Xây dựng và giải thích các đặc trưng cho mô hình

1 Các đặc trưng cho mô hình

1.1 Log Return

Khi sử dụng giá tuyệt đối (giá gốc), chênh lệch lợi nhuận được đo bằng:

$$\Delta P = P_t - P_{t-1}$$

Cách đo này phụ thuộc trực tiếp vào quy mô giá. Nếu toàn bộ mức giá được nhân lên bởi một hằng số k (do lạm phát hoặc sự tăng trưởng dài hạn của thị trường), sự chênh lệch cũng sẽ bị phóng đại theo:

$$\Delta(kP) = k(P_t - P_{t-1})$$

Ngược lại, Log Return (lợi nhuận logarit) đo lường tỉ lệ thay đổi tương đối của giá tài sản giữa hai thời điểm liên tiếp, được tính bằng công thức:

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right) = \ln(P_t) - \ln(P_{t-1})$$

Log Return chỉ phản ánh tỉ lệ thay đổi của giá, không bị ảnh hưởng bởi mức giá tuyệt đối. Đây là đặc tính quan trọng khi so sánh biến động giữa các giai đoạn có mức giá khác nhau. Ta có:

$$\ln(kP_t) - \ln(kP_{t-1}) = \ln(k) + \ln(P_t) - (\ln(k) + \ln(P_{t-1})) = \ln(P_t) - \ln(P_{t-1})$$

Ngoài ra, còn một vấn đề gặp phải đối với lợi nhuận thông thường là bất đối xứng giữa tăng và giảm giá. Giả sử ban đầu ta có giá ban đầu là 100. Khi giảm 50% giá trị (từ 100 về 50), ta cần phải tăng 100% giá trị của nó (từ 50 lên 100) để quay về mức ban đầu. Hai mức biến động -50% và $+100\%$ không đối xứng, gây khó khăn cho các mô hình học máy khi học động lực thị trường.

Ví dụ trên được biểu diễn bằng chỉ số Log Return như sau:

$$r_1 = \ln\left(\frac{50}{100}\right) = \ln(0.5) \quad r_2 = \ln\left(\frac{100}{50}\right) = \ln(2)$$

Về mặt toán học, $ln(2) = -\ln(0.5)$ cho thấy mức tăng và mức giảm để bù đắp cho nhau có tính đối xứng.

Nhờ tính bất biến theo tỷ lệ và khả năng xử lý bất đối xứng tăng – giảm, Log Return thường tạo ra chuỗi dữ liệu ổn định hơn về mặt thống kê và có xu hướng tiệm cận phân phối chuẩn hơn so với chuỗi giá gốc. Vì vậy, Log Return là biến đầu vào nền tảng trong các mô hình Machine Learning phân tích định lượng trong lĩnh vực tài chính.

1.2 Distance to Simple Moving Average (SMA)

Simple Moving Average (SMA) được tính bằng trung bình cộng giá đóng cửa của tài sản trong một khoảng thời gian xác định. Công thức tổng quát:

$$\text{SMA}_n(t) = \frac{1}{n}\sum_{i=0}^{n-1} P_{t-i}$$

Trong đó P_t là giá đóng cửa tại thời điểm t, và n là số phiên được sử dụng để tính trung bình. Khi số phiên càng lớn, SMA phản ánh xu hướng dài hạn; ngược lại, SMA với số phiên nhỏ sẽ phản ứng nhanh với biến động giá ngắn hạn.

Tuy nhiên, việc sử dụng trực tiếp giá trị của SMA chưa cung cấp đầy đủ thông tin cho mô hình Machine Learning. Chỉ số có ý nghĩa hơn là mức độ sai lệch giữa giá hiện tại và mức giá cân bằng này. Vì vậy, đặc trưng Distance to SMA được xây dựng nhằm biểu diễn khoảng cách tương đối giữa giá và đường trung bình động:

