ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN – CƠ – TIN HỌC

BÁO CÁO TIỂU LUẬN

HỌC PHẦN: PHÂN TÍCH CHUỖI THỜI GIAN

APPLE STOCK PRICE PREDICTION

Giảng viên:	TS. Hoàng Thị Phương Thảo
Ngành:	Khoa học dữ liệu
Khóa:	QH.2022.T
Mã lớp:	MAT3388
Nhóm:	11
Sinh viên thực hiện:	Nguyễn Tất Huấn
	Nguyễn Văn Khải
	Nguyễn Văn Chiến

Hà Nội, 2025

Giới thiệu thành viên nhóm

Thành viên	Mã sinh viên	Công việc
Nguyễn Văn Chiến	22001238	Phân tích và tiền xử lý dữ liệu
Nguyễn Tất Huấn	22001261	Tìm hiểu và thực hiện mô hình SARIMA
Nguyễn Văn Khải	22001264	Tìm hiểu và thực hiện mô hình Prophet

Mục lục

1 Lời Mở Đầu

Trong kỷ nguyên của công nghệ thông tin và dữ liệu lớn (Big Data), lĩnh vực tài chính và đầu tư chứng khoán đang dần chuyển mình để thích ứng với tốc độ phát triển vượt bậc của khoa học dữ liệu. Việc ứng dụng các kỹ thuật phân tích dữ liệu hiện đại, đặc biệt là phân tích chuỗi thời gian (Time Series Analysis), ngày càng trở nên phổ biến và cần thiết trong quá trình ra quyết định đầu tư. Với khả năng khai thác các quy luật tiềm ẩn trong dữ liệu lịch sử để dự đoán xu hướng tương lai, phân tích chuỗi thời gian trở thành một công cụ mạnh mẽ giúp nhà đầu tư đánh giá rủi ro và nắm bắt cơ hội trên thị trường tài chính.

Trong số các mã cổ phiếu đang được quan tâm, cổ phiếu của **Apple Inc.** – một trong những tập đoàn công nghệ hàng đầu thế giới – là đối tượng lý tưởng để áp dụng các phương pháp phân tích chuỗi thời gian. Apple không chỉ nổi bật bởi các sản phẩm mang tính biểu tượng như iPhone, iPad, MacBook, mà còn nhờ vào chiến lược phát triển bền vững, tăng trưởng doanh thu ổn định và sức ảnh hưởng toàn cầu. Những yếu tố này khiến cho cổ phiếu Apple thể hiện rõ các đặc điểm của một chuỗi thời gian điển hình: xu hướng tăng trưởng dài hạn, biến động ngắn hạn mang tính mùa vụ (đặc biệt là vào quý IV hằng năm), cùng với sự nhạy cảm trước các sự kiện kinh tế – xã hội lớn như ra mắt sản phẩm mới, khủng hoảng tài chính hay đại dịch toàn cầu.

Việc lựa chọn cổ phiếu Apple làm đối tượng nghiên cứu vì thế không chỉ mang tính thực tiễn cao, mà còn tạo điều kiện thuận lợi để kiểm chứng hiệu quả của các mô hình dự báo trong điều kiện thị trường thực tế. Trong bối cảnh mà mỗi quyết định đầu tư đòi hỏi sự hỗ trợ từ dữ liệu và mô hình phân tích chính xác, việc kết hợp giữa kiến thức tài chính và các kỹ thuật phân tích chuỗi thời gian mang đến một cách tiếp cận khoa học, hệ thống và hiệu quả cho các nhà đầu tư hiện đại.

Nghiên cứu này được thực hiện nhằm phân tích và dự báo giá cổ phiếu Apple thông qua việc ứng dụng các mô hình chuỗi thời gian hiện đại, từ đó góp phần hỗ trợ quá trình ra quyết định đầu tư một cách khách quan, khoa học và dựa trên cơ sở dữ liệu thực tế.

2 Giới Thiệu Chung

2.1 Giới thiệu đề tài

Dự báo giá cổ phiếu là một trong những thách thức lớn trong lĩnh vực tài chính và khoa học dữ liệu. Sự biến động liên tục của thị trường chứng khoán dưới tác động của các yếu tố kinh tế vĩ mô, chính sách tiền tệ, hoạt động doanh nghiệp, và cả tâm lý đám đông, khiến việc dự báo trở nên khó khăn nhưng lại vô cùng cần thiết. Các nhà đầu tư, nhà quản lý danh mục và các tổ chức tài chính đều có nhu cầu cấp thiết trong việc dự đoán xu hướng giá để ra quyết định mua, bán hay nắm giữ tài sản một cách hiệu quả và chính xác.

Trong bối cảnh đó, phân tích chuỗi thời gian (Time Series Analysis) đã trở thành một công cụ quan trọng và được ứng dụng rộng rãi trong dự báo tài chính. Phân tích chuỗi thời gian cho phép mô hình hóa dữ liệu theo thời gian và đưa ra dự báo dựa trên những mẫu (pattern) lịch sử như xu hướng (trend), mùa vụ (seasonality), chu kỳ (cycle) và nhiễu loạn (noise). Khi áp dụng đúng cách, các mô hình chuỗi thời gian không chỉ giúp giải thích sự thay đổi trong dữ liệu quá khứ mà còn cung cấp những dự đoán có giá trị thực tiễn về tương lai.

Trong báo cáo này, đối tượng được lựa chọn để phân tích và dự báo là cổ phiếu của công ty **Apple Inc.** (mã chứng khoán: AAPL) – một trong những công ty dẫn đầu thế giới trong lĩnh vực công nghệ. Apple là một lựa chọn lý tưởng vì:

- Là công ty có mức độ phổ biến toàn cầu và sức ảnh hưởng mạnh mẽ đến thị trường.
- Giá cổ phiếu của Apple thường thể hiện các yếu tố đặc trưng của chuỗi thời gian như xu hướng tăng dài hạn, mùa vụ cao điểm vào cuối năm (gắn liền với chu kỳ ra mắt sản phẩm mới và kỳ nghỉ lễ), và phản ứng với các sự kiện kinh tế lớn.
- Dữ liệu cổ phiếu Apple có sẵn, đáng tin cậy và kéo dài trong nhiều năm, tạo điều kiện thuận lợi cho việc áp dụng và kiểm chứng các mô hình dự báo.

Để thực hiện dự báo giá cổ phiếu Apple, nghiên cứu này tập trung vào hai mô hình chuỗi thời gian phổ biến và hiệu quả: **SARIMA** (Seasonal AutoRegressive Integrated Moving Average) và **Prophet** – mô hình được phát triển bởi Facebook, nổi bật nhờ khả năng xử lý dữ liệu có tính mùa vụ mạnh, tính linh hoạt và dễ sử dụng.

Quá trình nghiên cứu bao gồm các bước: thu thập và tiền xử lý dữ liệu giá cổ phiếu AAPL, phân tích đặc điểm chuỗi thời gian, xây dựng và hiệu chỉnh mô hình **SARIMA** và **Prophet**, đánh giá hiệu quả dự báo thông qua các chỉ số lỗi (RMSE, MAE, MAPE), và so sánh kết quả giữa hai mô hình.

2.2 Mục tiêu của dự án

Mục tiêu của dự án là nghiên cứu, hiểu và áp dụng các phương pháp mô hình hóa chuỗi thời gian lên bộ dữ liệu giá cổ phiếu của Apple. Một số vấn đề đặt ra:

- Cách để tiền xỷ lý dữ liệu: xử lý dữ liệu trống, xủ lý dữ liệu ngoại lai, ...
- Phương pháp chọn mô hình và mô hình phù hợp cho dữ liệu.
- Nhận xét và đánh giá mô hình.
- Hướng cải tiến của dự án trong tương lai.

Như vậy để thực hiện theo đúng mục tiêu của dự án cần xác định một số công việc phải giải quyết như sau:

- Tìm kiếm và đánh giá các kỹ thuật tiền xử lý dữ liệu. Áp dụng những kỹ thuật phù hợp lên dữ liệu.
- Tìm hiểu các phương pháp để mô hình hóa dữ liệu dạng chuỗi thời gian.
- Triển khai, so sánh và đánh giá các mô hình để tìm ra mô hình phù hợp.
- Sử dụng mô hình để dự báo tương lai.

3 Tập Dữ Liệu

Tập dữ liệu có tên là **AAPL.csv**, đại diện cho lịch sử giao dịch cổ phiếu của công ty **Apple Inc.**, được thu thập từ trang Yahoo Finance. Đây là một trong những nguồn dữ liệu phổ biến, cung cấp thông tin tài chính phục vụ cho các nghiên cứu và phân tích thị trường chứng khoán.

3.1 Cấu trúc tập dữ liệu

Tập dữ liệu bao gồm các bản ghi được ghi nhận theo **tần suất hàng ngày**, trải dài từ **tháng 1 năm 2012 đến tháng 12 năm 2019**, chỉ bao gồm các ngày giao dịch trên thị trường chứng khoán (không bao gồm cuối tuần và ngày nghỉ lễ), với tổng cộng **2011 quan sát** và 7 trường dữ liệu. Mỗi quan sát thể hiện thông tin chi tiết về giao dịch cổ phiếu của Apple trong một ngày cụ thể.

Các trường dữ liệu trong tập dữ liệu bao gồm:

- **Date**: Ngày giao dịch.
- **Open**: Giá mở cửa của cổ phiếu vào đầu phiên giao dịch.

- **High**: Giá cao nhất trong phiên giao dịch.
- **Low**: Giá thấp nhất trong phiên giao dịch.
- **Close**: Giá đóng cửa vào cuối phiên giao dịch.
- **Adj Close**: Giá đóng cửa đã điều chỉnh theo cổ tức hoặc chia tách cổ phiếu.
- **Volume**: Khối lượng cổ phiếu được giao dịch trong ngày.

Bảng 1 trình bày một ví dụ gồm 5 quan sát đầu tiên của tập dữ liệu AAPL, từ ngày 03/01/2012 đến 09/01/2012.

Bảng 1: Ví dụ mô tả dữ liệu AAPL (5 quan sát đầu tiên)

Date	Open	High	Low	Close	Adj Close	Volume
03/01/2012	58.485714	58.928570	58.428570	58.747143	50.765709	75555200
04/01/2012	58.571430	59.240002	58.468571	59.062859	51.038536	65005500
05/01/2012	59.278572	59.792858	58.952858	59.718571	51.605175	67817400
06/01/2012	59.967144	60.392857	59.888573	60.342857	52.144630	79573200
09/01/2012	60.785713	61.107143	60.192856	60.247143	52.061932	98506100

3.2 Thống kê mô tả

Để có cái nhìn tổng quan về tập dữ liệu, các thống kê mô tả được trình bày trong Bảng 2. Các giá trị này bao gồm số lượng quan sát, giá trị trung bình, độ lệch chuẩn, trung vị, các phân vị 25% và 75%, cũng như giá trị nhỏ nhất và lớn nhất của từng trường dữ liệu.

Bảng 2: Thống kê mô tả của tập dữ liệu AAPL

	Open	High	Low	Close	Adj Close	Volume
Count	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000
Mean	126.707469	127.827594	125.580258	126.741235	119.505548	5.949670e+07
Std	50.483753	50.926301	50.124940	50.578369	52.438444	4.683856e+07
25%	85.882858	86.717858	85.056427	86.202145	75.056679	2.758565e+07
50%	113.050003	114.190002	111.870003	113.050003	105.222908	4.346900e+07
75%	165.190002	167.409996	163.424995	165.245002	160.047111	7.471030e+07
Min	55.424286	57.085712	55.014286	55.790001	48.921928	1.136200e+07
Max	291.119995	293.970001	288.119995	291.519989	289.522614	3.765300e+08

Từ bảng thống kê mô tả, có thể thấy giá cổ phiếu AAPL dao động mạnh trong khoảng thời gian nghiên cứu. Giá đóng cửa điều chỉnh trung bình là 119.505548, với giá trị nhỏ nhất là 48.921928 và giá trị lớn nhất là 289.522614. Khối lượng giao dịch trung bình đạt khoảng 59.50 triệu cổ phiếu mỗi ngày, với độ lệch chuẩn khá lớn (46.84 triệu), cho thấy sự biến động đáng kể trong hoạt động giao dịch.

3.3 Phân tích sơ bộ chuỗi thời gian

Để hiểu rõ hơn về đặc điểm của dữ liệu, chúng tôi vẽ biểu đồ giá đóng cửa điều chỉnh (Adj Close) theo thời gian, được trình bày trong Hình 1. Biểu đồ này biểu diễn **Adj Close** của AAPL từ 03/01/2012 đến 30/12/2019 giúp quan sát xu hướng và biến động của giá cổ phiếu AAPL trong khoảng thời gian nghiên cứu.

Hình 1: Biểu đồ giá đóng cửa điều chỉnh của AAPL từ 03/01/2012 đến 30/12/2019

Từ Hình 1, có thể thấy giá cổ phiếu AAPL có xu hướng tăng dài hạn từ năm 2012 đến năm 2019, mặc dù có những giai đoạn biến động mạnh, đặc biệt vào khoảng năm 2015 và 2018. Xu hướng tăng này phản ánh sự tăng trưởng ổn định của Apple trong giai đoạn này, nhưng cũng cho thấy sự nhạy cảm của giá cổ phiếu với các yếu tố thị trường.

