ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

□□□□□

PHÂN TÍCH ĐỊNH LƯỢNG VÀ CÁC NHÂN TỐ ẢNH HƯỞNG ĐẾN BIẾN ĐỘNG GIÁ CỔ PHIẾU TẬP ĐOÀN FPT

Nhóm 31			
Sinh viên thực hiện:			
STT	Họ tên	MSSV	Ngành
1	Nguyễn Hoàng Long	23520882	KHDL
2	Hồ Tấn Dũng	23520327	KHDL

TP. HỒ CHÍ MINH – 12/2025

1. GIỚI THIỆU

Đồ án tập trung phân tích diễn biến và các yếu tố tác động đến giá đóng cửa của cổ phiếu *Tập đoàn FPT* trong giai đoạn từ 01/01/2020 đến 30/09/2025 nhằm cung cấp những góc nhìn khách quan dựa trên dữ liệu về xu hướng và độ biến động thị trường. Về phương pháp luận, nghiên cứu thực hiện phân tích đơn biến qua việc phân rã chuỗi thời gian, kết hợp mô hình ARIMA - GARCH(1,1) và mô phỏng Monte-Carlo để xác định xu hướng biến động dài hạn của chuỗi lợi suất logarit. Song song đó, đồ án cũng thử nghiệm dự báo giá dựa trên các chỉ số vĩ mô, dữ liệu ngành và báo cáo tài chính bằng mô hình XGBoost. Kết quả thực nghiệm cho thấy cổ phiếu tập đoàn FPT duy trì đà tăng trưởng ổn định trong dài hạn với biến động có kiểm soát trong hầu hết thời gian được khảo sát. Đối với phần mô hình hóa dự đoán biến mục tiêu, XGBoost thể hiện ưu thế vượt trội với chỉ số ***R2*** đạt 0.4976.

Bộ dữ liệu phân tích do nhóm tự thu thập trên các nguồn dữ liệu thị trường, bao gồm CafeF, Vietstock và Investing (chi tiết tại danh mục ***Tài liệu tham khảo***). Sau đó, chúng tôi tiến hành chọn lọc các biến cụ thể (xem phần 2) và tổng hợp vào một bảng dữ liệu thống nhất. Trong quá trình thực hiện, chúng tôi đã tiến hành đồng bộ hóa chuỗi thời gian giữa giá đóng cửa cổ phiếu FPT với các tác nhân ngoại cảnh như giá vàng và tỷ giá USD/VND, đồng thời thực hiện các bước xử lý sau đó (điền dữ liệu trống, kiểm định tính dừng, chuẩn hóa) nhằm phục vụ cho nhiệm vụ phân tích cụ thể sau đó.

2. MÔ TẢ BỘ DỮ LIỆU

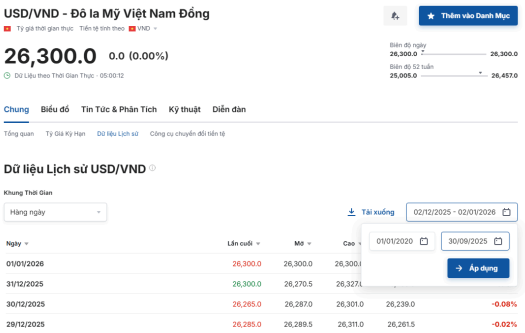
Dữ liệu thực nghiệm trong đồ án được tổng hợp từ các nguồn tài chính uy tín trên Internet, sau đó được chuẩn hóa và nhất quán hóa thành một bộ dữ liệu đồng bộ. Về quy ước trình bày, các chỉ số về giá trị tiền tệ trong bảng dữ liệu được mặc định tính theo đơn vị Đồng Việt Nam (VND), trừ trường hợp có ghi chú cụ thể khác.

STT	Tên cột	Kiểu dữ liệu	Giải thích
1	cpi_rate	float	Chỉ số giá tiêu dùng (Consumer Price Index) tại Việt Nam
2	gdp_value	float	Tổng sản phẩm quốc nội (Gross Domestic Product) của Việt Nam
3	usd_vnd_rate	float	Tỷ giá Đô la Mỹ - Đồng Việt Nam (USD-VND)
4	xau_usd_rate	float	Giá vàng tính theo giá đô la Mỹ (USD)
5	market_cap	float	Vốn hóa thị trường ngành ICT
6	pe_ratio	float	Giá thị trường của cổ phiếu (P) và thu nhập trên mỗi cổ phiếu (EPS)
7	fpt_net_revenue	float	Khoản doanh thu từ bán hàng và cung cấp dịch vụ của FPT
8	fpt_gross_profit	float	Số tiền còn lại sau khi lấy doanh thu thuần trừ đi giá vốn hàng bán
9	fpt_operating_profit	float	Khoản lợi nhuận thu được từ hoạt động cốt lõi của FPT
10	fpt_net_profit	float	Phần lợi nhuận thực tế sau thuế
11	fpt_stock_price	float	Giá cổ phiếu FPT
12	fpt_stock_volume	float	Khối lượng giao dịch cổ phiếu FPT

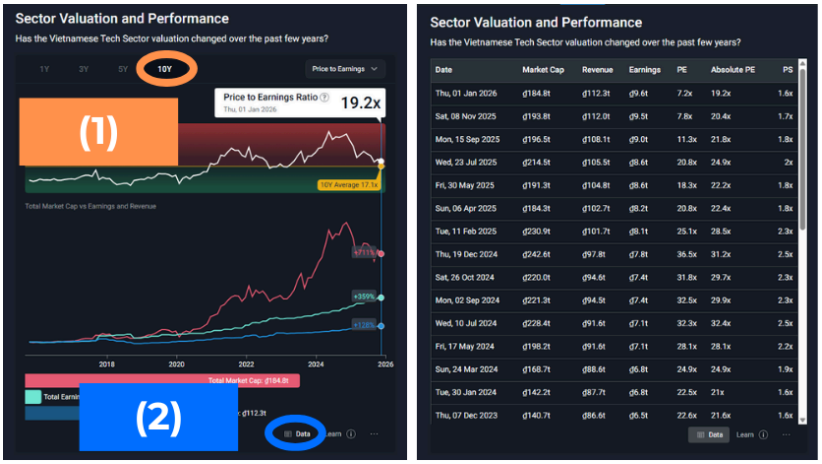
Bảng 1. Mô tả bộ dữ liệu được sử dụng trong đồ án