$$\text{Dist\_SMA}_n(t) = \frac{P_t - \text{SMA}_n(t)}{\text{SMA}_n(t)}$$

Đặc trưng này có tính chất bất biến theo tỷ lệ giá. Biểu diễn sai lệch dưới dạng tỷ lệ giúp mô hình so sánh được trạng thái thị trường một cách nhất quán qua nhiều giai đoạn mà không bị ảnh hưởng bởi mức giá tuyệt đối, tương tự như ưu điểm của Log Return. Trong bài thực hành này, hai đường trung bình động được sử dụng là SMA_20 đại diện cho xu hướng ngắn hạn và SMA_50 phản ánh xu hướng trung hạn. Tuy nhiên,

tôi cho rằng việc sử dụng SMA cũng tồn tại hạn chế, do đây là chỉ báo dựa hoàn toàn trên dữ liệu quá khứ nên thường có độ trễ nhất định trong việc phản ứng với các biến động giá đột ngột.

1.3 MACD Histogram

Moving Average Convergence Divergence (MACD) đo lường mối quan hệ giữa hai đường trung bình động lũy thừa (Exponential Moving Average – EMA) của giá. MACD được xây dựng dựa trên EMA ngắn hạn 12 phiên và EMA dài hạn 26 phiên:

$$\text{MACD}(t) = \text{EMA}_{12}(t) - \text{EMA}_{26}(t)$$

Đường tín hiệu (Signal Line) là trung bình động lũy thừa 9 phiên của chuỗi MACD, được sử dụng để giảm ảnh hưởng của các dao động ngắn hạn:

$$\text{Signal}(t) = \text{EMA}_9(\text{MACD}(t))$$

Từ hai đại lượng trên, MACD Histogram được định nghĩa là hiệu số giữa MACD Line và Signal Line:

$$\text{MACD\_Hist}(t) = \text{MACD}(t) - \text{Signal}(t)$$

MACD phản ánh tốc độ thay đổi của xu hướng giá, trong khi MACD Histogram đo lường sự thay đổi của động lượng đó. Có thể hiểu rằng MACD Histogram đóng vai trò như gia tốc của chuyển động giá. Trong mô hình Machine Learning, MACD Histogram cung cấp thông tin sớm về sự thay đổi động lượng mà bản thân giá hoặc đường trung bình chưa kịp phản ánh, giúp mô hình hiểu được giá đang tăng là do một xu hướng thực sự chuẩn bị tới hay chỉ là những nhịp tăng ngắn hạn nhất thời.

1.4 Relative Strength Index (RSI)

Relative Strength Index (RSI) được sử dụng để đánh giá mức độ mạnh – yếu của biến động giá trong ngắn hạn. Chỉ báo này giúp nhận diện trạng thái quá mua (overbought) và quá bán (oversold) của thị trường thông qua việc so sánh cường độ các

phiên tăng giá và giảm giá trong một khoảng thời gian xác định.

$$RSI = 100 - \frac{100}{1 + RS} \quad \text{với} \quad RS = \frac{\text{Mức tăng trung bình}}{\text{Mức giảm trung bình}}$$

Trong đó, mức tăng trung bình và mức giảm trung bình phản ánh cường độ biến động giá theo hai chiều tăng và giảm trong khoảng thời gian quan sát. Khi các phiên tăng chiếm ưu thế, giá trị RSI sẽ tăng; ngược lại, khi áp lực bán mạnh hơn, RSI có xu hướng giảm. $RSI > 70$ thường cho thấy thị trường đang ở trạng thái quá mua, còn $RSI < 30$ thường phản ánh trạng thái quá bán.

Trong các mô hình Machine Learning, RSI được sử dụng như một đặc trưng bổ trợ giúp mô hình nắm bắt trạng thái tâm lý ngắn hạn của thị trường và hạn chế các tín hiệu nhiễu xuất hiện trong những giai đoạn giá biến động quá mức.

1.5 Rolling Volatility

Volatility (độ biến động) là mức độ dao động của lợi suất trong một khoảng thời gian nhất định, phản ánh mức độ rủi ro và sự bất ổn của thị trường. Độ biến động được tính dựa trên độ lệch chuẩn của log-return trong một cửa sổ trượt 20 phiên:

$$\text{Volatility}_{20}(t) = \sqrt{\frac{1}{20-1} \sum_{i=0}^{19} (r_{t-i} - \bar{r})^2}$$

Trong đó r_t là log-return tại thời điểm t, và $\bar{r}$ là giá trị trung bình của log-return trong 20 phiên gần nhất. Việc sử dụng log-return, như đã giải thích ở trên, giúp đặc trưng này có tính bất biến theo tỷ lệ và phản ánh chính xác hơn mức độ biến động thực tế của thị trường.