3.4 Kết luận sơ bộ

Dựa trên thống kê mô tả và biểu đồ giá đóng cửa điều chỉnh, chúng tôi nhận thấy giá cổ phiếu AAPL có xu hướng tăng dài hạn, nhưng cũng có những giai đoạn biến động mạnh, đặc biệt vào các năm 2015 và 2018. Khối lượng giao dịch có sự biến động lớn, có thể phản ánh sự thay đổi trong tâm lý thị trường. Trong các phần tiếp theo, chúng tôi sẽ tiến hành phân tích sâu hơn để kiểm tra tính dừng, mô hình hóa chuỗi thời gian, và đưa ra dự báo giá cổ phiếu.

4 Tiền Xử Lý Dữ Liệu

Trước khi tiến hành phân tích chuỗi thời gian, việc tiền xử lý dữ liệu là cần thiết để đảm bảo chất lượng và tính toàn vẹn của tập dữ liệu. Chúng tôi đã thực hiện các bước sau

- **Chuyển đổi cột Date thành Date Time Index:** Đảm bảo cột Date được định dạng đúng và sử dụng làm chỉ số thời gian.
- **Kiểm tra dữ liệu trống:** Xác định xem có giá trị trống (missing values) trong các cột dữ liệu hay không.
- **Loại bỏ các đặc trưng dư thừa:** Loại bỏ các cột `Open`, `High`, `Low`, `Close`, và `Volume` do tính đa cộng tuyến và mức độ tương quan thấp với biến mục tiêu `Adj Close`.
- **Tái lấy mẫu dữ liệu thành tần suất hàng tháng:** Chuyển đổi dữ liệu từ tần suất hàng ngày sang tần suất hàng tháng để phù hợp với mục tiêu phân tích.

4.1 Chuyển đổi cột Date thành Date Time Index

Trong tập dữ liệu **AAPL.csv**, cột **Date** ban đầu được lưu dưới dạng chuỗi (string) với kiểu dữ liệu `object` và định dạng `DD/MM/YYYY` (ví dụ: `03/01/2012`). Để phù hợp với phân tích chuỗi thời gian, chúng tôi chuyển đổi cột **Date** thành định dạng `datetime` và đặt nó làm chỉ số (index) cho tập dữ liệu.

Bảng 3 thể hiện 1 phần dữ liệu sau khi đổi cột Date thành Date Time Index

Open	High	Low	Close	Adj Close	Volume	Date
58.485714	58.928570	58.428570	58.747143	50.765709	75555200	2012-01-03
58.571430	59.240002	58.468571	59.062859	51.038536	65005500	2012-01-04
59.278572	59.792858	58.952858	59.718571	51.605175	67817400	2012-01-05
59.967144	60.392857	59.888573	60.342857	52.144630	79573200	2012-01-06
60.785713	61.107143	60.192856	60.247143	52.061932	98506100	2012-01-09

Bảng 3: Một phần dữ liệu sau khi đổi cột Date thành Date Time Index

Quá trình này đảm bảo rằng các quan sát được sắp xếp theo thứ tự thời gian và có thể được sử dụng để vẽ biểu đồ hoặc thực hiện các phép toán liên quan đến thời gian.

4.2 Kiểm tra dữ liệu trống

Chúng tôi kiểm tra tập dữ liệu **AAPL.csv** để xác định xem có giá trị trống (missing values) trong các cột dữ liệu hay không.

Bảng 4 thể hiện số lượng giá trị trống trong mỗi trường dữ liệu của tập dữ liệu. Kết quả kiểm tra cho thấy không có giá trị trống trong bất kỳ cột nào, bao gồm **Date**, **Open**, **High**, **Low**, **Close**, **Adj Close** và **Volume**

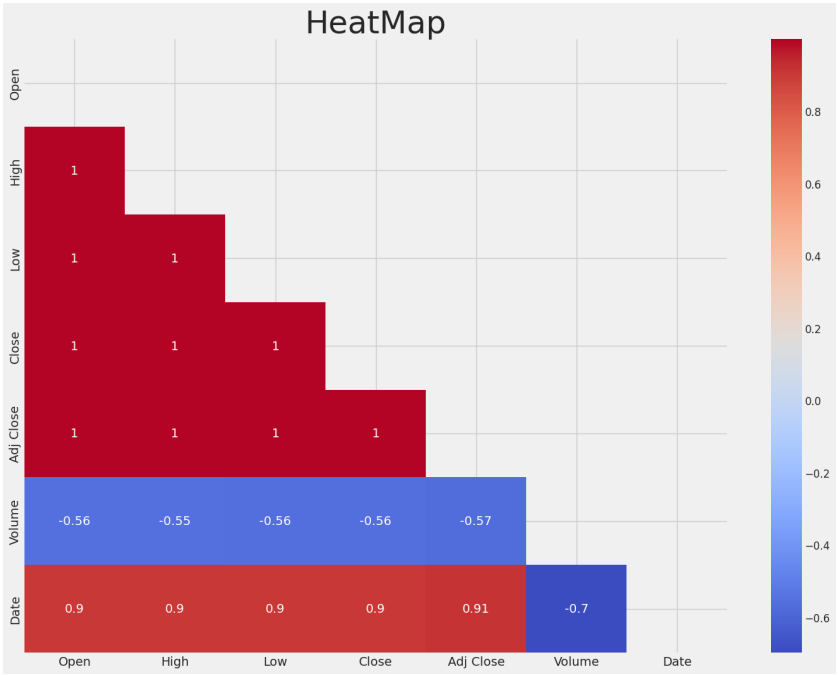
Thuộc tính	Số giá trị null
Open	0
High	0
Low	0
Close	0
Adj Close	0
Volume	0
Date	0

Bảng 4: Số lượng giá trị null trong mỗi thuộc tính

4.3 Loại bỏ các đặc trưng dư thừa

Để tập trung vào phân tích chuỗi thời gian của giá cổ phiếu và tránh các vấn đề về đa cộng tuyến, chúng tôi đã kiểm tra mức độ tương quan giữa các đặc trưng trong tập dữ liệu.

Hình 2 thể hiện ma trận tương quan giữa các trường dữ liệu trong tập dữ liệu AAPL.csv



Hình 2: Ma trận tương quan giữa các cột trong tập dữ liệu AAPL.csv

Từ biểu đồ này, chúng tôi nhận thấy các cột **Open**, **High**, **Low**, và **Close** có tính đa cộng tuyến cao với **Adj Close**, với hệ số tương quan lần lượt là 1.0, 1.0, 1.0, và 1.0. Ngoài ra, cột **Date** có tương quan dương mạnh với các cột giá (khoảng 0.9 đến 0.91), cho thấy giá cổ phiếu của Apple có xu hướng tăng dần theo thời gian trong giai đoạn từ 03/01/2012 đến 30/12/2019. Do đó, chúng tôi quyết định loại bỏ các cột **Open**, **High**, **Low**, và **Close**, đồng thời chọn **Adj Close** làm biến mục tiêu, vì **Adj Close** đã được điều chỉnh để phản ánh các quyết định của công ty như chia tách cổ phiếu (stock split) và cổ tức (dividends), mang lại giá trị đại diện chính xác hơn cho giá cổ phiếu.

Ngoài ra, cột **Volume** (khối lượng giao dịch) có mức độ tương quan âm vừa phải với

Adj Close (hệ số tương quan = -0.57), và không đóng góp đáng kể vào việc dự báo giá cổ phiếu. Hơn nữa, **Volume** có tương quan âm mạnh với **Date** (-0.7), cho thấy khối lượng giao dịch có xu hướng giảm dần theo thời gian. Vì vậy, cột **Volume** cũng được loại bỏ. Sau bước này, tập dữ liệu chỉ còn lại cột **Adj Close** để phân tích.

Bảng 5 thể hiện dữ liệu sau khi dữ liệu loại bỏ các đặc trưng dư thừa như **Open**, **High**, **Low**, **Close**, **Volume**

Bảng 5: Dữ liệu sau khi loại bỏ đặc trưng dư thừa

Date	**Adj Close**
2012-01-31	50.765709
2012-02-29	51.038536
2012-03-31	51.605175
2012-04-30	52.144630
2012-05-31	52.61932
⋮	⋮
2019-12-23	282.054138
2019-12-24	282.322266
2019-12-26	287.923645
2019-12-27	287.814392
2019-12-30	289.522614

4.4 Tái lấy mẫu dữ liệu thành tần suất hàng tháng

Tập dữ liệu ban đầu được ghi nhận theo tần suất hàng ngày, với 2011 quan sát từ 03/01/2012 đến 30/12/2019. Tuy nhiên, để giảm độ nhiễu và tập trung vào xu hướng dài hạn, chúng tôi tái lấy mẫu dữ liệu thành tần suất hàng tháng bằng cách lấy giá trị trung bình của **Adj Close** trong mỗi tháng.

Để tái lấy mẫu dữ liệu thành tần suất hàng tháng, chúng tôi nhóm dữ liệu theo tần suất hàng tháng, sau đó áp dụng hàm `mean()` để tính giá trị trung bình của **Adj Close** cho mỗi tháng.

Bảng 6 thể hiện một phần dữ liệu sau khi tái lấy mẫu thành tần suất hàng tháng. Sau khi tái lấy mẫu, tập dữ liệu giảm xuống còn 96 quan sát, tương ứng với 96 tháng từ tháng 01/2012 đến tháng 12/2019. Quá trình này giúp làm mượt dữ liệu và phù hợp hơn với mục tiêu phân tích xu hướng dài hạn của giá cổ phiếu AAPL.

Bảng 6: Một phần dữ liệu sau khi tái lấy mẫu thành tần suất hàng tháng

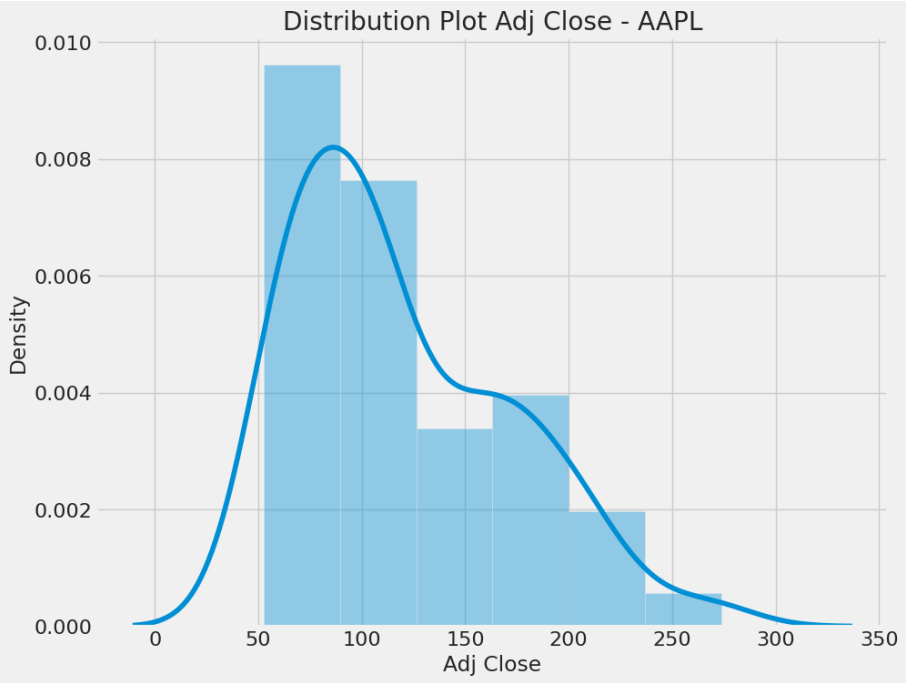
Index	**Adj Close**
2012-01-31	52.907298
2012-02-29	61.424381
2012-03-31	71.292448
2012-04-30	74.810151
2012-05-31	69.708045
⋮	⋮
2019-08-31	202.738817
2019-09-30	215.853332
2019-10-31	232.974974
2019-11-30	260.569057
2019-12-31	273.780717

Để trực quan hóa xu hướng dài hạn của dữ liệu sau khi tái lấy mẫu, chúng tôi vẽ biểu đồ đường biểu diễn giá **Adj Close** trung bình hàng tháng (hệ số tương quan giữa **Date** và **Adj Close** là 0.91). Kết quả của biểu đồ đường được trình bày trong hình 3.

Hình 3: Biểu đồ đường biểu diễn giá Adj Close trung bình hàng tháng của cổ phiếu Apple.

Biểu đồ trên cho thấy một xu hướng tăng trưởng dài hạn của giá cổ phiếu Apple từ năm 2012 đến năm 2019. Quá trình tái lấy mẫu đã làm mượt dữ liệu, giúp dễ dàng nhận diện xu hướng dài hạn của giá cổ phiếu Apple mà không bị ảnh hưởng bởi các biến động ngắn hạn hàng ngày. Điều này phù hợp với mục tiêu phân tích chuỗi thời gian, đặc biệt khi áp dụng các mô hình dự báo như ARIMA, vốn yêu cầu dữ liệu ít nhiễu hơn để đạt hiệu quả cao.