Dữ liệu nghiên cứu được thu thập thủ công thông qua các tệp định dạng excel hoặc csv từ nhiều nguồn khác nhau (xem thêm ở phần tài liệu tham khảo) thông qua các phương pháp khác nhau, cụ thể:

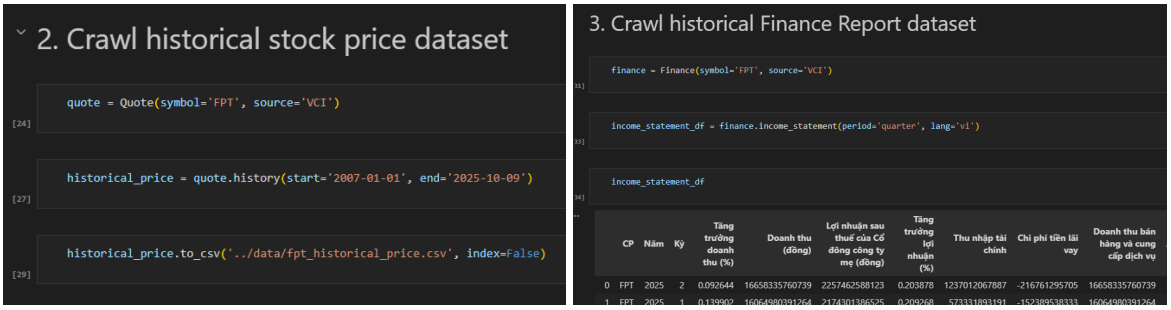
- *Phương pháp tải xuống file CSV* được áp dụng để thu thập dữ liệu cho các biến 'usd_vnd_rate', 'cpi_rate', 'gdp_value' và 'xau_usd_rate'. Cụ thể hơn, sau khi truy cập vào nguồn dữ liệu, nhóm đồ án tiến hành lọc chọn khoảng ngày tháng phù hợp rồi nhấn vào nút "Tải xuống" để tiến hành lấy dữ liệu (xem Hình 1).
- *Phương pháp chép lại thủ công dữ liệu* được áp dụng để thu thập dữ liệu cho các biến market_cap và pe_ratio. Tại nguồn dữ liệu, nhóm chọn khung thời gian 10 năm ('10Y') và thẻ 'data' để hiển thị bảng số liệu chi tiết (như minh họa tại Hình 2). Dữ liệu sau đó được sao chép và chuẩn hóa định dạng (ngày tháng, giá trị số) thông qua công cụ hỗ trợ. Để đảm bảo tính chính xác, nhóm thực hiện bước đối chiếu thủ công cuối cùng với nguồn gốc.
- *Phương pháp sử dụng API* được dùng để thu thập các biến còn lại liên quan tới FPT, bao gồm fpt_net_revenue, fpt_gross_profit, fpt_operating_profit, fpt_net_profit, fpt_stock_price, fpt_stock_volume thông qua API vnstock (Hình 3).



Hình 1: Cách thu thập dữ liệu bằng cách tải xuống file csv



Hình 2: Cách thu thập dữ liệu bằng cách thu thập thủ công (trái) và kết quả hiển thị (phải)



Hình 3: Sử dụng API thu thập giá cổ phiếu (trái) và báo cáo tài chính (phải) của FPT Corp

Sau khi thu thập xong, nhóm đồ án sử dụng Python để thực hiện hợp nhất các trường dữ liệu dựa trên khóa chính là biến thời gian. Bộ dữ liệu cuối cùng được chuẩn hóa và giới hạn trong giai đoạn từ ngày 01/01/2020 đến ngày 30/09/2025 theo đúng mục tiêu thực hiện ban đầu.

Bộ dữ liệu là một bảng dữ liệu chuỗi thời gian gồm 12 cột (như Bảng 1, Hình 4) với 1529 quan sát, tuy nhiên có sự chênh lệch lớn về tần suất xuất hiện giữa các nhóm biến số: trong khi dữ liệu giao dịch (tỷ giá, giá cổ phiếu) có độ bao phủ khá đầy đủ, thì các chỉ số vĩ mô và tài chính doanh nghiệp (CPI, GDP, kết quả kinh doanh FPT) lại có mật độ thưa thớt do đặc thù được công bố theo tháng, quý hoặc năm.

```
==================================================
KẾT QUẢ ĐIỀU TRA DATASET
==================================================
Tổng số cột là: 12
Tổng số mẫu là: 1529

<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 1529 entries, 2020-01-01 to 2025-09-30
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   cpi_rate             68 non-null     float64
 1   gdp_value            5 non-null      float64
 2   usd_vnd_rate         1501 non-null   float64
 3   xau_usd_rate         1495 non-null   float64
 4   market_cap           19 non-null     float64
 5   pe_ratio             19 non-null     float64
 6   fpt_net_revenue      22 non-null     float64
 7   fpt_gross_profit     22 non-null     float64
 8   fpt_operating_profit 22 non-null     float64
 9   fpt_net_profit       22 non-null     float64
 10  fpt_stock_price      1433 non-null   float64
 11  fpt_stock_volume     1433 non-null   float64
dtypes: float64(12)
memory usage: 155.3 KB
None
```

Hình 4: Bảng kết quả mô tả kiểu dữ liệu và số giá trị không rỗng từ bộ dữ liệu

3. PHƯƠNG PHÁP PHÂN TÍCH

Quy trình phân tích được thiết kế theo cấu trúc dòng chảy tuần tự kết hợp song song (xem Hình 5). Giai đoạn đầu của đồ án bắt đầu bằng việc thu thập dữ liệu, sau đó tiến hành tổng hợp dữ liệu để nhất quán hóa nguồn thông tin và thực hiện phân tích khám phá (EDA) nhằm có cái nhìn tổng quan ban đầu.

