38

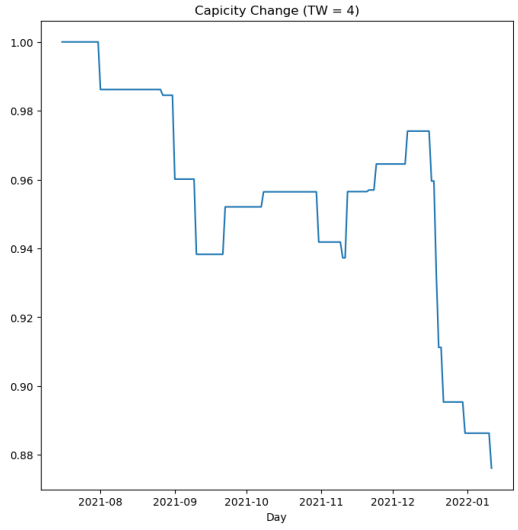


Hình 3.29: TW = 4
CC = -13.95%
Invest time = 18

Bear market

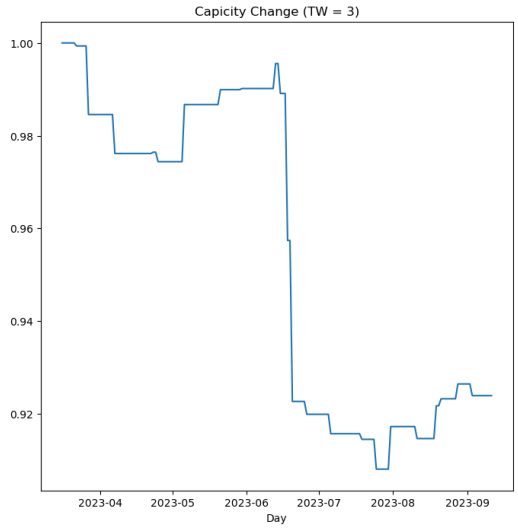


Hình 3.30: TW = 3
CC = 2.95%
Invest time = 21

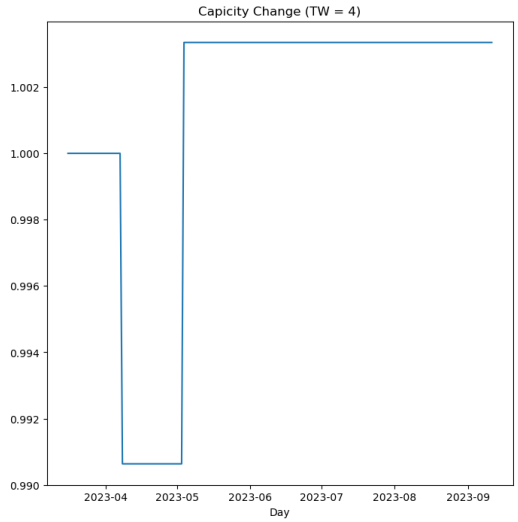


Hình 3.31: TW = 4
CC = 1.72%
Invest time = 6

Sideway market



Hình 3.32: TW = 3
CC = -9.62%
Invest time = 22



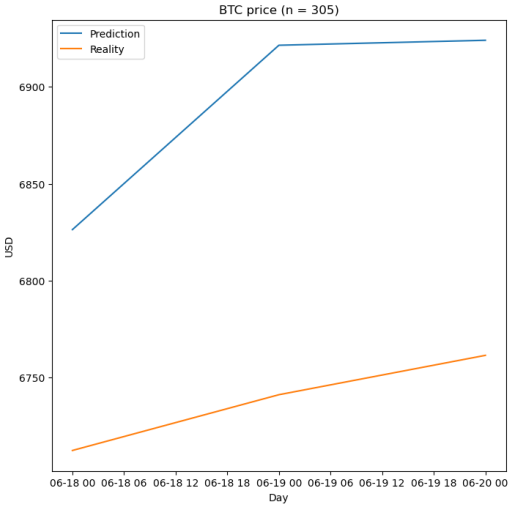
Hình 3.33: TW = 4
CC = 0.13%
Invest time = 2