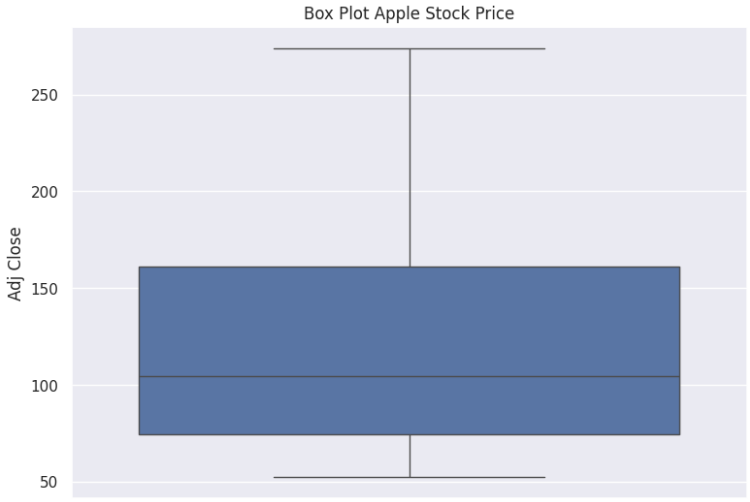
Để phân tích thêm về phân phối của dữ liệu **Adj Close** trung bình cổ phiếu hàng tháng của Apple từ 01/2012 đến 12/2019, chúng tôi vẽ biểu đồ phân phối (distribution plot). Kết quả được trình bày trong hình 4



Hình 4: Biểu đồ phân phối của giá Adj Close trung bình hàng tháng của cổ phiếu Apple.

Biểu đồ trên cho thấy phân phối của **Adj Close** trung bình hàng tháng có dạng lệch phải (right-skewed) và không tuân theo phân phối chuẩn, điều này có thể ảnh hưởng đến việc áp dụng các mô hình chuỗi thời gian như ARIMA, vốn thường giả định sai số (residuals) hoặc dữ liệu đã biến đổi tuân theo phân phối chuẩn.

Để phân tích sự phân tán và các giá trị ngoại lai của dữ liệu **Adj Close** trung bình hàng tháng, chúng tôi vẽ biểu đồ hộp (box plot) biểu diễn biểu đồ hộp cung cấp thông tin về phạm vi, trung vị, và sự phân tán của dữ liệu **Adj Close**. Kết quả được trình bày trong hình 5.

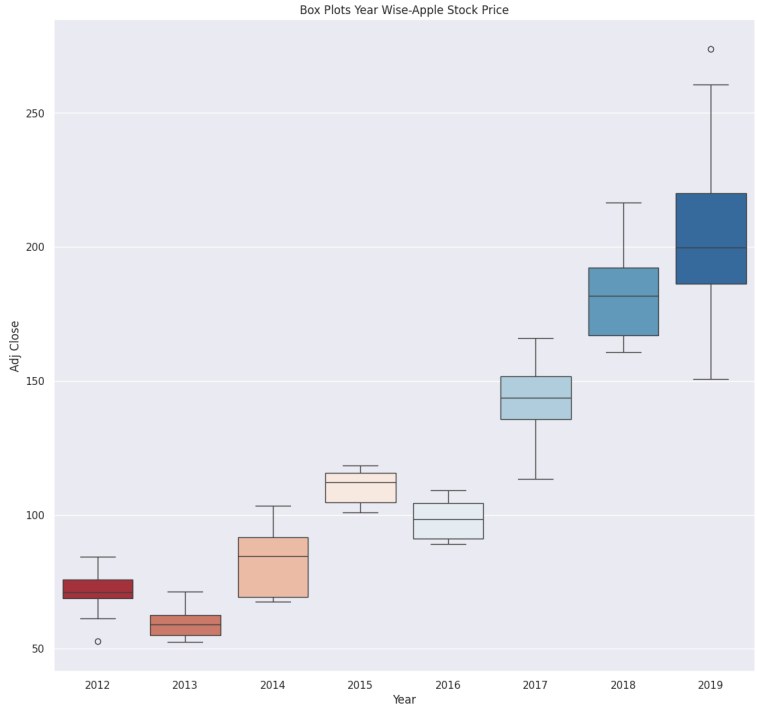


Hình 5: Biểu đồ hộp biểu diễn sự phân tán và giá trị ngoại lai của giá Adj Close trung bình hàng tháng của cổ phiếu Apple.

Adj Close

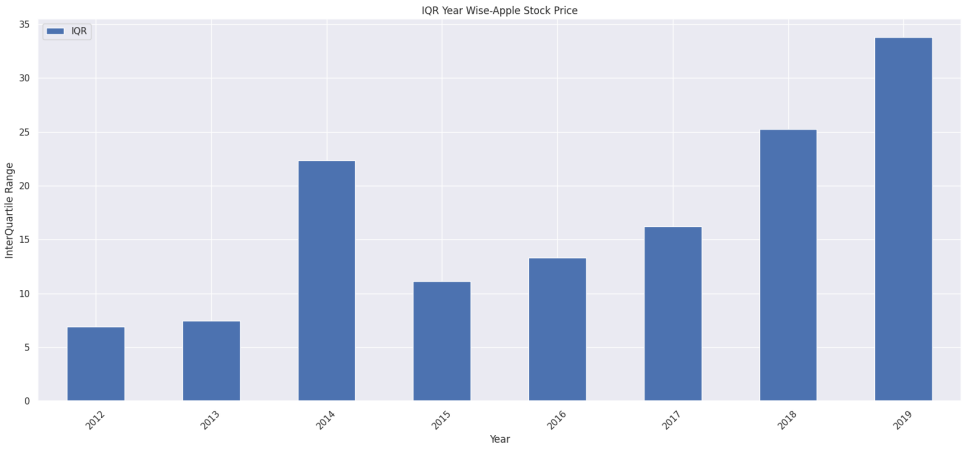
Biểu đồ hộp trên bổ sung cho biểu đồ phân phối, xác nhận đặc điểm lệch phải của dữ liệu **Adj Close** và không có giá trị ngoại lai, trung vị của trung bình hàng tháng nằm khoảng 120 USD, với khoảng tứ phân vị (IQR) kéo dài từ khoảng 80 USD (Q1) đến 170 USD (Q3).

Để phân tích sự phân tán và các giá trị ngoại lai của **Adj Close** theo từng năm, chúng tôi vẽ các biểu đồ hộp theo các năm từ năm 2012 đến năm 2019. Kết quả được trình bày trong hình 6



Hình 6: Biểu đồ hộp theo năm biểu diễn sự phân tán và giá trị ngoại lai của giá Adj Close trung bình hàng tháng của cổ phiếu Apple.

Biểu đồ trên cho thấy trung vị của **Adj Close** tăng dần qua các năm, từ khoảng 60 USD (2012) lên 220 USD (2019), xác nhận xu hướng tăng trưởng dài hạn của giá cổ phiếu Apple, phản ánh sự gia tăng biến động giá cổ phiếu qua thời gian.

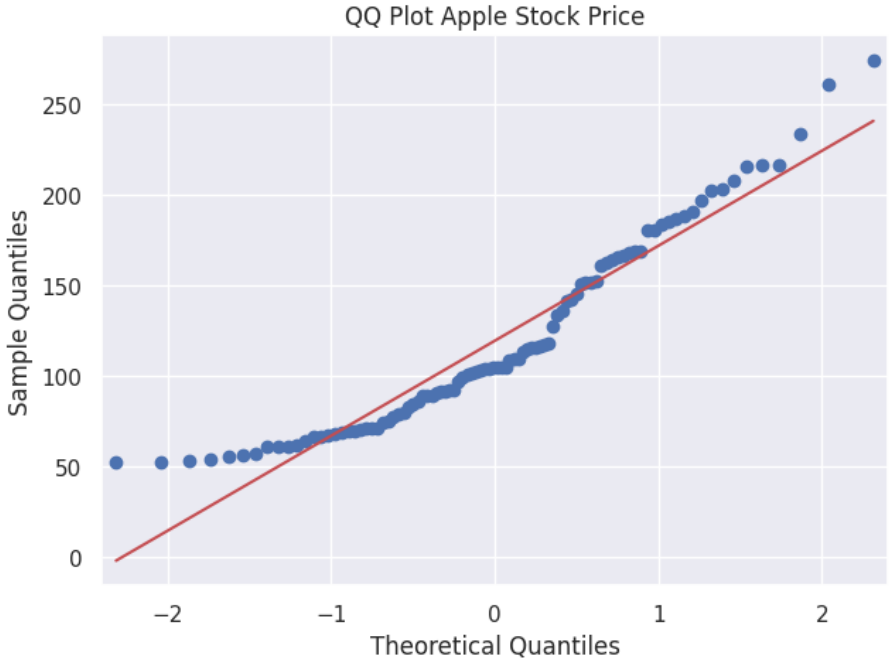


Hình 7: Biểu đồ cột biểu diễn khoảng tứ phân vị (IQR) của giá Adj Close trung bình hàng tháng của cổ phiếu Apple theo từng năm.

Để định lượng sự phân tán của **Adj Close** trung bình hàng tháng của cổ phiếu Apple theo từng năm từ 2012 đến 2019, chúng tôi tính khoảng tứ phân vị (IQR) cho mỗi năm và vẽ biểu đồ cột. Kết quả được trình bày trong hình 7.

Biểu đồ trên cho thấy IQR của **Adj Close** trung bình hàng tháng tăng dần qua các năm, các năm 2014, 2018, và 2019 có IQR cao hơn đáng kể,phản ánh các giai đoạn giá cổ phiếu biến động mạnh, sự gia tăng của IQR qua thời gian cho thấy giá cổ phiếu Apple không chỉ tăng về giá trị trung bình mà còn biến động mạnh hơn.

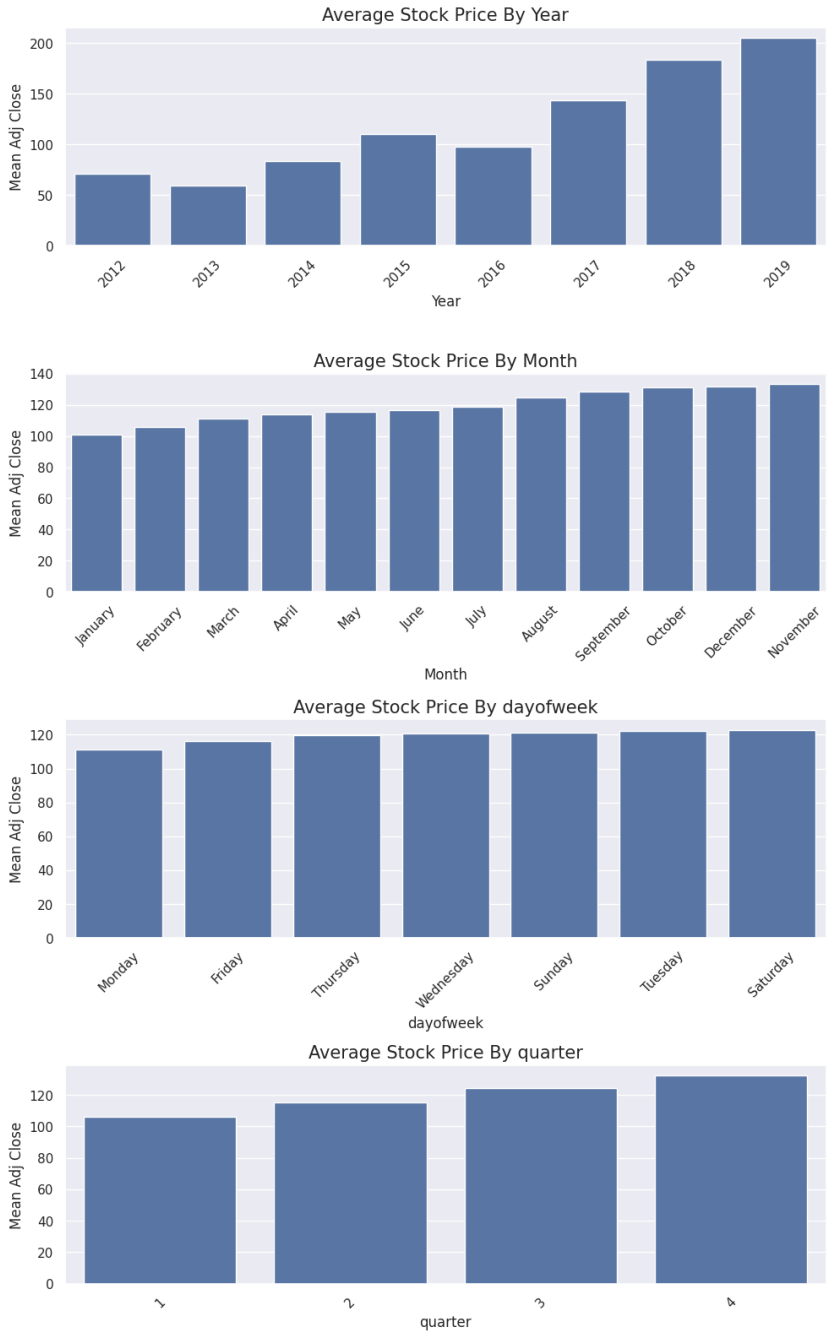
Để đánh giá xem dữ liệu **Adj Close** trung bình hàng tháng có tuân theo phân phối chuẩn hay không, chúng tôi vẽ biểu đồ Quantile-Quantile (QQ) của **Adj Close** trung bình hàng tháng của cổ phiếu Apple theo từng năm. Kết quả được trình bày trong hình 8. của Adj Close trung bình hàng tháng của cổ phiếu Apple theo từng năm



Hình 8: Biểu đồ QQ-Plot của giá Adj Close trung bình hàng tháng của cổ phiếu Apple theo từng năm.