Hình 5: Kịch bản phân tích dữ liệu

Sau bước EDA, quy trình phân nhánh thành hai luồng xử lý độc lập. *Luồng thứ nhất* có nhiệm vụ phân tích nội tại biến giá đóng cửa. Tại đây, sau bước tiền xử lý, dữ liệu được khai thác theo ba hướng tiếp cận đồng thời: (1) thực hiện thống kê mô tả; (2) phân rã chuỗi thời gian và kiểm định tính hợp lệ của chuỗi phân rã; (3) kiểm định tính dừng đối với chuỗi lợi suất logarit, tiếp nối bằng kiểm định tính tự tương quan làm cơ sở cho việc mô hình hóa với

ARIMA-GARCH và mô phỏng Monte Carlo. *Luồng còn lại* là luồng mô hình hóa biến mục tiêu (giá đóng cửa). Quy trình này tuân thủ trật tự tuyến tính chặt chẽ: bắt đầu từ tiền xử lý và tạo sinh biến mới, đi qua các bước kiểm định tính dừng và kiểm định tính tương quan để sàng lọc đặc trưng, cuối cùng đi vào mô hình hóa với XGBoost.

Kết quả đầu ra từ tất cả các nhánh phân tích trên (bao gồm thống kê mô tả, kết quả phân rã, mô hình ARIMA-GARCH kết hợp mô phỏng Monte Carlo và mô hình XGBoost) đều được hội tụ về bước cuối cùng để tổng hợp thành kết luận cuối cùng, từ đó tổng kết lại những điểm mạnh, điểm cần cải thiện và đề xuất phát triển.

4. PHÂN TÍCH ĐƠN BIẾN GIÁ ĐÓNG CỬA

Trong phần này, chúng tôi tập trung phân tích dữ liệu giá cổ phiếu FPT thông qua các phương pháp tiếp cận đa chiều, bao gồm: thống kê mô tả giá cổ phiếu và phân tích xu hướng chuỗi giá; mô hình hóa biến động lợi suất bằng phương pháp kết hợp ARIMA-GARCH và mô phỏng Monte Carlo.

4.1. Thực hiện thống kê mô tả

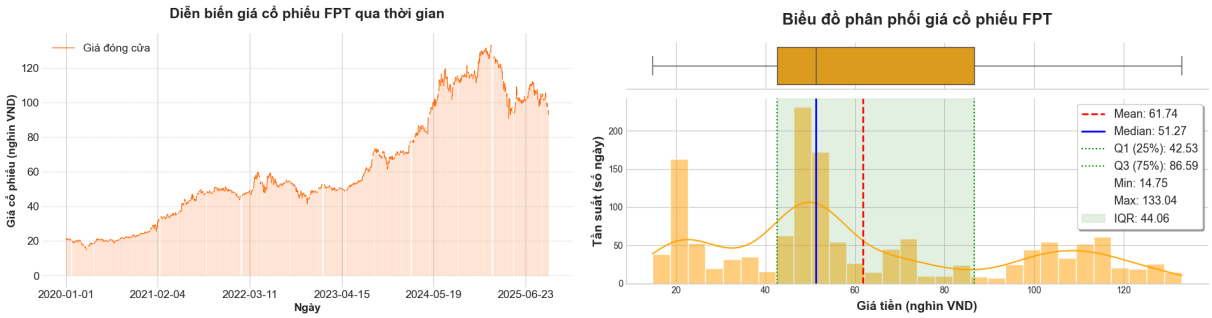
Dữ liệu được thu thập từ đầu năm 2020 đến tháng 09/2025, bao gồm tổng cộng 2100 quan sát với tần suất hàng ngày (Hình 6). Cột giá đóng cửa (close) ghi nhận 1433 ngày giao dịch thực tế sau khi đã loại trừ các ngày cuối tuần và ngày lễ theo đặc thù của thị trường chứng khoán. Toàn bộ dữ liệu được lưu trữ dưới dạng số thực (float64), đảm bảo tính chính xác và thuận tiện cho quá trình mô hình hóa tiếp theo.