Rolling Volatility cho phép mô hình nắm bắt trạng thái rủi ro động của thị trường. Giá trị volatility cao thường gắn với các giai đoạn thị trường biến động mạnh, tâm lý nhà đầu tư bất ổn hoặc xuất hiện thông tin mới. Ngược lại, volatility thấp phản ánh thị trường ổn định, ít dao động và chưa có xu hướng hay động lực nào làm thay đổi.

1.6 Bollinger Bands

Bollinger Bands là một chỉ báo dùng để mô tả mức độ biến động của giá xoay quanh một mức trung bình. Chỉ báo này bao gồm ba thành phần: dải giữa (Middle Band), dải trên (Upper Band) và dải dưới (Lower Band).

Dải giữa được xác định bằng đường trung bình động giản đơn 20 phiên của giá đóng cửa:

$$\mathrm{BB}_{\mathrm{Middle}}(t) = \mathrm{SMA}_{20}(t)$$

Hai dải trên và dưới được xây dựng bằng cách cộng và trừ một bội số của độ lệch chuẩn 20 phiên vào dải giữa:

$$\mathrm{BB}_{\mathrm{Upper}}(t) = \mathrm{SMA}_{20}(t) + 2 \cdot \sigma_{20}(t)$$

$$\mathrm{BB}_{\mathrm{Lower}}(t) = \mathrm{SMA}_{20}(t) - 2 \cdot \sigma_{20}(t)$$

Trong đó $\sigma_{20}(t)$ là độ lệch chuẩn của giá đóng cửa trong 20 phiên gần nhất.

Bollinger Band Width được tính bằng công thức:

$$\mathrm{BB\_Width}(t) = \frac{\mathrm{BB}_{\mathrm{Upper}}(t) - \mathrm{BB}_{\mathrm{Lower}}(t)}{\mathrm{BB}_{\mathrm{Middle}}(t)}$$

Chỉ số này đánh giá trạng thái giãn nở hoặc thu hẹp của biến động thị trường. Khi giá dao động mạnh, khoảng cách giữa hai dải mở rộng; ngược lại, trong các giai đoạn thị trường đi ngang hoặc thiếu động lực, hai dải có xu hướng thu hẹp lại. Việc chuẩn hóa độ rộng dải Bollinger bằng giá trị trung bình giúp đặc trưng này không phụ thuộc vào mức giá tuyệt đối của tài sản, cho phép so sánh trạng thái biến động giữa các giai đoạn thị trường khác nhau. BB Width cao thường phản ánh các pha thị trường biến động mạnh, rủi ro gia tăng hoặc xuất hiện thông tin mới. Ngược lại, BB Width thấp cho thấy thị trường đang trong giai đoạn tích lũy, ít biến động và có khả năng chuẩn bị cho một chuyển động giá lớn trong tương lai.

1.7 High–Low Range

High–Low Range là một đặc trưng đơn giản dùng để đo lường biên độ dao động của giá trong một phiên giao dịch. Chỉ số này phản ánh mức chênh lệch giữa giá cao nhất và giá thấp nhất trong phiên:

$$\text{HL\_Range}(t) = \frac{P_t^{\text{High}} - P_t^{\text{Low}}}{P_t^{\text{Close}}}$$

Trong đó P_t^{High} và P_t^{Low} lần lượt là giá cao nhất và thấp nhất tại thời điểm t, còn P_t^{Close} là giá đóng cửa của phiên giao dịch.

Việc chuẩn hóa biên độ dao động bằng giá đóng cửa giúp đặc trưng này có tính bất biến theo tỷ lệ giá, tương tự như Log Return hay Bollinger Band Width. Nhờ đó, mô hình có thể so sánh mức độ biến động nội phiên giữa các giai đoạn thị trường có mức giá khác nhau.