Biểu đồ trên cho thấy các điểm dữ liệu không nằm thẳng trên đường màu đỏ, đặc biệt ở hai đuôi (trái và phải), cho thấy dữ liệu **Adj Close** không tuân theo phân phối chuẩn. Sự lệch ở hai đuôi, đặc biệt là đuôi phải dài hơn, xác nhận rằng dữ liệu **Adj Close** có phân phối lệch phải.

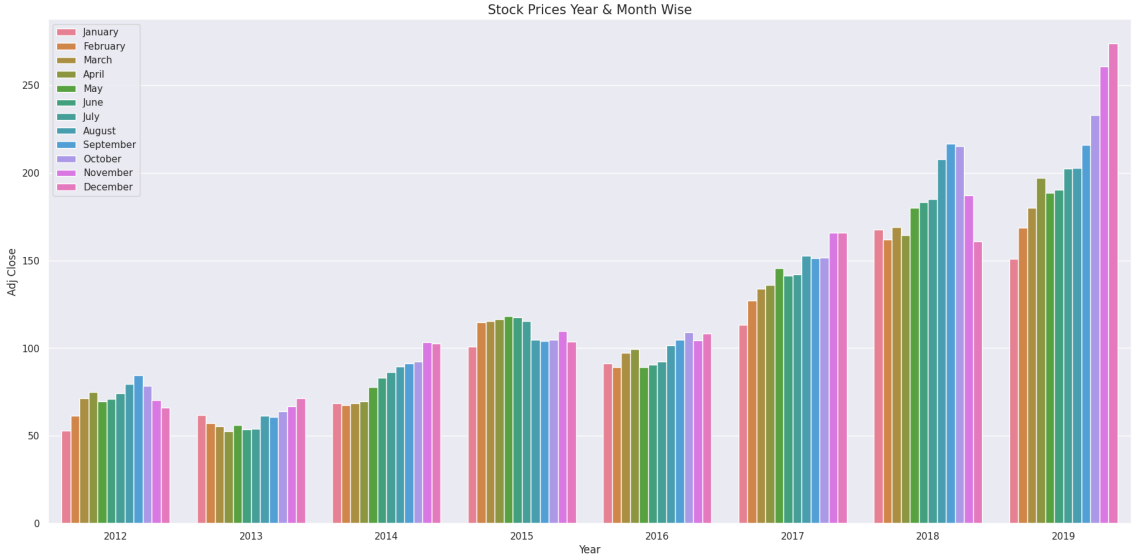
Để phân tích giá trị trung bình của **Adj Close** theo nhiều tiêu chí nhóm khác nhau (năm, tháng, ngày trong tuần và quý), chúng tôi vẽ một tập hợp các biểu đồ cột trên. Kết quả được trình bày trong hình 9.



Hình 9: Biểu đồ cột biểu diễn giá trị trung bình của **Adj Close** theo nhiều tiêu chí nhóm khác nhau của cổ phiếu Apple.

Biểu đồ đầu tiên (theo năm) cho thấy giá trung bình của **Adj Close** có xu hướng tăng trưởng dài hạn. Biểu đồ thứ hai (theo tháng) cho thấy giá trung bình dao động nhẹ giữa các tháng, nhưng không có xu hướng mùa vụ rõ ràng. Biểu đồ thứ ba (theo ngày trong tuần) cho thấy giá trung bình khá đồng đều giữa các ngày, nhưng sự chênh lệch không đáng kể, cho thấy không có hiệu ứng ngày trong tuần rõ rệt. Cuối cùng, biểu đồ thứ tư (theo quý) cho thấy giá trung bình tăng dần, cho thấy một xu hướng tăng nhẹ theo quý, với giá cổ phiếu cao hơn vào cuối năm. Các biểu đồ cột này cung cấp một cái nhìn toàn diện về giá trị trung bình của **Adj Close** theo nhiều khía cạnh thời gian khác nhau.

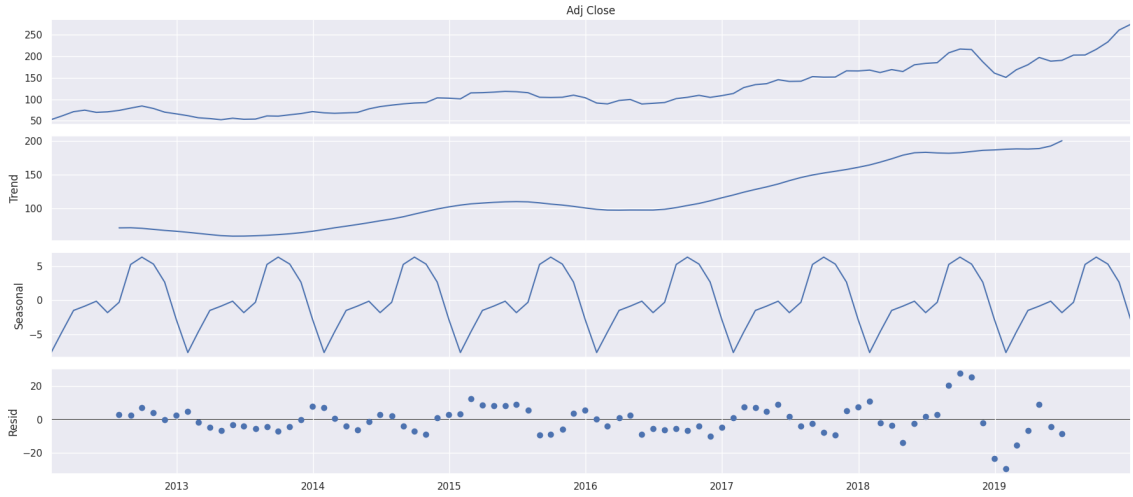
Để phân tích chi tiết hơn về giá cổ phiếu Apple theo cả năm và tháng, chúng tôi vẽ biểu đồ cột phân nhóm, trong đó mỗi năm được biểu diễn bằng một nhóm cột, và mỗi cột trong nhóm đại diện cho một tháng. Kết quả được trình bày trong hình 10



Hình 10: Biểu đồ cột phân nhóm biểu diễn giá đóng cửa điều chỉnh trung bình của cổ phiếu Apple theo năm và tháng.

Biểu đồ trên cho thấy **Adj Close** trung bình tăng dần qua các năm, các tháng cuối năm (October, November, December) thường có giá cao hơn trong hầu hết các năm. Biểu đồ này cung cấp cái nhìn chi tiết hơn về giá cổ phiếu Apple, phân tích đồng thời cả yếu tố năm và tháng.

Để hiểu rõ hơn về các thành phần cơ bản của chuỗi thời gian giá cổ phiếu Apple, chúng tôi thực hiện phân rã chuỗi thời gian (time series decomposition) trên dữ liệu **Adj Close** hàng ngày, nhằm tách biệt các thành phần xu hướng (trend), mùa vụ (seasonality), và nhiễu (residual)



Hình 11: Biểu đồ phân rã chuỗi thời gian của giá đóng cửa điều chỉnh hàng ngày của cổ phiếu Apple, bao gồm dữ liệu gốc, xu hướng, mùa vụ, và nhiễu.

Qua biểu đồ trên cho thấy phân rã chuỗi thời gian cho thấy giá cổ phiếu Apple chủ yếu được điều khiển bởi xu hướng tăng dài hạn, với yếu tố mùa vụ nhỏ nhưng lặp lại (giá tăng vào cuối năm). Nhiễu lớn hơn trong các năm 2013, 2015, 2018, và 2019 phản ánh các biến động bất thường, có thể liên quan đến các sự kiện thị trường hoặc nội tại của Apple (như báo cáo doanh thu, ra mắt sản phẩm).

4.5 Tiền xử lý tính dừng chuỗi thời gian

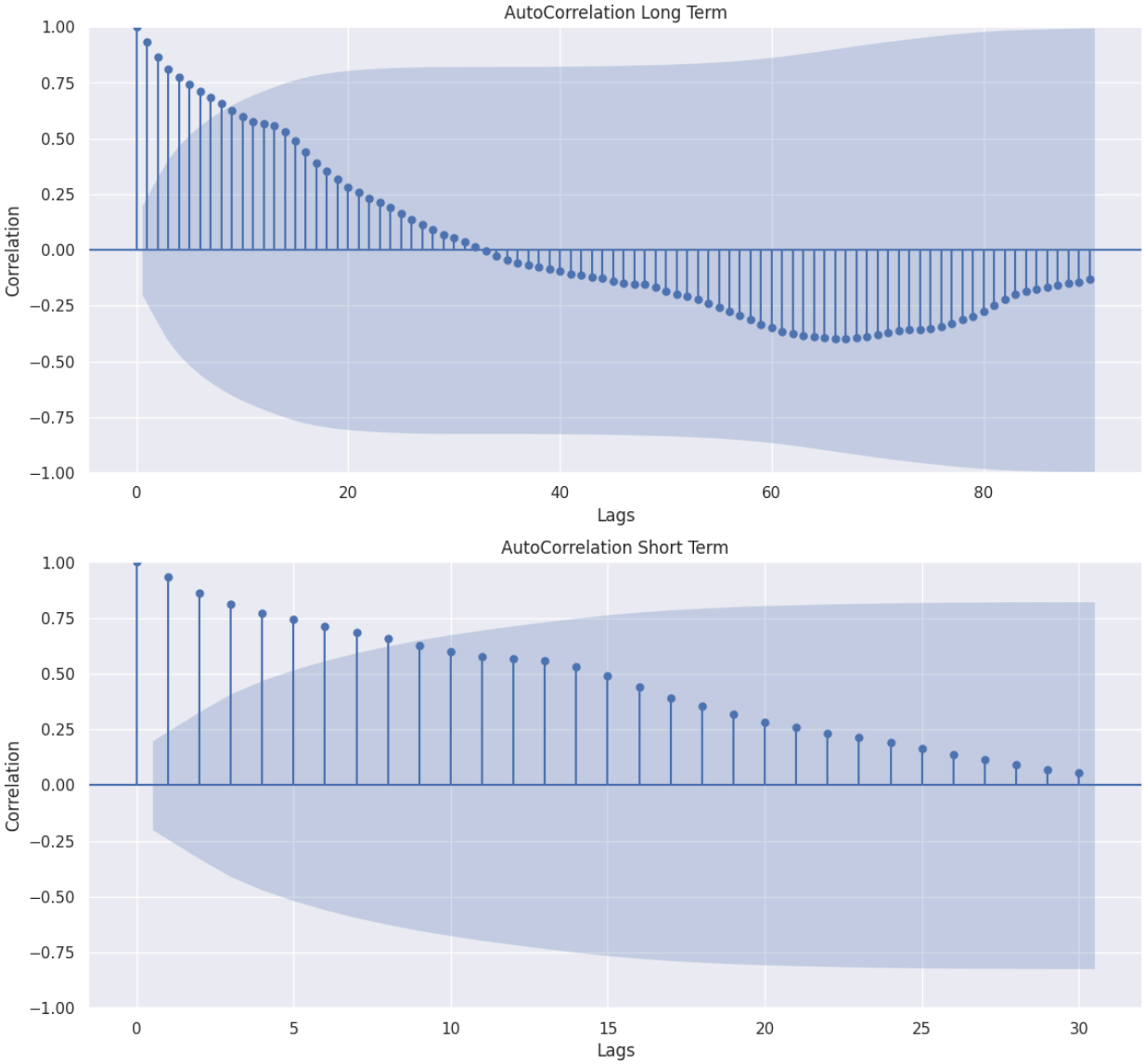
4.5.1 Kiểm tra tính dừng của chuỗi ban đầu

Trong phân tích chuỗi thời gian, tính dừng là điều kiện quan trọng nhằm đảm bảo tính hiệu quả và độ chính xác của các mô hình dự báo. Một chuỗi được gọi là dừng khi các đặc trưng thống kê như kỳ vọng, phương sai và hiệp phương sai không thay đổi theo thời gian.