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 2100 entries, 2020-01-01 to 2025-09-30
Freq: D
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   close   1433 non-null   float64
 1   volume  1433 non-null   float64
dtypes: float64(2)
memory usage: 49.2 KB
```

Hình 6: Mô tả thống kê biến giá đóng cửa cổ phiếu tập đoàn FPT

Thông qua việc kết hợp các kỹ thuật trực quan hóa và thống kê mô tả, nhóm đồ án đã xác lập được bức tranh toàn cảnh về cấu trúc dữ liệu của cổ phiếu FPT giai đoạn 2020 – 2025. Biểu đồ diễn biến thời gian tại Hình 7 (trái) cho thấy một quỹ đạo tăng trưởng dài hạn bền vững; bắt đầu giai đoạn nghiên cứu với mức giá quanh ngưỡng 20,000 đồng và **từng chạm đáy tại 14,750 đồng**, cổ phiếu này đã bứt phá mạnh mẽ từ đầu năm **2024 để thiết lập mức đỉnh lịch sử 133,040 đồng**.

Về đặc điểm phân phối, biểu đồ tần suất tại Hình 7 (phải) xác nhận chuỗi giá không tuân theo phân phối chuẩn mà có dạng đa đỉnh, phản ánh rõ nét các giai đoạn dịch chuyển của mặt bằng giá theo thời gian. Cụ thể, các đỉnh mật độ tập trung lần lượt xuất hiện tại các ngưỡng khoảng 20, 50 và 110 nghìn đồng, tương ứng với các cột mốc đầu năm 2020, cuối năm 2022 và từ giữa năm 2024 về sau. Bên cạnh đó, các chỉ số đo lường độ phân tán cho thấy khoảng biến thiên tứ phân vị (IQR) đạt mức 44,060 đồng, với giới hạn dưới Q1 là 42,530 đồng, trung vị (median) là 51,270 đồng và giới hạn trên Q3 là 86,590 đồng. Điều này minh chứng rằng 50% dữ liệu giá đóng cửa tập trung mật độ cao tại vùng giá trung bình thấp.



Hình 7: Diễn biến giá đóng cửa (trái) và phân phối tần suất giá cổ phiếu FPT (phải) giai đoạn 2020 – 2025 (đơn vị: nghìn VNĐ)

Tổng kết lại, khoảng biến thiên rộng giữa các giá trị tứ phân vị phối hợp cùng sự chênh lệch rất lớn giữa mức giá cao nhất và thấp nhất đã phản ánh biên độ dao động lớn của mã cổ phiếu FPT trong giai đoạn khảo sát. Đặc biệt, sự chuyển dịch qua ba mặt bằng giá với xu hướng tăng dần theo thời gian không chỉ minh chứng cho sự bứt phá về giá trị cổ phiếu mà còn cho thấy đặc tính không dừng của chuỗi dữ liệu gốc. Những đặc điểm về phân phối đa đỉnh và sự biến động mạnh này là những tín hiệu thực nghiệm quan trọng làm tiền đề để nhóm nghiên cứu tiến hành lấy sai phân và áp dụng các mô hình định lượng phức tạp nhằm bắt kịp các quy luật biến động trong các phần tiếp theo.

4.2. Phân tích xu hướng

Hình 8: Phân rã chuỗi thời gian đóng cửa

```
--- Kết quả kiểm định ADF cho chuỗi Phần dư (Residual) ---
H0: Chuỗi Phần dư (Residual) là chuỗi không dừng
H1: Chuỗi Phần dư (Residual) là chuỗi dừng
Giá trị thống kê ADF (ADF Statistic): -3.4735
Giá trị p (p-value): 0.0087
Số lượng độ trễ (Lags used): 0
Các giá trị giới hạn quan trọng (Critical Values):
   1%: -3.434125
   5%: -2.863207
   10%: -2.567658