High–Low Range phản ánh mức độ giằng co giữa lực mua và lực bán trong một phiên. Giá trị HL_Range cao thường xuất hiện trong các giai đoạn thị trường biến động mạnh, tâm lý nhà đầu tư thiếu ổn định hoặc khi có thông tin mới tác động đến kỳ vọng giá. Ngược lại, HL_Range thấp cho thấy thị trường đang giao dịch trong biên độ hẹp, phản ánh trạng thái cân bằng.

So với Rolling Volatility đo lường biến động theo chuỗi thời gian, High–Low Range tập trung vào biến động tức thời trong từng phiên. Việc kết hợp hai đặc trưng này giúp mô hình phân biệt được các pha biến động bền vững kéo dài nhiều phiên, đồng thời vẫn có thể tập trung nắm bắt các dao động ngắn hạn tức thời.

1.8 Volume Ratio

Khối lượng giao dịch (Volume) phản ánh mức độ tham gia của dòng tiền vào thị trường, tuy nhiên, giá trị khối lượng tuyệt đối thường phụ thuộc mạnh vào quy mô thị trường và thay đổi đáng kể theo thời gian, khiến việc so sánh trực tiếp giữa các giai đoạn trở nên kém hiệu quả.

Để khắc phục vấn đề này, đặc trưng Volume Ratio được xây dựng nhằm đo lường mức độ bất thường của khối lượng giao dịch hiện tại so với mức trung bình gần đây:

$$\text{Vol\_Ratio}(t) = \frac{V_t}{\text{MA}_{20}(V_t)}$$

Trong đó V_t là khối lượng giao dịch tại thời điểm t, và $\text{MA}_{20}(V_t)$ là trung bình động 20 phiên của khối lượng giao dịch.

Volume Ratio cho biết khối lượng hiện tại đang cao hay thấp hơn bao nhiêu lần so với trạng thái bình thường của thị trường trong ngắn hạn. Khi Vol_Ratio > 1, thị trường đang ghi nhận sự gia tăng về mức độ tham gia của dòng tiền; giá trị càng lớn cho thấy dòng tiền vào thị trường càng mạnh. Ngược lại, Vol_Ratio < 1 phản ánh trạng thái giao dịch trầm lắng.

1.9 Time Encoding (Cyclical Time Features)

Dữ liệu tài chính nói chung và dữ liệu giá ETHUSDT nói riêng là dữ liệu chuỗi thời gian (time series), trong đó các quan sát không độc lập và phân phối của dữ liệu có thể thay đổi theo thời gian (non-IID). Thực tế thường cho thấy các chuỗi giá tài chính không thể hiện rõ tính chu kỳ cố định như các hiện tượng vật lý hay khí tượng do chịu ảnh hưởng mạnh từ hành vi con người. Tuy nhiên, giả thuyết rằng thị trường có thể tồn tại các yếu tố mang tính chu kỳ yếu hoặc cấu trúc thời gian tiềm ẩn, ví dụ, cường độ giao dịch và mức độ biến động có thể khác nhau giữa các khung giờ trong ngày, giữa ngày trong tuần và cuối tuần, hoặc giữa các giai đoạn khác nhau trong năm. Các hiện tượng này không mang tính ổn định và không phải lúc nào cũng xuất hiện nhất quán trong dữ liệu tài chính. Vì vậy, các đặc trưng thời gian được đưa vào với mục tiêu kiểm chứng liệu mô hình có thể khai thác được cấu trúc chu kỳ tiềm ẩn nào từ dữ liệu hay không, kiểm chứng xem liệu các mẫu chu kỳ này có tồn tại và có ý nghĩa thống kê hay không.

Thay vì sử dụng trực tiếp các biến rời rạc như giờ, thứ trong tuần hoặc tháng trong năm, các đặc trưng thời gian được mã hóa dưới dạng hàm sin và cos nhằm phản ánh bản chất chu kỳ liên tục của thời gian:

$$\text{Time}_{\sin} = \sin\left(2\pi\frac{t}{T}\right), \quad \text{Time}_{\cos} = \cos\left(2\pi\frac{t}{T}\right)$$

Trong đó t là vị trí thời gian hiện tại trong chu kỳ, và T là độ dài của chu kỳ tương

ứng. Cách mã hóa này cho phép biểu diễn thời gian như một đại lượng tuần hoàn, tương tự cách các góc được biểu diễn lên đường tròn lượng giác trong toán học.