Để kiểm tra tính dừng, nghiên cứu sử dụng kiểm định Augmented Dickey-Fuller (ADF) nhằm phát hiện đơn vị gốc trong chuỗi. Kết quả kiểm định ADF như sau:

- **ADF Statistic**: 1.339253
- **p-value**: 0.996820
- **Critical Values**:
  - 1%: -3.504
  - 5%: -2.894
  - 10%: -2.584

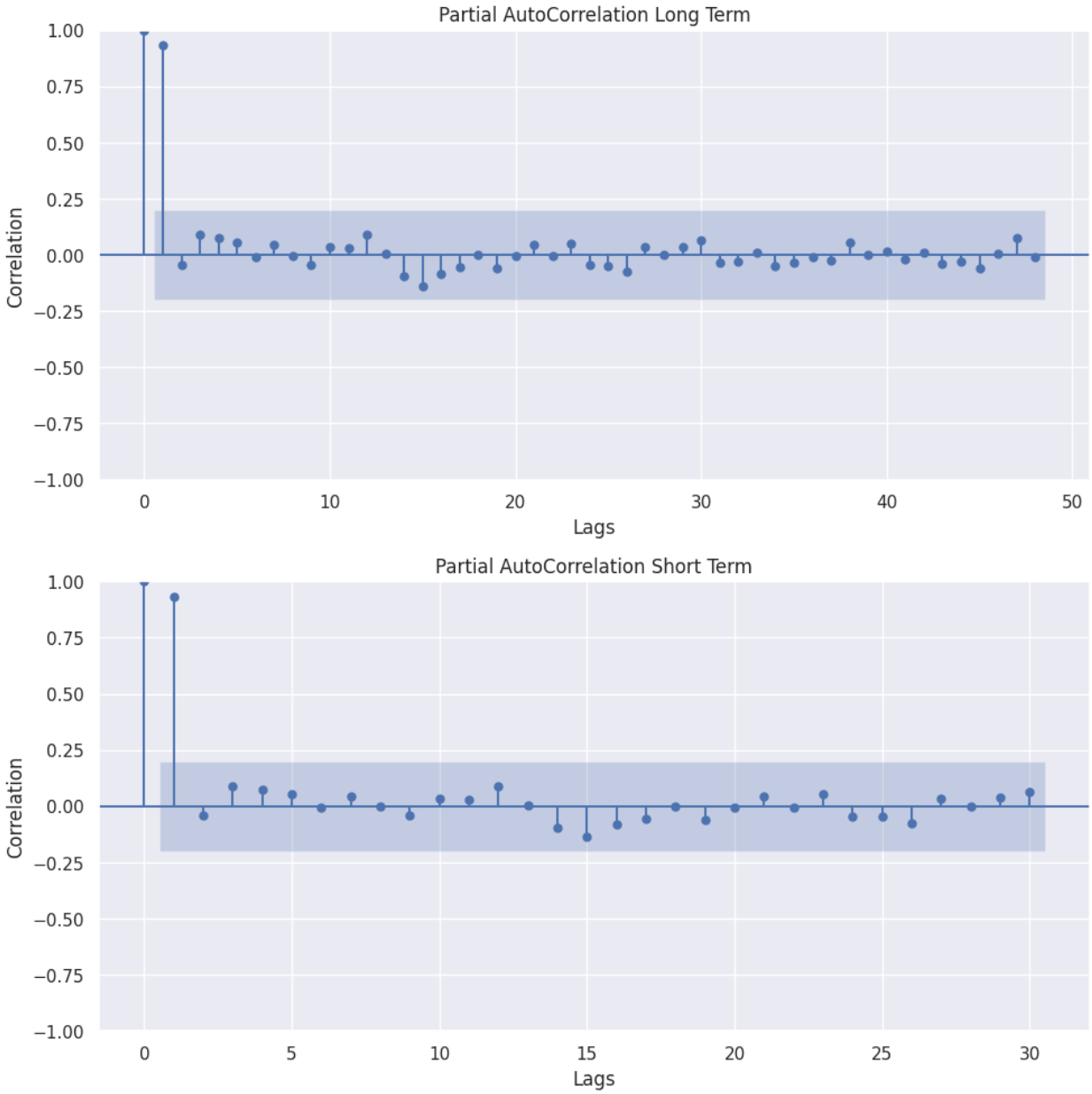
Vì ADF Statistic lớn hơn tất cả các mức giá trị tới hạn và p-value > 0.05, chúng tôi không thể bác bỏ giả thuyết không (H_0), tức là chuỗi không có tính dừng. Chuỗi cần được biến đổi (chẳng hạn lấy sai phân) để phù hợp cho các mô hình chuỗi thời gian như SARIMA.



Hình 12: Đồ thị ACF

- Đồ thị ACF dài hạn cho thấy hệ số tương quan tự động giảm dần rất chậm theo số trễ, đặc trưng của chuỗi không dừng. Một số giá trị trễ xa vẫn còn tương quan đáng kể.
- ACF ngắn hạn thể hiện các giá trị tương quan cao dần đều từ trễ 1 đến khoảng trễ 25, không suy giảm nhanh về 0, cho thấy hiện tượng "dư âm" trong chuỗi, dấu hiệu rõ ràng của chuỗi không dừng.

Chuỗi thể hiện sự tự tương quan mạnh mẽ qua nhiều bước trễ, điều này là dấu hiệu đặc trưng của một chuỗi có xu hướng hoặc chuỗi không dừng.



Hình 13: Đồ thị PACF

- PACF dài hạn cho thấy hệ số tương quan từng phần giảm nhanh và dao động quanh 0 sau trễ thứ nhất hoặc thứ hai, điều này thường chỉ ra chuỗi có thể được mô hình hóa bằng AR (Autoregressive).
- Ở PACF ngắn hạn, hệ số tại trễ 1 và 2 là đáng kể, sau đó nhanh chóng suy giảm về gần 0, phù hợp với mô hình AR(1) hoặc AR(2).

Dù PACF cho thấy khả năng mô hình hóa bằng AR, nhưng kết hợp với ADF và ACF, chuỗi hiện tại vẫn cần được chuyển thành chuỗi dừng trước khi mô hình hóa.

Từ kiểm định ADF và phân tích đồ thị ACF/PACF, có thể kết luận rằng chuỗi hiện tại là chuỗi không dừng, có thể chứa xu hướng hoặc phương sai không ổn định theo thời gian.

Trước khi xây dựng mô hình dự báo, chuỗi cần được lấy sai phân (differencing) hoặc biến đổi log/sqrt để đảm bảo tính dừng, giúp các mô hình như ARIMA, SARIMA hoạt động hiệu quả và chính xác hơn.

4.5.2 Lấy sai phân để biến chuỗi thành chuỗi dừng

Sau khi xác định chuỗi ban đầu không có tính dừng thông qua kiểm định Augmented Dickey-Fuller (ADF) và phân tích đồ thị tự tương quan (ACF) cũng như tự tương quan riêng phần (PACF), nghiên cứu tiến hành lấy sai phân bậc nhất nhằm loại bỏ xu hướng và đảm bảo các đặc tính thống kê của chuỗi không thay đổi theo thời gian – điều kiện cần để áp dụng hiệu quả các mô hình chuỗi thời gian như ARIMA hoặc SARIMA.

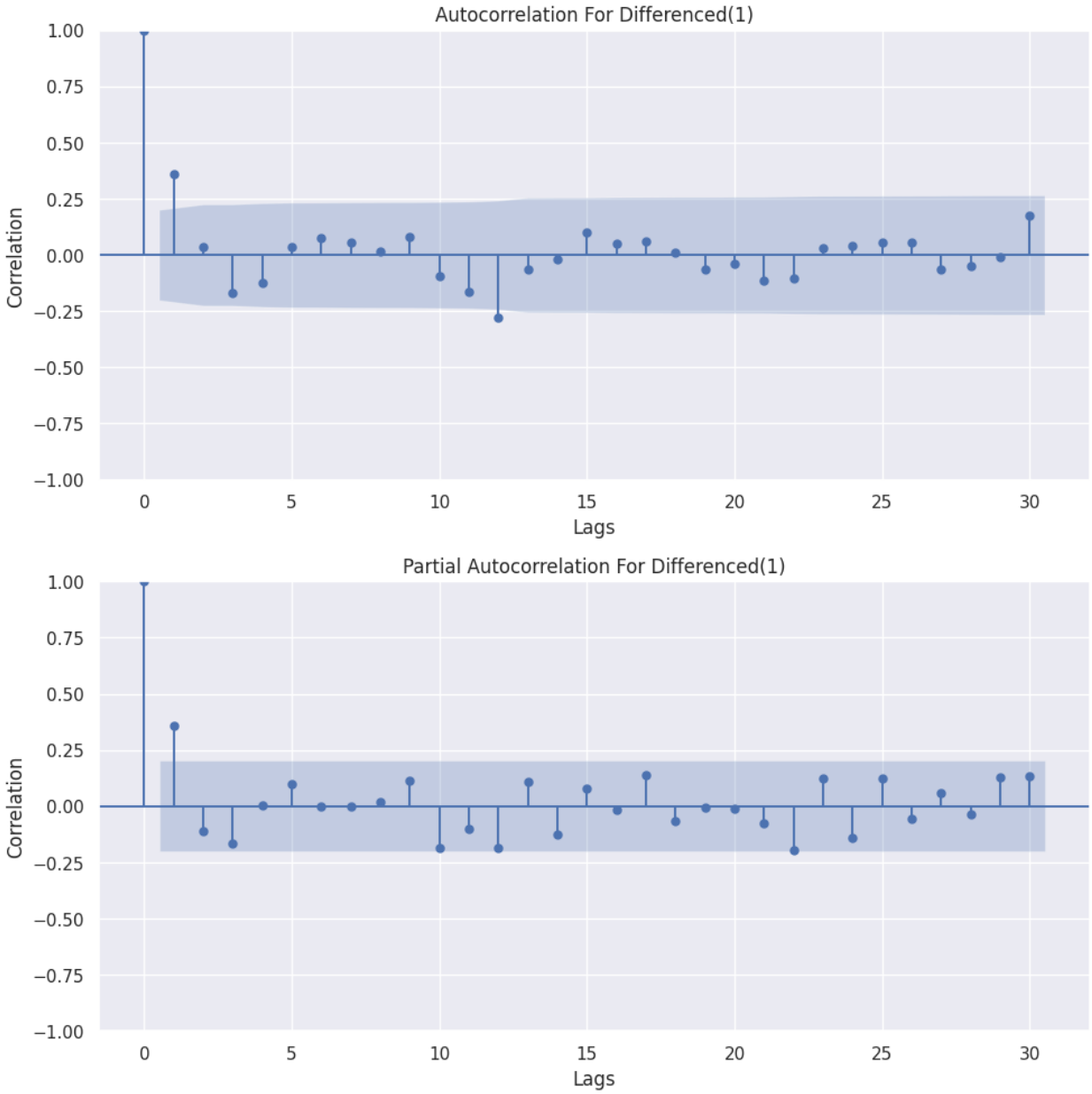
Để đánh giá tính dừng của chuỗi sau khi lấy sai phân, nghiên cứu tiếp tục sử dụng kiểm định Augmented Dickey-Fuller (ADF). Kết quả kiểm định được trình bày như sau:

- **ADF Statistic**: -6.501865
- **p-value**: 0.000000
- **Các giá trị tới hạn (Critical Values)**:
  - Mức 1%: -3.502
  - Mức 5%: -2.893
  - Mức 10%: -2.583

Dễ dàng nhận thấy rằng giá trị thống kê ADF nhỏ hơn đáng kể so với tất cả các ngưỡng tới hạn ở mức ý nghĩa 1%, 5% và 10%. Đồng thời, giá trị p-value xấp xỉ bằng 0 (p-value ≈ 0) cho thấy bằng chứng rất mạnh để bác bỏ giả thuyết không (H_0) - tức là chuỗi có đơn vị gốc. Điều này đồng nghĩa với việc chuỗi sau khi sai phân đã đạt được tính dừng về mặt thống kê.

Ngoài ra, phân tích đồ thị ACF và PACF của chuỗi sau khi lấy sai phân cũng củng cố kết luận trên:

- Các hệ số tương quan tự động (ACF) và tương quan từng phần (PACF) giảm nhanh về gần 0 sau trễ thứ nhất, đồng thời dao động ngẫu nhiên trong khoảng tin cậy 95%, cho thấy không còn hiện tượng tự tương quan có hệ thống.
- Biểu hiện này là dấu hiệu rõ ràng cho thấy chuỗi sau biến đổi không còn xu hướng và có tính ổn định theo thời gian – tức là đã trở thành chuỗi dừng.



Hình 14: Đồ thị ACF và PACF của chuỗi sau khi lấy sai phân

Tổng hợp cả kết quả kiểm định định lượng (ADF) và phân tích định tính (ACF/PACF), có thể kết luận rằng chuỗi sau khi lấy sai phân bậc nhất hoàn toàn đạt được tính dừng, và do đó đủ điều kiện để bước vào giai đoạn xây dựng mô hình dự báo chuỗi thời gian như ARIMA hoặc SARIMA.

5 Cơ Sở Lý Thuyết Của Các Mô Hình

5.1 SARIMA

5.1.1 Autoregressive Models (AR)

Trong phân tích chuỗi thời gian, mô hình hồi quy bội (Multiple Regression Model) được sử dụng để dự báo một biến phụ thuộc dựa trên tổ hợp tuyến tính của một hoặc nhiều biến độc lập. Tuy nhiên, khi biến phụ thuộc có mối quan hệ với chính các giá trị trong quá khứ của nó, mô hình tự hồi quy (AutoRegressive - AR) trở thành một lựa chọn phù hợp hơn. Mô hình này dựa trên giả định rằng giá trị hiện tại của chuỗi có thể được biểu diễn như là một tổ hợp tuyến tính của các giá trị trễ (Lagged Values) của chính nó, kèm theo một thành phần nhiễu trắng (White Noise).
Một mô hình tự hồi quy bậc p, ký hiệu là **AR(p)**, được mô tả theo công thức tổng quát như sau:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

Trong đó:

- y_t: giá trị của chuỗi tại thời điểm t;
- c: hằng số (intercept);
- $\phi_1, \phi_2, \ldots, \phi_p$: các hệ số tự hồi quy;
- ε_t: nhiễu trắng tại thời điểm t, được giả định là phân phối chuẩn với trung bình bằng 0 và phương sai không đổi.