Kết luận: Chuỗi Phần dư (Residual) là dừng.
```

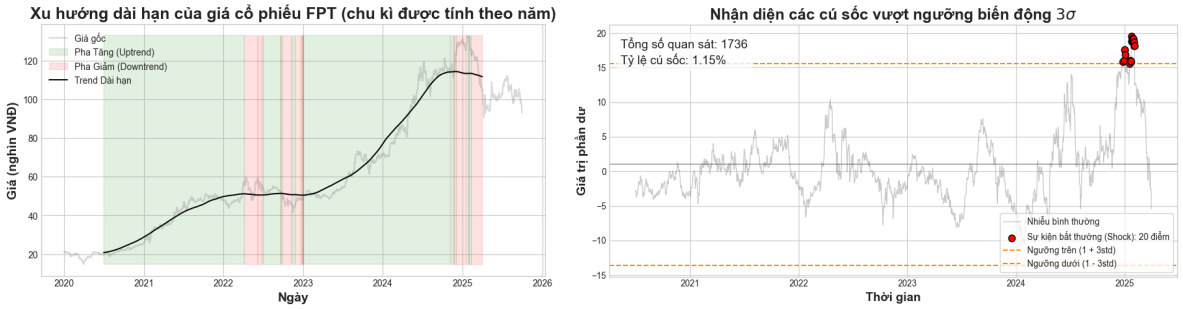
Hình 9: Kiểm định tính dừng đối với chuỗi phần dư

Nhằm khắc phục các biến động ngẫu nhiên ngắn hạn đặc trưng của dữ liệu gốc, chúng tôi áp dụng mô hình phân rã cộng (additive decomposition) với chu kỳ năm (365 ngày). Trước đó, các giá trị khuyết đã được xử lý bằng kỹ thuật bfill. Phương pháp này cho phép tách chuỗi giá đóng cửa thành ba thành phần độc lập: xu hướng, mùa vụ và nhiễu (xem Hình 8). Quá

trình này đóng vai trò quan trọng trong việc trích xuất các quy luật hệ thống ra khỏi dữ liệu, đảm bảo thành phần phần dư chỉ còn chứa đựng các biến động ngẫu nhiên thuần túy.

Để xác thực tính hợp lệ của mô hình phân rã, kết quả kiểm định Augmented Dickey-Fuller (ADF) trên chuỗi phần dư được trình bày tại Hình 9 cho thấy giá trị p-value=.0087 (nhỏ hơn mức ý nghĩa .01). Với kết quả này, giả thuyết H0 bị bác bỏ, khẳng định chuỗi phần dư đã đạt trạng thái dừng. Điều này minh chứng rằng các thành phần xu hướng và mùa vụ đã được bóc tách triệt để, tạo lập một nền tảng dữ liệu tin cậy để triển khai các phân tích chuyên sâu và xây dựng mô hình hóa dự báo ở các giai đoạn tiếp theo.

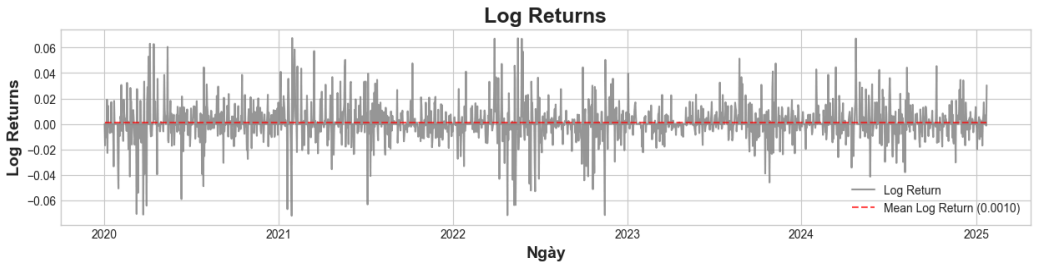
Sau khi xác thực tính hợp lệ của mô hình phân rã, nhóm tiến hành phân tích sâu về cấu trúc xu hướng và thành phần nhiễu của cổ phiếu FPT. Kết quả tại Hình 10 minh chứng cho một xu hướng tăng trưởng dài hạn vững chắc, trong đó các pha điều chỉnh giảm (màu đỏ) chỉ diễn ra ngắn hạn với biên độ thấp, không làm phá vỡ cấu trúc tăng giá chủ đạo (màu xanh). Việc định tính hóa phần dư qua ngưỡng 3std, khẳng định tính ổn định của chuỗi dữ liệu khi phần lớn biến động đều nằm trong phạm vi kiểm soát. Đáng chú ý, các "cú sốc" vượt ngưỡng (điểm đỏ) xuất hiện với tỷ lệ cực thấp, chỉ chiếm 1.15% tổng số quan sát khả thi và tập trung chủ yếu vào giai đoạn biến động mạnh cuối năm 2024 - đầu 2025. Tỷ lệ nhiễu cực đoan không đáng kể này là cơ sở quan trọng để đảm bảo độ tin cậy của dữ liệu trước khi đưa vào các mô hình dự báo chuyên sâu.



Hình 10: Chuỗi xu hướng giá và định tính hóa các cú sốc vượt ngưỡng 3std

4.3. Phân tích chuỗi lợi suất logarit và mô phỏng chuỗi biến động

Sau khi tạo ra biến chuỗi lợi suất logarit từ giá cổ phiếu FPT, nhóm đồ án đã tiến hành trực quan hóa dữ liệu để đánh giá các đặc điểm phân phối. Kết quả tại Hình 11 cho thấy rõ xu hướng co cụm biến động, trong đó các khoảng dao động biên độ lớn thường xuất hiện nối tiếp nhau. Hiện tượng này là cơ sở dẫn đến giả thuyết trong lĩnh vực tài chính rằng phương sai của chuỗi không hằng định mà thay đổi theo thời gian (phương sai có điều kiện).



Hình 11: Hiện tượng kết cụm của chuỗi lợi suất logarit

Trước khi xây dựng mô hình, các kiểm định thống kê đã được thực hiện để xác định cấu trúc phù hợp. Kiểm định Ljung-Box trên chuỗi lợi suất logarit cho thấy các giá trị p-value đều lớn hơn 0.05 ở mọi độ trễ, xác nhận chuỗi này là "nhiễu trắng" (ngẫu nhiên), do đó mô hình ARIMA(0, 0, 0) (mô hình trung bình hằng số) được lựa chọn là đủ để mô tả thành phần trung

bình (Hình 12 trái). Tuy nhiên, khi kiểm định trên bình phương lợi suất, các kết quả cho thấy sự tự tương quan rất mạnh (p-value < .0001, Hình 12 phải). Điều này chứng minh sự tồn tại của hiệu ứng ARCH, tức hiện tượng các giai đoạn biến động mạnh đi kèm với nhau, làm cơ sở vững chắc cho việc áp dụng mô hình GARCH.

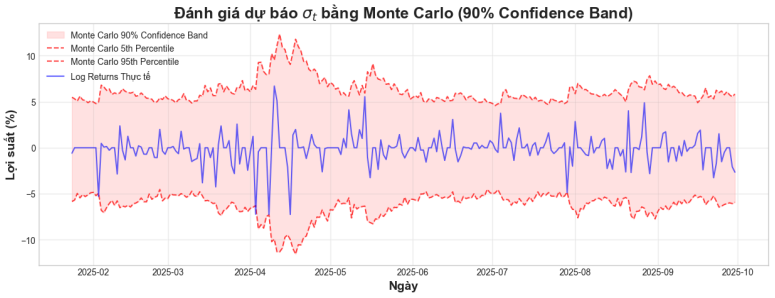
```
Kiểm định Ljung-Box cho chuỗi lợi suất
============================================================
KẾT QUẢ KIỂM ĐỊNH LJUNG-BOX (VỚI LAG = [1, 2, 3, 4, 5, 10])
============================================================
    Test Statistic   p-value Kết luận (H0)
1         1.107832  0.292554    Ngẫu nhiên
2         3.516296  0.172364    Ngẫu nhiên
3         6.449124  0.091691    Ngẫu nhiên
4         6.463932  0.167077    Ngẫu nhiên
5         8.222437  0.144395    Ngẫu nhiên
10       11.775777  0.300343    Ngẫu nhiên
------------------------------------------------------------
TỔNG KẾT: Chuỗi dữ liệu là NGẪU NHIÊN (White Noise) ở mọi độ trễ kiểm tra.
============================================================
```