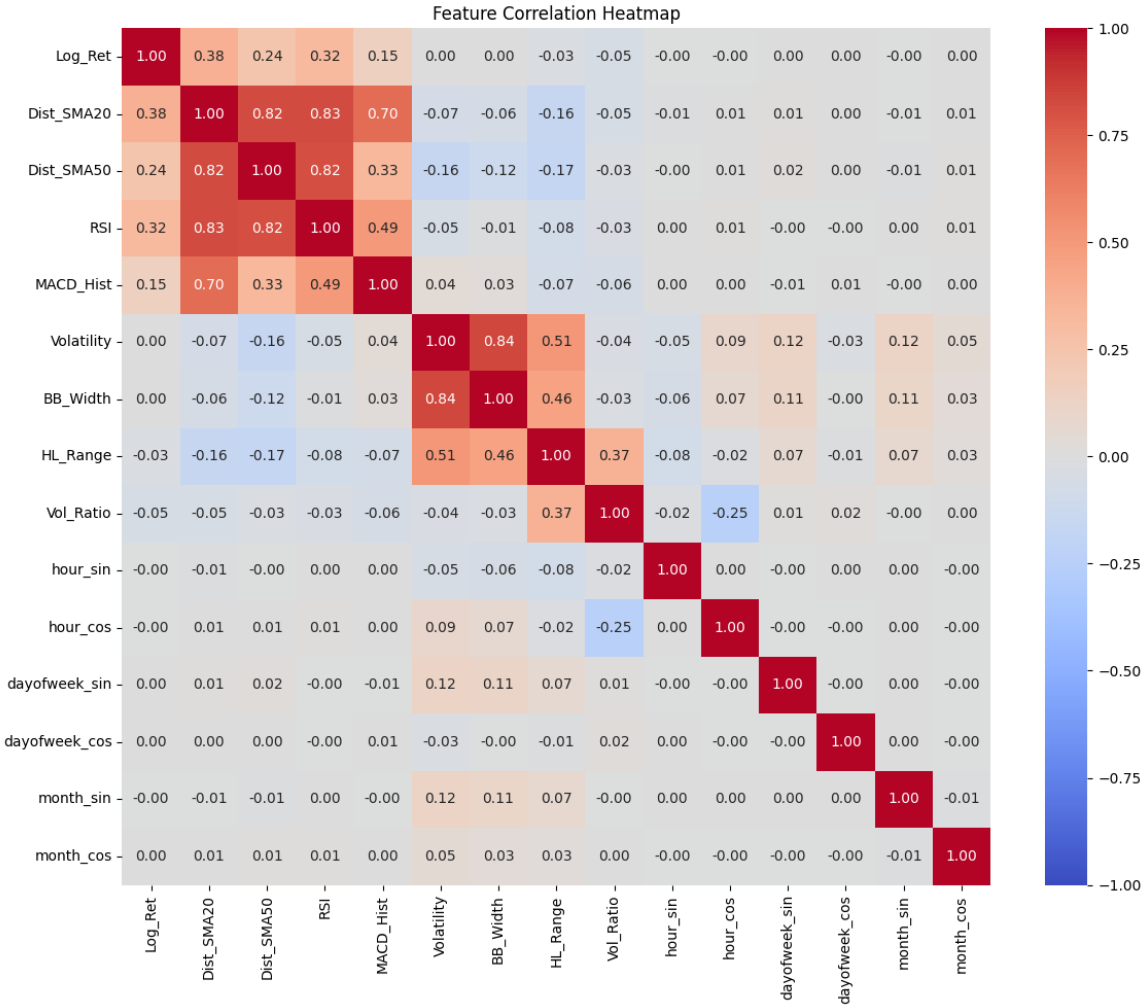
Ba nhóm đặc trưng thời gian được xây dựng trong nghiên cứu này bao gồm:

- Chu kỳ trong ngày: mã hóa theo phút trong ngày với chu kỳ $T = 1440$ phút.
- Chu kỳ theo tuần: mã hóa ngày trong tuần với chu kỳ $T = 7$.
- Chu kỳ theo năm: mã hóa tháng trong năm với chu kỳ $T = 12$

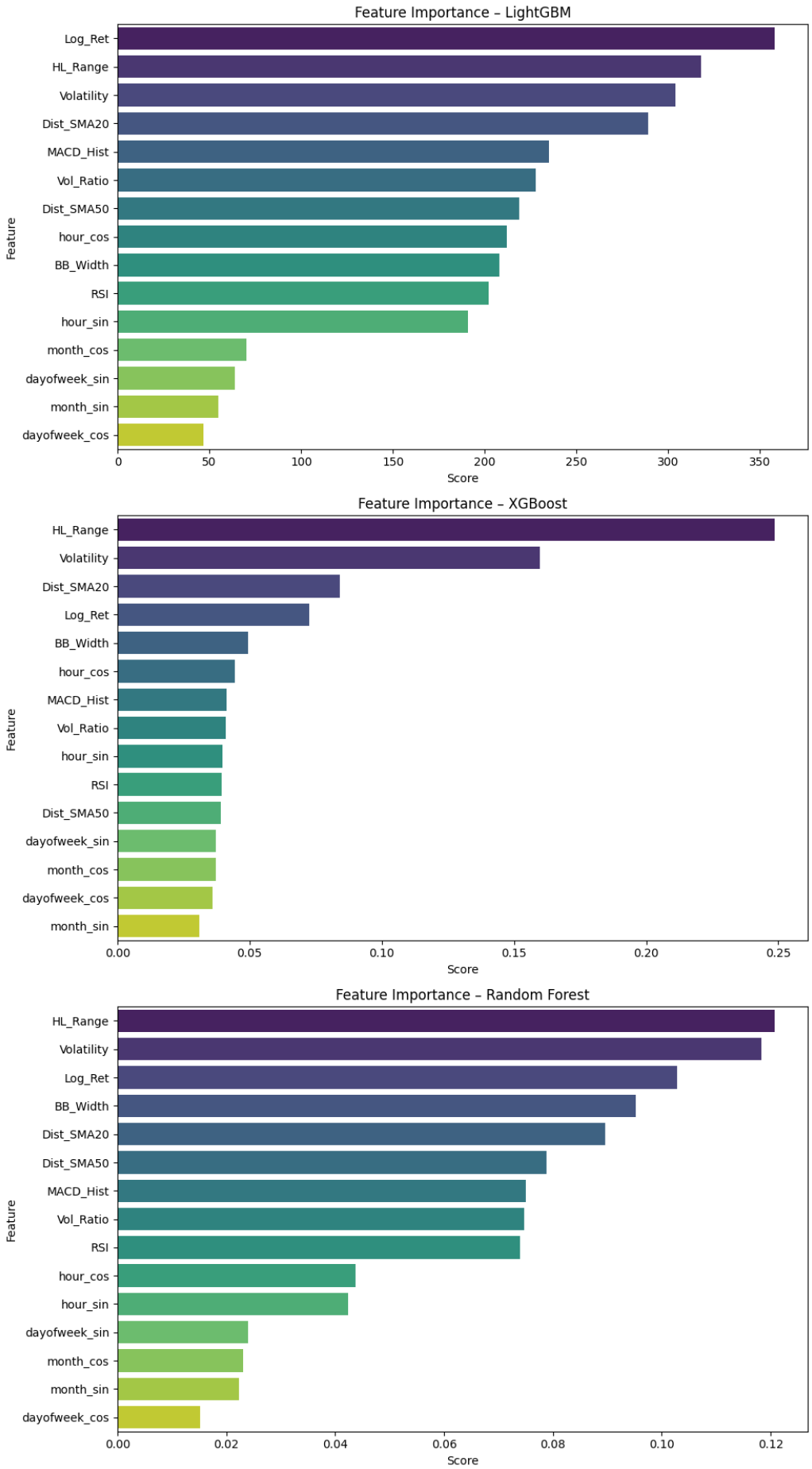
Việc mã hóa thời gian bằng cặp đặc trưng sin và cos cho mỗi chu kỳ giúp bảo toàn tính liên tục của trục thời gian, tránh hiện tượng đứt gãy. Chẳng hạn, hai thời điểm 23 giờ khuya và 0 giờ ngày hôm sau là liền kề trong thực tế nhưng lại cách xa nhau nếu biểu diễn trực tiếp bằng giá trị số nguyên.