Mô hình AR(p) có cấu trúc tương tự như một mô hình hồi quy tuyến tính bội, nhưng với các biến độc lập là các giá trị trong quá khứ của chính biến phụ thuộc. Đây là một mô hình có tính linh hoạt cao, cho phép biểu diễn nhiều dạng hành vi của chuỗi thời gian khác nhau chỉ bằng cách thay đổi các hệ số ϕ_i.
Mô hình AR(1):
Trong trường hợp đơn giản nhất, khi $p = 1$, ta thu được mô hình **AR(1)**:

$$y_t = c + \phi_1 y_{t-1} + \varepsilon_t$$

Ý nghĩa của các giá trị ϕ_1 trong mô hình AR(1) có thể được diễn giải như sau:

- Nếu $\phi_1 = 0$, mô hình trở thành $y_t = c + \varepsilon_t$, tương đương với **nhiễu trắng**, tức là các quan sát là độc lập và không có sự phụ thuộc theo thời gian.

- Nếu $\phi_1 = 1$ và $c = 0$, mô hình trở thành một **bước ngẫu nhiên** (random walk), trong đó giá trị hiện tại phụ thuộc hoàn toàn vào giá trị liền trước cộng với nhiễu, và quá trình không dừng.
- Nếu $\phi_1 = 1$ và $c \neq 0$, mô hình là **bước ngẫu nhiên có trôi** (random walk with drift), thể hiện xu hướng tăng hoặc giảm dần theo thời gian.
- Nếu $\phi_1 < 0$, quá trình có xu hướng dao động xung quanh giá trị trung bình, với hiện tượng điều chỉnh ngược chiều mỗi khi lệch khỏi trung bình.

Điều kiện dừng của mô hình AR(p) Để một quá trình AR(p) là **dừng** (stationary), các nghiệm của đa thức đặc trưng liên quan đến phương trình:

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0$$

phải **nằm bên ngoài vòng tròn đơn vị** trong mặt phẳng phức, tức là **mọi nghiệm** z **phải thỏa mãn** $|z| > 1$. Một điều kiện cần thiết đơn giản hơn nhưng thường dùng trong thực hành là: $|\phi_i| < 1$ đối với mọi $i = 1, 2, \ldots, p$. Điều kiện này đảm bảo rằng ảnh hưởng của các giá trị trong quá khứ sẽ giảm dần theo thời gian và chuỗi sẽ dao động quanh một giá trị kỳ vọng ổn định.

Tóm lại, mô hình tự hồi quy là một công cụ mạnh mẽ và linh hoạt trong phân tích chuỗi thời gian, đặc biệt khi dữ liệu thể hiện sự phụ thuộc vào các giá trị trong quá khứ. Việc lựa chọn bậc p và kiểm định tính dừng là các bước quan trọng nhằm đảm bảo mô hình được xây dựng là hợp lý và có khả năng dự báo tốt.

5.1.2 Integrated (I)

Thành phần "Integrated" (tích phân) đại diện cho quá trình lấy sai phân (differencing) nhằm biến đổi chuỗi thời gian ban đầu trở nên dừng (stationary).

Một chuỗi thời gian được gọi là dừng khi các đặc trưng thống kê của nó như kỳ vọng (mean) và phương sai (variance) không thay đổi theo thời gian. Việc đảm bảo tính dừng là điều kiện tiên quyết để áp dụng hiệu quả các mô hình chuỗi thời gian như ARIMA, bởi vì phần lớn các giả định trong mô hình đều yêu cầu dữ liệu đầu vào phải là chuỗi dừng.

Để đánh giá tính dừng của chuỗi, ta có thể sử dụng kiểm định Dickey-Fuller mở rộng (Augmented Dickey-Fuller Test - ADF). Kiểm định này giúp xác định liệu chuỗi có chứa unit root (rễ đơn vị) - dấu hiệu của chuỗi không dừng - hay không. Nếu chuỗi không dừng, ta cần áp dụng phép lấy sai phân với các bậc khác nhau (d = 1, 2, ...) để làm cho chuỗi trở nên dừng.

Cụ thể, với hệ số sai phân $d = 1$, ta thực hiện phép lấy hiệu giữa giá trị hiện tại và giá

trị ngay trước đó, tức là:

$$y'_t = y_t - y_{t-1}$$

Sau khi áp dụng phép lấy sai phân bậc 1, chuỗi kết quả thường có xu hướng ổn định hơn về trung bình và phương sai theo thời gian. Việc này giúp loại bỏ xu hướng (trend) hoặc các thành phần gây nhiễu dài hạn trong chuỗi, từ đó giúp chuỗi trở nên phù hợp hơn để xây dựng mô hình dự báo.

Trong trường hợp sai phân bậc 1 vẫn chưa đủ để làm chuỗi dừng, ta có thể thử với các bậc sai phân cao hơn. Tuy nhiên, việc lấy sai phân quá nhiều cũng có thể làm mất thông tin quan trọng trong dữ liệu, vì vậy cần cân nhắc kỹ lưỡng và đánh giá cẩn thận thông qua kiểm định cũng như trực quan hóa dữ liệu sau khi sai phân.

5.1.3 Moving Average (MA)

Mô hình Trung bình trượt (Moving Average - MA) là một trong những mô hình nền tảng trong phân tích chuỗi thời gian, được sử dụng để dự báo giá trị tương lai của chuỗi bằng cách khai thác thông tin từ các sai số dự báo trong quá khứ. Khác với mô hình tự hồi quy (Autoregressive - AR), vốn dựa vào các giá trị quan sát trước đó trong chuỗi thời gian, mô hình MA sử dụng một tổ hợp tuyến tính của các sai số tại những thời điểm trước đó nhằm mô tả và dự đoán các biến động trong tương lai.

Cấu trúc tổng quát của mô hình MA bậc q, ký hiệu là $MA(q)$, được biểu diễn dưới dạng:

$$m_t = \theta_0 + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q}$$

Trong đó:

- m_t là giá trị dự báo tại thời điểm t
- e_{t-i} là sai số tại thời điểm $t-i$, được định nghĩa là phần chênh lệch giữa giá trị thực tế và giá trị dự báo trong quá khứ,
- $\theta_0, \theta_1, \ldots, \theta_q$ là các tham số của mô hình, thể hiện trọng số ảnh hưởng của các sai số trong quá khứ đến giá trị dự báo hiện tại.

Do chỉ sử dụng các sai số để dự báo thay vì giá trị quan sát, mô hình MA có khả năng phản ánh các biến động ngẫu nhiên và những yếu tố bất định ảnh hưởng đến chuỗi thời gian. Điều này đặc biệt hữu ích trong các trường hợp chuỗi chứa nhiều nhiễu hoặc không có cấu trúc rõ ràng theo thời gian.

Tuy nhiên, một điểm cần lưu ý là các sai số e_t không thể quan sát trực tiếp từ dữ liệu, mà chỉ có thể được suy ra sau khi mô hình đã được ước lượng. Điều này làm cho việc ước lượng các tham số của mô hình MA trở nên phức tạp hơn so với mô hình AR. Thay vì sử

dụng phương pháp bình phương tối thiểu thông thường (Ordinary Least Squares - OLS), mô hình MA thường phải được ước lượng thông qua các phương pháp nâng cao hơn như Ước lượng khả năng tối đa (Maximum Likelihood Estimation - MLE).

Mô hình MA thường được sử dụng như một phần của các mô hình chuỗi thời gian tổng quát hơn, chẳng hạn như mô hình ARMA (Autoregressive Moving Average) hoặc ARIMA (Autoregressive Integrated Moving Average), giúp tăng khả năng mô hình hóa các chuỗi có tính chất phức tạp hơn.

5.1.4 Autoregressive Moving Average (ARMA)

Trong phân tích chuỗi thời gian, cơ chế sản sinh ra giá trị của chuỗi Y_t không chỉ có thể được mô hình hóa bằng một trong hai thành phần **tự hồi quy** (AR) hoặc **trung bình trượt** (MA), mà còn có thể kết hợp cả hai yếu tố này. Khi kết hợp hai thành phần AR và MA, mô hình chuỗi thời gian được gọi là mô hình **trung bình trượt tự hồi quy** (ARMA).

Cụ thể, mô hình **ARMA(1,1)** là mô hình đơn giản nhất trong họ ARMA, được biểu diễn dưới dạng:

$$Y_t = \theta + \phi_1 Y_{t-1} + \theta_0 \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

Trong đó:

- Y_t là giá trị của chuỗi tại thời điểm t,
- ε_t là nhiễu trắng (white noise), với các đặc tính không có tự tương quan và kỳ vọng bằng 0,
- ϕ_1 là tham số tự hồi quy,
- θ_0 và θ_1 là các tham số của phần trung bình trượt.

Tổng quát, mô hình **ARMA(p, q)** có thể được mô tả bằng phương trình sau:

$$Y_t = \theta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \theta_0 \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

Trong đó:

- Y_t là giá trị của chuỗi tại thời điểm t,
- $\phi_1, \phi_2, \ldots, \phi_p$ là các tham số tự hồi quy (AR),
- $\theta_0, \theta_1, \ldots, \theta_q$ là các tham số trung bình trượt (MA),
- ε_t là nhiễu trắng tại thời điểm t.

Mô hình **ARMA(p, q)** kết hợp các yếu tố tự hồi quy và trung bình trượt để mô tả sự phụ thuộc tuyến tính giữa các giá trị trong quá khứ của chuỗi Y_t và các sai số ngẫu nhiên. Việc xác định các tham số $\phi_1, \phi_2, \ldots, \phi_p$ và $\theta_0, \theta_1, \ldots, \theta_q$ là rất quan trọng trong việc tối ưu hóa mô hình, giúp cải thiện độ chính xác của dự báo chuỗi thời gian.

5.1.5 Autoregressive Integrated Moving Average (ARIMA)

Trong phân tích chuỗi thời gian, một chuỗi có thể có tính **dừng** (stationary) hoặc **không dừng** (non-stationary). Một chuỗi không dừng được gọi là **tích hợp bậc 1** và ký hiệu là $I(1)$, nếu sai phân bậc 1 của chuỗi đó là một chuỗi dừng. Mặt khác, một chuỗi được gọi là **tích hợp bậc** d, ký hiệu là $I(d)$, nếu sai phân bậc d của chuỗi đó là một chuỗi dừng, với d là số lần cần thiết để sai phân chuỗi cho đến khi chuỗi trở thành dừng. Cụ thể, nếu $d = 0$, chuỗi ban đầu là chuỗi dừng.

Khi một chuỗi Y_t tích hợp d, mô hình **ARMA(p, q)** có thể được áp dụng cho chuỗi sai phân bậc d. Mô hình này được gọi là **ARIMA(p, d, q)**, trong đó:

- d là số lần sai phân chuỗi Y_t để biến nó thành chuỗi dừng,
- p là bậc tự hồi quy (AutoRegressive) của chuỗi dừng,
- q là bậc trung bình trượt (Moving Average) của chuỗi dừng.

Mô hình **ARIMA(p, d, q)** có thể được diễn tả theo phương trình sau:

$$Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Trong đó:

- Y_t là giá trị của chuỗi tại thời điểm t,
- c là hằng số (thành phần sai lệch),
- $\phi_1, \phi_2, \ldots, \phi_p$ là các tham số tự hồi quy (AR),
- $\theta_1, \theta_2, \ldots, \theta_q$ là các tham số trung bình trượt (MA),
- ε_t là nhiễu trắng (white noise), giả định có kỳ vọng bằng 0, phương sai không đổi và không có tự tương quan theo thời gian.

Mô hình **ARIMA(p, d, q)** là sự kết hợp của ba thành phần chính:

- **AR(p)**: Phần tự hồi quy, mô hình hóa sự phụ thuộc tuyến tính giữa giá trị hiện tại và các giá trị trong quá khứ.
- **I(d)**: Phần tích hợp, xử lý tính không dừng của chuỗi thông qua sai phân bậc d.

- **MA(q)**: Phần trung bình trượt, mô hình hóa sự phụ thuộc của giá trị hiện tại vào các sai số trong quá khứ.

Hai mô hình đặc biệt của **ARIMA(p, d, q)** bao gồm:

- **AR(p)** là trường hợp đặc biệt khi $d = 0$ và $q = 0$, tức là chỉ có phần tự hồi quy mà không có phần sai phân và trung bình trượt.
- **MA(q)** là trường hợp đặc biệt khi $d = 0$ và $p = 0$, tức là chỉ có phần trung bình trượt mà không có phần tự hồi quy và sai phân.

Một trong các vấn đề quan trọng trong phân tích chuỗi thời gian là xác định các tham số d, p, q, cũng như các tham số $\theta_1, \theta_2, \ldots, \theta_q$ và $\phi_1, \phi_2, \ldots, \phi_p$. Việc lựa chọn các tham số này ảnh hưởng trực tiếp đến khả năng dự báo và tính chính xác của mô hình ARIMA trong việc mô phỏng và dự báo các chuỗi thời gian.

5.1.6 Seasonal ARIMA (SARIMA)

Mô hình ARIMA mô tả chuỗi thời gian dừng Y_t (sau khi đã thực hiện sai phân bậc d) theo phương trình tổng quát sau:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

Phương trình này có thể được biểu diễn dưới dạng toán tử trễ (lag operator) như sau:

$$\phi_p(L) Y_t = \phi_0 + \theta_q(L) \epsilon_t$$

Trong đó:

- L là toán tử trễ, với $L^k Y_t = Y_{t-k}$,
- $\phi_p(L)$ là đa thức tự hồi quy (AR),
- $\theta_q(L)$ là đa thức trung bình trượt (MA).