```
Kiểm định Ljung-Box cho bình phương chuỗi lợi suất
============================================================
KẾT QUẢ KIỂM ĐỊNH LJUNG-BOX (VỚI LAG = [1, 2, 3, 4, 5, 10])
============================================================
    Test Statistic       p-value  Kết luận (H0)
1        40.578392  1.888839e-10  Có tương quan
2        44.052734  2.716880e-10  Có tương quan
3        54.866344  7.332086e-12  Có tương quan
4        57.008476  1.232115e-11  Có tương quan
5        60.008216  1.210714e-11  Có tương quan
10      150.960883  2.364309e-27  Có tương quan
------------------------------------------------------------
TỔNG KẾT: Tồn tại sự tự tương quan ở một số độ trễ nhất định.
============================================================
```

Hình 12: Kiểm định LJung-Box cho chuỗi lợi suất (trái) và bình phương chuỗi lợi suất (phải)

Trong đồ án, chúng tôi chọn mô hình GARCH(1, 1) với thiết lập ***phân phối Student's t*** để xử lý đặc tính đuôi nặng (fat tails) của dữ liệu tài chính, nơi các biến động cực đoan xảy ra thường xuyên hơn phân phối chuẩn. Kết quả cho thấy hệ số alpha (phản ứng với cú sốc mới) là 0.1185 và hệ số beta (tính bền vững của biến động quá khứ) là 0.8815. Tổng hai hệ số này xấp xỉ bằng 1 cho thấy biến động của FPT có tính "nhớ" rất lâu, cụ thể là một cú sốc trên thị trường sẽ mất nhiều thời gian để ổn định trở lại. Theo tính toán, độ biến động (STD) trung bình hàng ngày của FPT đạt mức 2.12%, với những giai đoạn cao điểm lên tới 4.29%.

Để đánh giá khả năng dự báo rủi ro, một quy trình mô phỏng Monte Carlo với 3000 kịch bản đã được thực hiện dựa trên các tham số của mô hình GARCH. Kết quả tạo ra một dải băng tin cậy 90% cho lợi suất của 250 ngày giao dịch cuối cùng. Khi so sánh với dữ liệu thực tế, hầu hết các biến động giá của FPT đều nằm gọn trong dải băng này (Hình 13), chứng minh rằng mô hình GARCH(1, 1) đã nắm bắt rất tốt biên độ rủi ro của cổ phiếu. Việc sử dụng tham số nu xấp xỉ 2.16 (bậc tự do của phân phối t) giúp mô hình cảnh báo chính xác hơn về các rủi ro sụt giảm mạnh mà các mô hình truyền thống thường bỏ qua.



Hình 13: Mô phỏng Monte Carlo cho biến động chuỗi lợi suất logarit

5. PHÂN TÍCH BIẾN MỤC TIÊU

5.1. Tiền xử lý dữ liệu và tạo biến mới

Nhằm đảm bảo tính đồng nhất và tối ưu hóa hiệu suất cho các mô hình học máy, nhóm nghiên cứu đã thiết lập một quy trình tiền xử lý tự động hóa chia các biến đầu vào thành ba nhóm đặc thù dựa trên tính chất phân phối và ý nghĩa kinh tế của chúng.

Đối với nhóm các biến tài chính có xu hướng tăng trưởng theo thời gian như giá cổ phiếu, vốn hóa thị trường, tỷ giá và các chỉ số tài chính doanh nghiệp (doanh thu, lợi nhuận), chúng tôi áp dụng kỹ thuật biến đổi logarit lợi suất. Phương pháp này chuyển đổi chuỗi giá trị tuyệt đối sang chuỗi tỷ suất sinh lời liên tục bằng công thức

$$r_i = \ln\left(\frac{P_{i,t}}{P_{i,t-1}}\right)$$

nhằm khử bỏ xu hướng (trend) và ổn định phương sai, đưa dữ liệu về trạng thái dừng để phù hợp với giả định của các thuật toán thống kê.

Cuối cùng, đối với nhóm các biến vĩ mô đã ở dạng tỷ lệ phần trăm (như CPI), quy trình chỉ thực hiện chuẩn hóa trực tiếp để đồng bộ hóa quy mô với các biến khác trong tập dữ liệu.

5.2. Chọn biến huấn luyện

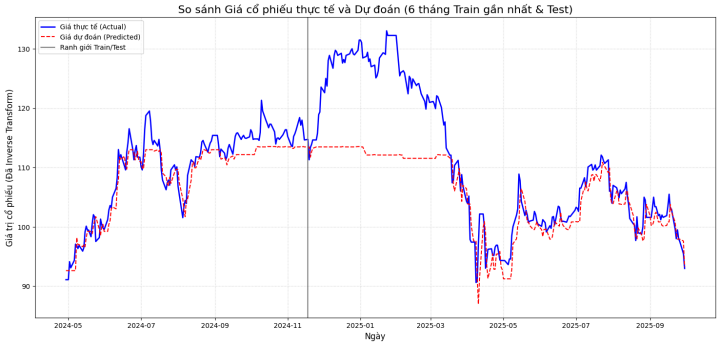
Để xử lý hiện tượng đa cộng tuyến, chúng tôi sàng lọc biến dựa trên ma trận Pearson và kiểm định p-value với ngưỡng |r| > 0.50. Phân tích cho thấy gdp_value tương quan cực mạnh (r = 0.96) với nhóm chỉ số tài chính FPT (doanh thu, lợi nhuận), nên nhóm quyết định giữ lại gdp_value và loại bỏ toàn bộ các biến tài chính trên để tránh làm mô hình mất ổn định do phương sai tăng cao. Đồng thời, biến usd_vnd_rate cũng bị lược bỏ do tương quan đáng kể (r=0.58) với fpt_net_revenue. Quá trình giảm chiều dữ liệu này đảm bảo các biến độc lập còn lại (như gdp_value, cpi_rate) cung cấp thông tin riêng biệt, tối ưu hóa khả năng tổng quát hóa của mô hình.

Dưới góc độ kinh doanh, sự tương quan cao (r > 0.95) giữa GDP và các chỉ số tài chính phản ánh sự đồng pha chặt chẽ giữa đà tăng trưởng của FPT và chu kỳ kinh tế vĩ mô. Do đó, biến gdp_value được chọn làm đại diện bao trùm thay thế cho nhóm biến nội bộ. Đồng thời, biến tỷ giá (usd_vnd_rate) cũng bị loại bỏ vì tác động của nó đã được phản ánh trực tiếp trong doanh thu xuất khẩu. Việc sàng lọc này giúp triệt tiêu thông tin thừa, đảm bảo mô hình tập trung vào các động lực tăng trưởng cốt lõi.