Các đặc trưng thời gian này được đưa vào mô hình như biến bổ trợ, nhằm kiểm chứng sự tồn tại của các cấu trúc chu kỳ tiềm ẩn trong dữ liệu, đồng thời cung cấp thêm ngữ cảnh thời gian để mô hình tự đánh giá mức độ hữu ích của chúng trong từng giai đoạn thị trường.

1.10 Mối quan hệ và tính bổ trợ giữa các đặc trưng



Hình 1: Ma trận tương quan giữa các features



Hình 2: Độ quan trọng của từng đặc trưng đối với mỗi mô hình

Từ ma trận tương quan và kết quả đánh giá mức độ quan trọng của đặc trưng trong các mô hình LightGBM, XGBoost và Random Forest, có thể nhận thấy rằng các đặc trưng hình thành những nhóm thông tin có tính bổ trợ lẫn nhau để mô tả trạng thái thị trường một cách toàn diện.

Nhóm đặc trưng liên quan đến xu hướng và động lượng bao gồm Distance to SMA (20, 50), RSI và MACD Histogram có mức tương quan tương đối cao. Điều này phản ánh việc các chỉ báo này cùng mô tả hành vi chuyển động của giá, nhưng ở các góc nhìn khác nhau. Distance to SMA thể hiện vị trí của giá so với mức cân bằng động của thị trường, RSI phản ánh cường độ mua–bán trong ngắn hạn, còn MACD Histogram đo lường sự thay đổi của động lượng xu hướng. Khi được sử dụng đồng thời, nhóm đặc trưng này giúp mô hình phân biệt rõ hơn giữa xu hướng dài hạn và các biến động mang tính điều chỉnh ngắn hạn.

Nhóm đặc trưng đo lường biến động thị trường như Rolling Volatility, Bollinger Band Width và High–Low Range thể hiện vai trò nổi bật trong cả ba mô hình. Kết quả feature importance cho thấy HL_Range và Volatility thường nằm trong nhóm đặc trưng quan trọng nhất. Các đặc trưng này có mức tương quan nội bộ tương đối cao nhưng tương quan thấp với nhóm xu hướng, cho thấy chúng cung cấp một chiều thông tin khác về thị trường là mức độ rủi ro và trạng thái biến động. Việc kết hợp các thước đo biến động ở nhiều thang thời gian giúp mô hình nhận diện tốt hơn các biến động của thị trường.

Mặc dù không có tương quan mạnh với các đặc trưng còn lại, Log Return lại có mức độ quan trọng cao trong mô hình. Điều này cho thấy mô hình vẫn dựa vào động lực giá cơ bản như một nguồn thông tin độc lập, đóng vai trò nền tảng giúp mô hình nắm bắt chuyển động thực tế của thị trường thay vì chỉ dựa vào các biến đổi của giá trong quá khứ.

Volume Ratio bổ sung thêm chiều thông tin về mức độ tham gia của dòng tiền. Mặc dù tương quan của đặc trưng này với các biến khác không cao, mức độ quan trọng trung bình của nó trong các mô hình cho thấy khối lượng giao dịch đóng vai trò xác nhận cho biến động giá và động lượng thị trường.

Cuối cùng, các đặc trưng mã hóa thời gian có tương quan gần như bằng không với

các nhóm đặc trưng khác và mức độ quan trọng tương đối thấp. Điều này cho thấy dữ liệu không thể hiện rõ cấu trúc chu kỳ thời gian mạnh, và các biến thời gian chủ yếu đóng vai trò bổ trợ cho các đặc trưng quan trọng nêu trên.