Mô hình này chưa xét đến yếu tố mùa vụ (seasonality). Để mở rộng mô hình ARIMA thành mô hình có yếu tố mùa vụ, ta có ba trường hợp chính sau:
Trường hợp 1: Chuỗi Y_t có yếu tố thời vụ
Khi chuỗi Y_t chứa thành phần mùa vụ, ta hiệu chỉnh chuỗi bằng cách áp dụng toán tử sai phân mùa vụ $(1 - L^s)$, trong đó s là chu kỳ mùa vụ (ví dụ: $s = 12$ cho chuỗi dữ liệu theo tháng, $s = 4$ cho dữ liệu theo quý). Phương trình mô hình lúc này sẽ trở thành:

$$\phi_p(L)(1 - L^s) Y_t = \phi_0 + \theta_q(L) \epsilon_t$$

Trường hợp 2: Thành phần nhiễu ϵ_t có yếu tố mùa vụ
Khi thành phần nhiễu ϵ_t có yếu tố mùa vụ, ta áp dụng toán tử sai phân mùa vụ vào phần nhiễu. Phương trình mô hình lúc này trở thành:

$$\phi_p(L)Y_t = \phi_0 + \theta_q(L)(1 - L^s)\epsilon_t$$

Trường hợp 3: Cả chuỗi Y_t và thành phần nhiễu ϵ_t đều có yếu tố mùa vụ
Khi cả chuỗi Y_t và phần nhiễu ϵ_t đều có yếu tố mùa vụ, ta cần áp dụng toán tử sai phân mùa vụ cho cả hai phần. Phương trình tổng quát sẽ được viết như sau:

$$\phi_p(L)(1 - L^s)Y_t = \phi_0 + \theta_q(L)(1 - L^s)\epsilon_t$$

Mô Hình SARIMA Tổng Quát
Mô hình có yếu tố mùa vụ được ký hiệu là:

$$SARIMA(p, d, q)(P, D, Q)_s$$

Trong đó:

- p, d, q là các bậc của phần không mùa (tự hồi quy, sai phân, trung bình trượt),
- P, D, Q là các bậc của phần mùa vụ (tự hồi quy mùa, sai phân mùa, trung bình trượt mùa),
- s là chu kỳ mùa vụ, là số đơn vị thời gian trong một mùa vụ (ví dụ: $s = 12$ cho dữ liệu theo tháng, $s = 4$ cho dữ liệu theo quý).

Mô hình SARIMA giúp xử lý hiệu quả các chuỗi thời gian có yếu tố mùa vụ rõ rệt. Theo *Diebold (2006)*, mô hình này đã chứng minh hiệu quả vượt trội trong việc dự báo chính xác các giá trị trong tương lai của chuỗi thời gian mùa vụ, nhờ khả năng linh hoạt kết hợp các yếu tố: xu hướng (trend), chu kỳ (seasonality) và ngẫu nhiên (random noise).

5.2 Prophet

Khác với mô hình SARIMA truyền thống dựa trên lý thuyết tự hồi quy và sai phân, Prophet sử dụng một cách tiếp cận hoàn toàn khác – đó là mô hình hồi quy cộng với các thành phần xu hướng, mùa vụ và sự kiện đặc biệt. Phần dưới đây sẽ làm rõ cấu trúc và cách hoạt động của Prophet trong việc xử lý chuỗi thời gian tài chính.

5.2.1 Tổng quan mô hình

Prophet là một mô hình dự báo chuỗi thời gian được phát triển bởi Facebook nhằm phục vụ các bài toán dự báo quy mô lớn trong kinh doanh. Mô hình này được thiết kế để dễ sử dụng, trực quan, và đặc biệt hiệu quả với các chuỗi thời gian có tính mùa vụ mạnh, xu hướng biến đổi theo thời gian và có ảnh hưởng bởi các sự kiện hoặc ngày lễ.

Prophet sử dụng mô hình hồi quy có thể diễn giải được dưới dạng tổng các thành phần:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

Trong đó:

- $g(t)$: thành phần xu hướng (trend), biểu diễn sự thay đổi dài hạn.
- $s(t)$: thành phần mùa vụ (seasonality), mô hình hóa các biến động định kỳ (tuần, năm).
- $h(t)$: thành phần ngày lễ/sự kiện, mô hình hóa ảnh hưởng của các ngày đặc biệt.
- ε_t: nhiễu ngẫu nhiên (noise), thường giả định có phân phối chuẩn.

5.2.2 Mô hình xu hướng $g(t)$

Co hai mô hình xu hướng phù hợp với nhiều ứng dụng của Facebook: một mô hình tăng trưởng bão hòa (saturating growth model) và một mô hình tuyến tính từng đoạn (piecewise linear model).

a) Mô hình tăng trưởng bão hòa (Logistic growth)

Dùng khi chuỗi thời gian thể hiện sự tăng trưởng giới hạn (như số người dùng). Hàm logistic cơ bản:

$$g(t) = \frac{C}{1 + \exp(-k(t - m))}$$

Với:

- C: sức chứa (carrying capacity).
- k: tốc độ tăng trưởng.
- m: điểm giữa thời gian (offset).

Mô hình Prophet mở rộng mô hình logistic truyền thống bằng cách cho phép thay đổi tốc độ tăng trưởng tại các thời điểm nhất định, gọi là *changepoints*. Có hai khía cạnh quan trọng được xét đến:

- Sức chứa không cố định (time-varying capacity): Prophet không sử dụng một giới hạn tăng trưởng cố định C, mà thay vào đó là một hàm theo thời gian $C(t)$ để phản ánh rằng giới hạn tăng trưởng có thể thay đổi, ví dụ như do nhiều người hơn có quyền truy cập Internet.
- Tốc độ tăng trưởng không cố định (time-varying growth rate): Tốc độ tăng trưởng có thể thay đổi mạnh do các sự kiện hoặc sản phẩm mới, do đó Prophet cho phép thay đổi tốc độ tăng tại các *changepoint*.

Giả sử có S changepoints xảy ra tại thời điểm s_j với $j = 1, \ldots, S$. Ta định nghĩa một vector điều chỉnh tốc độ $\boldsymbol{\delta} \in \mathbb{R}^S$, với δ_j là mức thay đổi tại thời điểm s_j. Tốc độ tăng trưởng tại thời điểm t được xác định bởi:

$$k + \sum_{j:t \geq s_j} \delta_j$$

Để biểu diễn dễ hơn, định nghĩa vector $\mathbf{a}(t) \in \{0, 1\}^S$ như sau:

$$a_j(t) = \begin{cases} 1, & \text{nếu } t \geq s_j \\ 0, & \text{ngược lại} \end{cases}$$

Tốc độ tại thời điểm t là:

$$k + \mathbf{a}(t)^\top \boldsymbol{\delta}$$

Khi tốc độ thay đổi, tham số m (offset) cũng cần điều chỉnh để đảm bảo tính liên tục giữa các đoạn. Điều chỉnh đúng tại changepoint j được tính là:

$$\gamma_j = \left(s_j - m - \sum_{l<j} \gamma_l \right) \left(1 - \frac{k + \sum_{l<j} \delta_l}{k + \sum_{l \leq j} \delta_l} \right)$$

Khi đó, mô hình logistic với các điểm thay đổi trở thành:

$$g(t) = \frac{C(t)}{1 + \exp\left(- \left(k + \mathbf{a}(t)^\top \delta \right) \left(t - \left(m + \mathbf{a}(t)^\top \gamma \right) \right) \right)}$$

Trong đó:

- δ: vector điều chỉnh tốc độ tăng trưởng tại các điểm thay đổi.
- γ: vector điều chỉnh độ lệch (offset) để đảm bảo tính liên tục.

- k: tốc độ tăng trưởng.
- m: điểm giữa thời gian (offset).
- $\mathbf{a}(t)$: vector chỉ thị xác định t có nằm sau điểm thay đổi nào không.

b) Mô hình xu hướng tuyến tính đoạn thẳng (Piecewise Linear Trend)

Phù hợp với chuỗi không có tăng trưởng bão hòa. Hàm xu hướng tuyến tính với changepoints:

$$g(t) = (k + \mathbf{a}(t)^\top \delta)t + (m + \mathbf{a}(t)^\top \gamma)$$

Trong đó:

- k: tốc độ tăng trưởng.
- m: điểm giữa thời gian (offset).
- δ: vector điều chỉnh tốc độ tăng trưởng tại các điểm thay đổi.
- γ: vector điều chỉnh độ lệch (offset) để đảm bảo tính liên tục.

c) Chọn điểm thay đổi tự động

Các điểm thay đổi có thể được chọn:

- Thủ công bởi nhà phân tích (ví dụ: ra mắt sản phẩm mới).
- Tự động bằng cách áp dụng phân phối tiên nghiệm thưa:

$$\delta_j \sim \text{Laplace}(0, \tau)$$

d) Ước lượng bất định dự báo (Forecast Uncertainty)

Prophet mô phỏng tương lai bằng cách giả định phân phối δ trong tương lai giống quá khứ:

$$\delta_j = \begin{cases} 0 & \text{với xác suất } \frac{T-S}{T} \\ \text{Laplace}(0, \lambda) & \text{với xác suất } \frac{S}{T} \end{cases}$$

Trong đó λ là ước lượng từ dữ liệu lịch sử:

$$\lambda = \frac{1}{S} \sum_{j=1}^{S} |\delta_j|$$

5.2.3 Thành phần mùa vụ $s(t)$

Được mô hình hóa bằng chuỗi Fourier:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right)$$

Hoặc viết dạng vector:

$$s(t) = \mathbf{X}(t)\boldsymbol{\beta}$$

Với:

- P: chu kỳ (365.25 ngày cho mùa vụ năm, 7 ngày cho tuần).
- $\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma^2)$: phân phối chuẩn để làm mượt.
- N: số lượng bậc Fourier.

5.2.4 Thành phần ngày lễ $h(t)$

Các ngày lễ ảnh hưởng đến chuỗi được thêm bằng cách chỉ định tập hợp các ngày D_i cho từng ngày lễ i. Thành phần này được mô hình hóa bằng:

$$h(t) = \mathbf{Z}(t)\boldsymbol{\kappa}$$

Với:

- $\mathbf{Z}(t) = [\mathbf{1}_{t \in D_1}, ..., \mathbf{1}_{t \in D_L}]$
- $\boldsymbol{\kappa} \sim \mathcal{N}(0, \nu^2)$

Có thể thêm hiệu ứng trước và sau ngày lễ để bao quát cả khoảng thời gian ảnh hưởng.

5.2.5 Huấn luyện mô hình

Toàn bộ mô hình được xây dựng và huấn luyện bằng công cụ **Stan** sử dụng thuật toán tối ưu L-BFGS. Prophet cung cấp hai dạng xu hướng (logistic hoặc tuyến tính), và fitting rất nhanh. Đặc biệt phù hợp để tương tác, thử nhiều cấu hình mô hình.

5.2.6 Hỗ trợ phân tích tương tác (Analyst-in-the-Loop)

Prophet được thiết kế để người dùng có thể dễ dàng điều chỉnh các thành phần sau:

- **Capacity:** xác định ngưỡng bão hòa $C(t)$.
- **Changepoints:** thêm các thời điểm thay đổi đã biết.

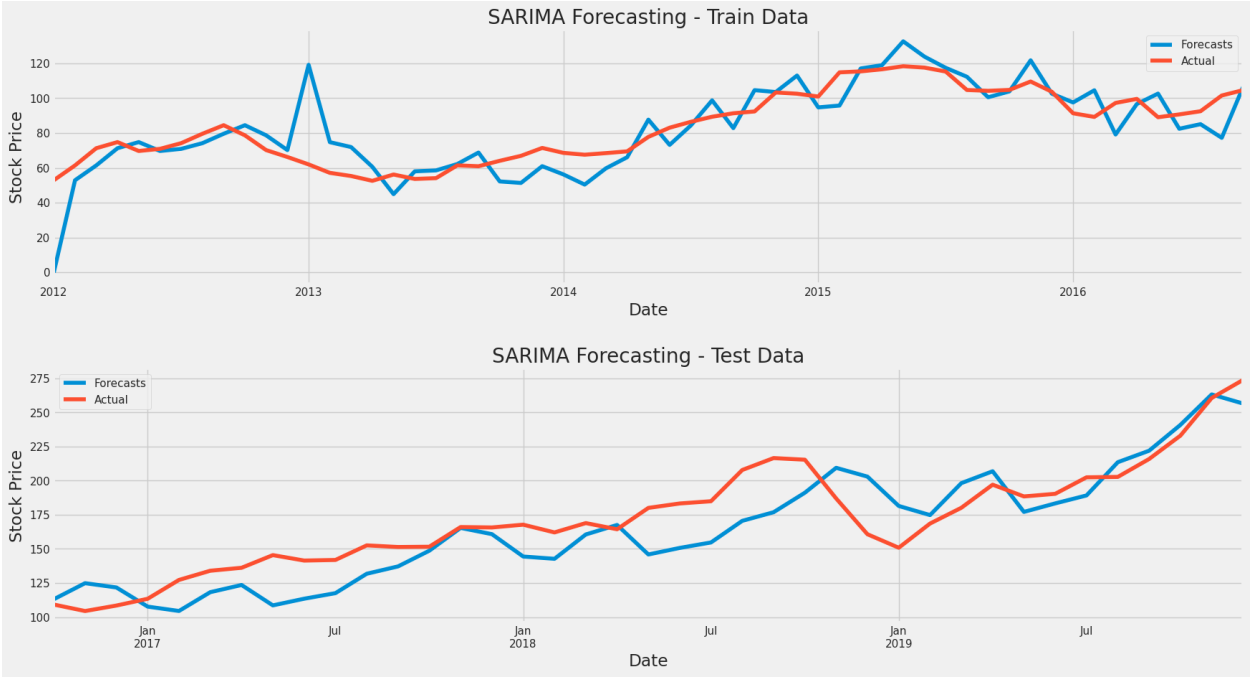
- **Holidays và Seasonality:** bổ sung sự kiện và quy luật mùa vụ đặc thù.
- **Tham số làm mượt:** điều chỉnh độ linh hoạt thông qua τ, σ, ν.