Cuối cùng, chúng tôi giữ lại 8 trường dữ liệu chính, bao gồm 'fpt_stock_price_log_return_scaled', 'market_cap_log_return', 'fpt_stock_price_log_return', 'gdp_value', 'cpi_rate', 'fpt_stock_price', 'fpt_stock_volume_log_return', 'fpt_net_revenue_log_return', 'usd_vnd_rate_log_return', 'xau_usd_rate_log_return'.

5.3. Huấn luyện và dự đoán

Mô hình học máy XGBoost được huấn luyện trên 85% dữ liệu và kiểm thử trên 15% dữ liệu còn lại với cấu hình 5000 cây quyết định và tốc độ học 0.01. Các chỉ số đo lường sai số trên tập kiểm thử cho kết quả cụ thể như sau: sai số tuyệt đối trung bình (MAE) là 5.95 nghìn VND; sai số bình phương trung bình căn (RMSE) là 8.53 nghìn VND và hệ số xác định (R2) đạt 0.4976. Biểu đồ so sánh giữa giá trị thực tế và giá trị dự báo cho thấy mô hình bám sát được các xu hướng lớn nhưng có xu hướng bị lệch tại các điểm có biến động cực đại đột ngột.



Hình 11: Mô hình XGBoost dự đoán phần test

Sự phân hóa về độ chính xác giữa các giai đoạn dự báo phản ánh hai khía cạnh của mô hình. Một mặt, việc nắm bắt thành công xu hướng chủ đạo đã khẳng định tính hợp lý và cơ sở

khoa học của tập biến đầu vào. Mặt khác, các sai lệch cục bộ cho thấy độ nhạy của thuật toán còn hạn chế trước những biến động phức tạp, hoặc thị trường đang chịu sự chi phối của các yếu tố ngoại sinh nằm ngoài phạm vi dữ liệu quan sát.

6. KẾT QUẢ PHÂN TÍCH

Quá trình thực hiện đồ án đã đem lại những kết quả định lượng cụ thể về diễn biến giá cổ phiếu FPT và các yếu tố tác động, cụ thể như dưới đây

6.1. Phân tích cấu trúc nội tại và xu hướng

Dựa trên dữ liệu của 1,433 phiên giao dịch thực tế từ năm 2020 đến tháng 09/2025, bức tranh toàn cảnh về diễn biến giá và cấu trúc rủi ro của cổ phiếu FPT đã được thiết lập qua hai khía cạnh chính: xu hướng dài hạn và đặc tính biến động.

Thứ nhất, dữ liệu lịch sử khẳng định xu hướng tăng trưởng bền vững của FPT với cấu trúc xu hướng được tách biệt rõ ràng, trong đó phần nhiễu hoàn toàn dừng và tỷ lệ biến động bất thường cực thấp (1.15%). Điều này khẳng định các quy luật vận động của giá là có thật và ít bị bóp méo bởi các cú sốc ngẫu nhiên, tạo tiền đề lý tưởng cho độ chính xác của các mô hình dự báo phía sau.

Thứ hai, kết quả phân tích cũng chỉ ra rằng nơi các nhịp tăng giá đóng vai trò chủ đạo lấn át các pha điều chỉnh ngắn hạn. Sự xuất hiện của dạng phân phối đa đỉnh tăng dần theo thời gian (tại các mốc 20, 50 và 110 nghìn đồng) thay vì phân phối chuẩn cho thấy cổ phiếu liên tục phá vỡ các vùng cân bằng cũ để thiết lập mặt bằng giá mới cao hơn. Điều này phản ánh nội lực tăng trưởng về vốn hóa doanh nghiệp theo thời gian.

Thứ ba, kết quả định lượng xác nhận rủi ro của cổ phiếu FPT có tính 'trí nhớ' dai dẳng: các cú sốc thị trường không tan biến ngay mà tạo ra dư âm kéo dài, đòi hỏi chiến lược quản trị rủi ro phải có tầm nhìn dài hạn. Một phát hiện đắt giá là trong khi hướng giá (tăng/giảm) mang tính ngẫu nhiên khó đoán thì cường độ biến động lại hoàn toàn có thể dự báo. Điều này trao cho nhà đầu tư một lợi thế lớn: dù khó đoán định mức giá cụ thể, ta hoàn toàn có thể tiên lượng được thị trường đang trong trạng thái bình ổn hay bước vào giai đoạn biến động cực đoan để điều chỉnh tỷ trọng danh mục phù hợp.

6.2. Phân tích đa biến với các biến thị trường

Thứ nhất, sức khỏe tài chính của FPT gắn chặt với chu kỳ kinh tế vĩ mô (ở đây là nền kinh tế Việt Nam). Bằng chứng là ma trận tương quan Pearson chỉ ra hệ số cực cao ($r \approx 0.96$) giữa GDP và các chỉ số nội tại của FPT (doanh thu, lợi nhuận). Điều này cho phép chúng ta đơn giản hóa mô hình bằng cách dùng GDP làm "biến đại diện" duy nhất mà vẫn bao quát được động lực tăng trưởng của doanh nghiệp.

Thứ hai, việc chuyển đổi dữ liệu sang dạng logarit lợi suất thay vì giá trị tuyệt đối là chiến lược giúp mô hình học được "cấu trúc sinh lời" thay vì chỉ "ghi nhớ mặt bằng giá". Cách tiếp cận này đưa dữ liệu về trạng thái dừng, đảm bảo thuật toán nhận diện đúng quy luật tăng/giảm theo phần trăm bất kể thị giá đang ở mức 20,000 đồng hay 130,000 đồng. Nhờ đó, mô hình tránh được việc dự báo sai lệch do sự thay đổi về quy mô giá trong dài hạn.

Thứ ba, con số $R2 \approx 50\%$ cho thấy trong khi khả năng bám sát xu hướng dài hạn xác thực độ tin cậy của tập biến đầu vào thì các sai số cục bộ lại bộc lộ những giới hạn nhất định. Cụ thể, sự chênh lệch này cho thấy thuật toán chưa đủ độ nhạy trước các biến động phức tạp, hoặc thị trường đang chịu sự chi phối của các yếu tố ngẫu nhiên và biến số ngoại sinh nằm ngoài phạm vi dữ liệu quan sát.