3.3.3. Nhận xét kết quả kiểm định mô hình

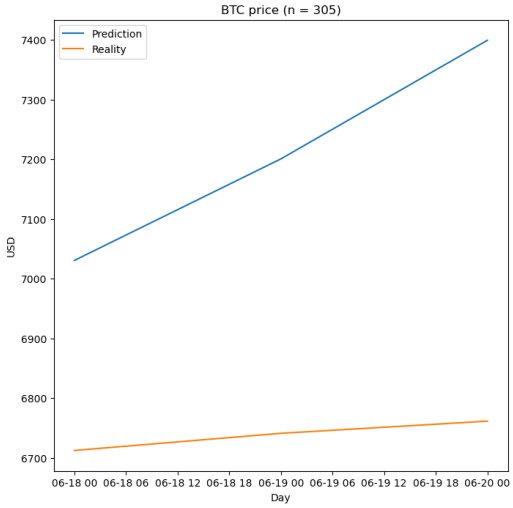
- Có thể thấy hiệu quả đầu tư ở cả 3 thị trường là chưa cao. Dựa vào kết quả đầu tư ở Mục 3.3.1 , đánh giá được rằng độ hiệu quả thấp là do dự đoán của mô hình LSTM còn yếu, chứ không phải do chiến thuật đầu tư.

- Do hạn chế phần cứng, nên em phải lựa chọn $\boldsymbol{epochs = 10}$ thay vì **100**. Dù đã chọn epochs thấp, trung bình thời gian chạy mô hình đầu tư của em là 400 phút, khoảng 6 - 7 tiếng, một khoảng thời gian khá dài, và quan sát thấy số lượt đầu tư (Invest time) ở $\boldsymbol{TW = 4}$ đã rất thấp, em không kiểm định mô hình với các $\boldsymbol{TW}$ cao hơn.

- Ngay cả với epochs cao, có thể tháy dự đoán của LSTM vẫn chưa tốt, lệch khỏi đường giá thực tế một khoảng khá lớn.



Hình 3.34: $\boldsymbol{epochs = 100}$



Hình 3.35: $\boldsymbol{epochs = 200}$

Vậy, ta cần tìm kiếm một thuật toán dự đoán giá BTC chính xác hơn.

Kết luận

Kết quả nghiên cứu

Đồ án đã đạt được mục tiêu đề ra. Đồ án đã trình bày 3 thuật toán tự động: một thuật toán lựa chọn các Crypto uy tín; một thuật toán phân bổ vốn đầu tư cho các Crypto dựa trên nguyên tắc đối sánh mẫu theo từng ngày và một thuật toán đầu tư bán khống ngắn hạn. Đồ án cho thấy sự phụ thuộc về giá của các đồng coin khác vào BTC. Đồ án đã xây dựng chương trình dựa trên Python, đồng thời kiểm định hiệu quả của thuật toán.

Định hướng nghiên cứu trong tương lai

Trong tương lai, em mong muốn có thể cải thiện hiệu quả thuật toán bằng các tham số khác hoặc áp dụng thuật toán cho các thị trường tài chính khác. Đồng thời, em cũng mong muốn nghiên cứu thêm về các thuật toán nhằm tối ưu hóa lợi nhuận khi đầu tư Crypto nhằm cải tiến thuật toán.

Tài liệu tham khảo

Tiếng Việt

[1] Nguyễn Thị Thu Thủy, Bài giảng Suy luận thống kê, 2023.

Tiếng Anh

[2] Lorenzo L, Arroyo J, "Online risk-based portfolio allocation on subsets of crypto assets applying a prototype-based clustering algorithm", *Financial Innovation*, 9: 25 (2023). https://doi.org/10.1186/s40854-022-00438-2.

[3] Khedmati M, Azin P, "An online portfolio selection algorithm using clustering approaches and considering transaction costs", 159: 30 (2020), 1135462020.

[4] G. Kapetanios,Y. Shin,A. Snell, "Testing for a unit root in the nonlinear star framework", *J Econom*, 112(2):359–379, 2003.

[5] Gujarati D, "Econometrics by Example", *Bloomsbury Publishing*, 2nd, 385 pages, 2011.

[6] Bruce M. Hill, "A simple general approach to inference about the tail of a distribution", *The Annals of Statistics*, 3(5):1163–1174, 1975.

Website

[7] https://www.scribbr.com/statistics/normal-distribution/

[8] https://seekingalpha.com/article/4088965-hunting-and-profiting-from-fat-tails-be-long-highvol-assets

[9] https://machinelearningcoban.com/2017/01/01/kmeans/

[10] https://in.tradingview.com/ideas/us/

[11] https://phamdinhkhanh.github.io/content

[12] https://drive.google.com/file/d/1lNjzISABdoc7SRq8tg-xkCRRZRABPCKi/view