7. KẾT LUẬN

Đồ án đã hoàn thành mục tiêu xây dựng một khung phân tích định lượng toàn diện đối với cổ phiếu Tập đoàn FPT trong giai đoạn từ năm 2020 đến tháng 09/2025, tuân thủ chặt chẽ quy trình chuẩn từ khâu thu thập, xử lý dữ liệu đa nguồn đến việc áp dụng các mô hình thống kê và học máy phức tạp. Thông qua việc tổng hợp dữ liệu từ các báo cáo tài chính, chỉ số vĩ mô và dữ liệu giao dịch thị trường, nhóm nghiên cứu đã thiết lập được một cơ sở dữ liệu đồng nhất, đảm bảo tính khoa học cho các bước kiểm định tiếp theo.

Về kết quả nghiên cứu, đồ án khẳng định FPT sở hữu cấu trúc tăng trưởng bền vững với tỷ lệ nhiễu cực thấp (1.15%) và gắn kết chặt chẽ với chu kỳ kinh tế vĩ mô và cường độ biến động. Trong khi mô hình ARIMA-GARCH lượng hóa thành công đặc tính rủi ro "dai dẳng", thì XGBoost (với R2 ≈ 50%) đã chứng minh vai trò hiệu quả trong việc định hướng xu hướng dài hạn, bất chấp các hạn chế về dự báo cú sốc tâm lý ngắn hạn.

Bên cạnh những kết quả khả quan, nghiên cứu vẫn còn tồn tại những giới hạn nhất định phản ánh tính phức tạp của thị trường tài chính. Mặc dù mô hình XGBoost bám sát tốt xu hướng dài hạn, nhưng kết quả thực nghiệm cho thấy sự lệch pha rõ rệt tại các điểm biến động cực đại, nơi đường dự báo thường không bắt kịp biên độ của các cú sốc giá ngắn hạn. Hạn chế này chỉ ra rằng tập dữ liệu hiện tại, dù đã bao gồm các yếu tố vĩ mô và tài chính, vẫn chưa đủ để giải thích toàn bộ sự biến thiên của giá cổ phiếu. Khoảng trống thông tin còn lại ("vùng xám" mà mô hình chưa nắm bắt được) nhiều khả năng chịu sự chi phối mạnh mẽ của các yếu tố phi cấu trúc như tâm lý đám đông, tin tức đột ngột hay các dòng tiền đầu cơ, tức những biến số ngoại sinh chưa được lượng hóa trong phạm vi đồ án này.

Từ những phân tích trên, nhóm nghiên cứu đề xuất các kiến nghị cụ thể trên cả hai phương diện đầu tư và phát triển mô hình. Đối với nhà đầu tư, kết quả định lượng gợi ý chiến lược tập trung vào quản trị xu hướng dài hạn thay vì cố gắng giao dịch tần suất cao tại các điểm đảo chiều, đồng thời cần tận dụng tín hiệu từ mô hình GARCH để chủ động phòng vệ khi thị trường bước vào chu kỳ co cụm biến động. Về hướng phát triển kỹ thuật, để khắc phục hạn chế trong việc dự báo các cú sốc ngắn hạn, các nghiên cứu tiếp theo cần mở rộng phạm vi dữ liệu bằng cách tích hợp thêm phân tích cảm xúc (Sentiment Analysis) từ tin tức và mạng xã hội thông qua kỹ thuật xử lý ngôn ngữ tự nhiên (NLP). Song song đó, việc thử nghiệm các kiến trúc học sâu tiên tiến hơn như Transformer, vốn có ưu thế trong việc xử lý các chuỗi phi tuyến phức tạp, sẽ là bước đi cần thiết để nâng cao độ nhạy của mô hình, giúp hệ thống không chỉ nắm bắt được xu hướng lớn mà còn thích ứng tốt hơn với những biến động bất thường của thị trường.

TÀI LIỆU THAM KHẢO

[1] (Tài liệu trên internet) USD/VND - Đô la Mỹ Đồng Việt Nam. Link: https://vn.investing.com/currencies/usd-vnd-historical-data (23/10/2025).

[2] (Tài liệu trên internet) Chỉ số giá tiêu dùng (CPI) Việt Nam. Link: https://vn.investing.com/economic-calendar/vietnamese-cpi-1851 (23/10/2025).

[3] (Tài liệu trên internet) GDP (current US$) - Vietnam | Data. Link: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=VN (23/10/2025).

[4] (Tài liệu trên internet) Real GDP Growth - Vietnam. Link: https://www.imf.org/external/datamapper/profile/VNM (23/10/2025).

[5] (Tài liệu trên internet) XAU/USD - Vàng Đô la Mỹ. Link: https://vn.investing.com/currencies/xau-usd-historical-data (23/10/2025).

[6] (Tài liệu trên internet) Tech Industry in Vietnam - Market Analysis. Link: https://simplywall.st/markets/vn/tech (23/10/2025).

[7] (Tài liệu trên internet) Dữ liệu tài chính doanh nghiệp FPT qua thư viện vnstock API. Link: https://github.com/thinh-vu/vnstock (23/10/2025).

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Hồ Tấn Dũng	- Xây dựng kịch bản phân tích - EDA - Phân tích thống kê phần phân tích đơn biến - Tiền xử lý dữ liệu cho phần mô hình hóa XGBoost - Soạn slide - Làm dashboard - Hỗ trợ làm báo cáo
2	Nguyễn Hoàng Long	- Xây dựng kịch bản phân tích - Thu thập dữ liệu - EDA - Xây dựng mô hình ARIMA-GARCH, mô phỏng Monte Carlo, mô hình XGBoost - Viết báo cáo - Hỗ trợ soạn slide