Việc này cho phép người dùng không chuyên về thống kê vẫn có thể can thiệp hiệu quả vào mô hình, nâng cao chất lượng dự báo.

6 Đánh Giá Kết Quả

6.1 SARIMA

Sau khi huấn luyện mô hình SARIMA, quá trình đánh giá mô hình được thực hiện qua nhiều chỉ số khác nhau để kiểm tra mức độ phù hợp và khả năng dự báo của mô hình. Chúng tôi đã kiểm tra hiện tượng overfitting và underfitting bằng cách so sánh độ khớp giữa tập huấn luyện (in-sample) và tập kiểm tra (out-of-sample). Kết quả cho thấy mô hình có độ khớp cân bằng giữa hai tập dữ liệu, không có hiện tượng học quá mức (overfitting) hay học thiếu (underfitting), chứng tỏ rằng mô hình có tính tổng quát tốt.



Hình 15: So sánh giữa giá trị thực tế và giá trị dự báo từ mô hình SARIMA

Metric	Train Data	Test Data
R^2 Score	0.491	0.704
Mean Squared Error (MSE)	199.567	441.120
Mean Absolute Error (MAE)	9.797	17.557
Mean Absolute Percentage Error (MAPE)	13.39%	10.65%
Accuracy (100 - MAPE)	86.61%	89.35%

Bảng 7: So sánh các độ đo đánh giá mô hình SARIMA trên tập huấn luyện và kiểm tra

Mặc dù có sự gia tăng nhẹ trong các lỗi đo lường như MSE và MAE khi dự báo trên tập test, các chỉ số như R^2, MAPE và Accuracy đều cho thấy mô hình duy trì được độ chính xác cao. Cụ thể, giá trị R^2 trên tập test đạt 0.704, cho thấy mô hình có khả năng giải thích khoảng 70% sự biến động của dữ liệu. MAPE và Accuracy trên tập test lần lượt là 10.65% và 89.35%, cho thấy mức độ dự báo tương đối chính xác và có thể áp dụng hiệu quả trong các dự báo thực tế.

Mặc dù các lỗi MSE và MAE tăng trên tập test, sự cải thiện về độ chính xác và giảm MAPE trên tập test chứng tỏ rằng mô hình không bị overfitting và có thể tổng quát tốt đối với dữ liệu mới. Điều này cũng cho thấy khả năng áp dụng mô hình trong dự báo các chuỗi thời gian chưa thấy trước.

Phương trình của mô hình SARIMA:

Để biểu diễn phương trình của mô hình SARIMA$(1,1,1) \times (2,2,[],12)$, ta sử dụng các tham số thu được từ bảng kết quả như sau:

- **Phần AR không mùa vụ**: $\phi_1 = 0.7677$
- **Phần MA không mùa vụ**: $\theta_1 = -0.5014$
- **Phần AR mùa vụ**: $\Phi_1 = -0.5420$, $\Phi_2 = -0.3440$
- **Phương sai của sai số**: $\sigma^2 = 92.9283$

Phương trình tổng quát

$$(1 - \phi_1 B)(1 - \Phi_1 B^{12} - \Phi_2 B^{24})(1 - B)(1 - B^{12})^2 y_t = (1 + \theta_1 B)\varepsilon_t \tag{1}$$

Thay các hệ số vào phương trình (1)

$$(1 - 0.7677B)(1 + 0.5420B^{12} + 0.3440B^{24})(1 - B)(1 - B^{12})^2 y_t = (1 - 0.5014B)\varepsilon_t \tag{2}$$

Khai triển chi tiết các toán tử

- Phần không mùa vụ:

$$(1 - 0.7677B)(1 - B) = 1 - 1.7677B + 0.7677B^2$$

- Phần mùa vụ:

$$(1 - B^{12})^2 = 1 - 2B^{12} + B^{24}$$

- Phần SAR (tự hồi quy mùa vụ):

$$1 + 0.5420B^{12} + 0.3440B^{24}$$

Phương trình đầy đủ sau khi khai triển

$$(1-1.7677B+0.7677B^2)(1+0.5420B^{12}+0.3440B^{24})(1-2B^{12}+B^{24})y_t = (1-0.5014B)\varepsilon_t \quad (3)$$

Ghi chú

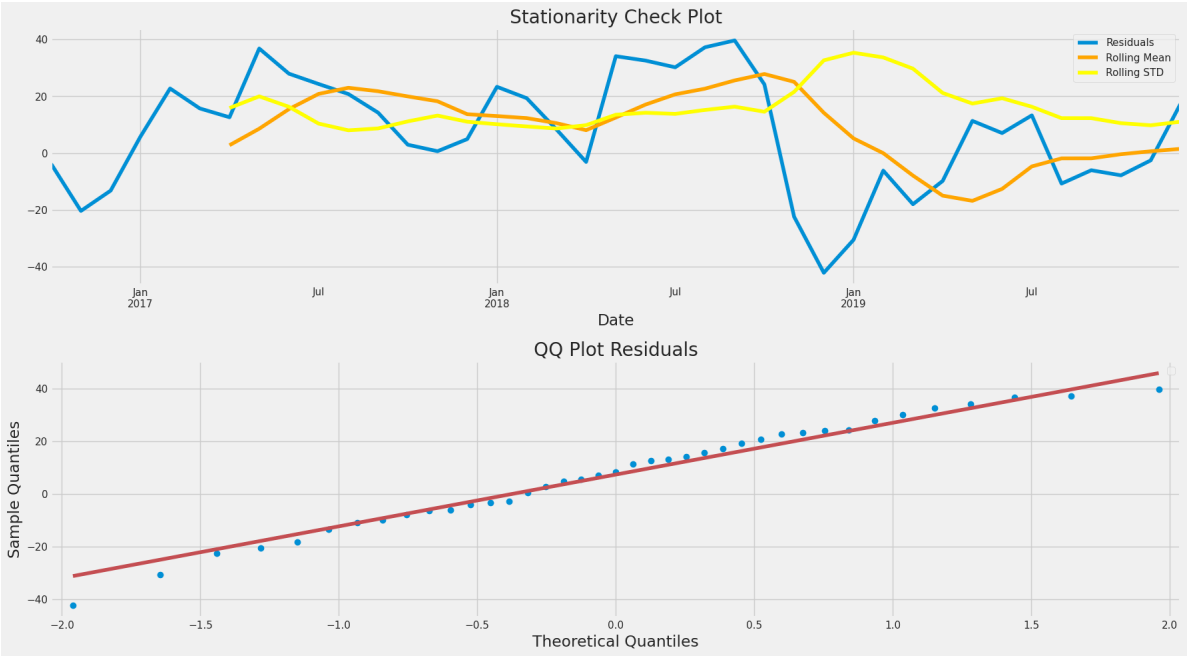
- y_t là chuỗi thời gian đã được sai phân:

$$y'_t = \nabla\nabla^2_{12} y_t$$

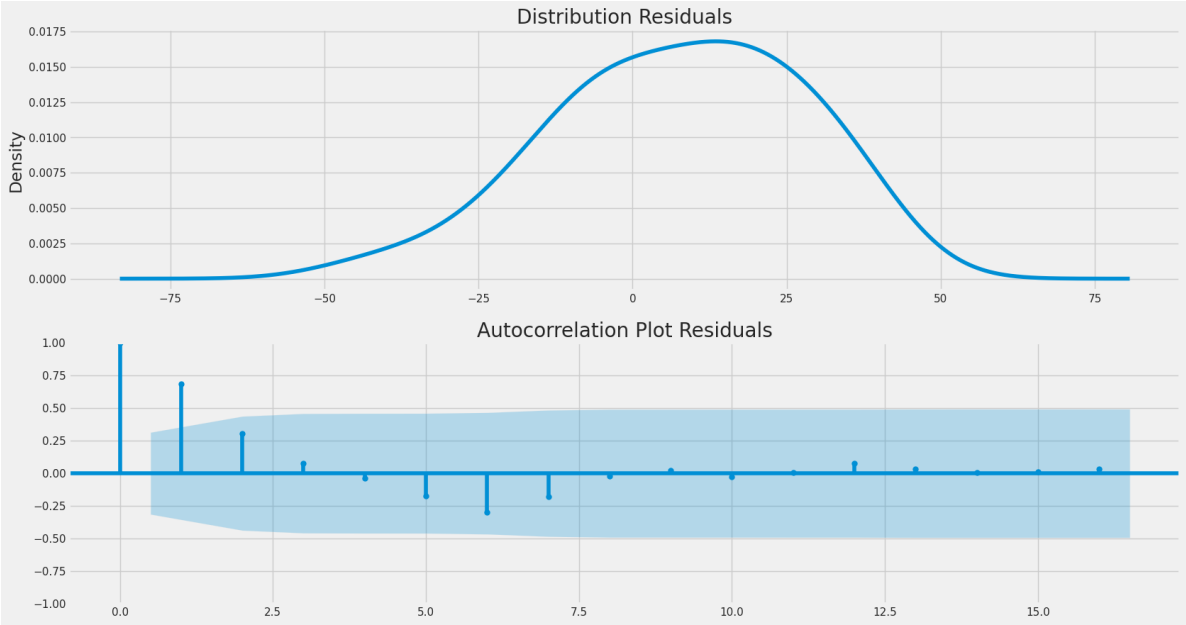
  tức là sai phân bậc 1 (không mùa vụ) và sai phân mùa vụ bậc 2 với chu kỳ 12.

- $\varepsilon_t \sim \mathcal{N}(0, \sigma^2 = 92.9283)$ là nhiễu trắng (sai số ngẫu nhiên tuân theo phân phối chuẩn).

Kiểm định phần dư



Hình 16: Biểu đồ kiểm tra tính dừng và phân phối chuẩn của chuỗi dư từ mô hình SARIMA



Hình 17: Biểu đồ phân phối và tự tương quan của chuỗi dư SARIMA.

Từ kết quả phân tích chuỗi dư, có thể thấy rằng mô hình SARIMA đã thực hiện việc dự báo khá tốt. Các residuals gần như tuân theo phân phối chuẩn, điều này cho thấy mô hình đã nắm bắt phần lớn sự biến động trong dữ liệu, với độ lệch nhỏ, không có hiện tượng bias đáng kể. Biểu đồ ACF của chuỗi dư không cho thấy sự tồn tại của mối tương quan tự động ở các độ trễ khác nhau, điều này xác nhận rằng các residuals là ngẫu nhiên, không có cấu trúc dư thừa mà mô hình chưa khai thác.

Sự ngẫu nhiên và phân phối chuẩn của các residuals chứng tỏ rằng mô hình đã loại bỏ được tất cả các mẫu tương quan có thể có trong dữ liệu, đồng thời không có thông tin quan trọng nào bị bỏ sót. Kết quả này củng cố thêm khả năng tổng quát của mô hình, cho thấy mô hình SARIMA không chỉ phù hợp với tập huấn luyện mà còn có thể áp dụng hiệu quả cho các chuỗi thời gian chưa quan sát.

Nhìn chung, mô hình SARIMA cho thấy khả năng dự báo ổn định và chính xác với các chỉ số đánh giá tốt trên cả tập train và tập test. Kết quả này phản ánh sự thành công trong việc phát triển một mô hình dự báo chuỗi thời gian có thể giải thích được các biến động quan trọng trong dữ liệu và có khả năng dự báo chính xác trong các tình huống thực tế.

Với khả năng dự báo cao, mô hình SARIMA có thể là một công cụ hữu ích trong việc dự đoán các chuỗi thời gian chưa quan sát, đặc biệt khi dữ liệu có tính chu kỳ, xu hướng, và yếu tố ngẫu nhiên như trong bài toán nghiên cứu của chúng tôi.

6.2 Prophet

Thực nghiệm được thực hiện bằng cách sử dụng thư viện Prophet trong Python để dự báo giá cổ phiếu dựa trên dữ liệu giá điều chỉnh hàng tháng (Adj Close). Chúng tôi tiến

hành hai giai đoạn chính: xây dựng một mô hình Prophet cơ bản và tối ưu hóa mô hình thông qua tìm kiếm siêu tham số.
Dữ liệu được chia thành hai tập:

- Mô hình cơ bản: 60% dữ liệu cho huấn luyện (57 tháng), 40% cho kiểm tra (39 tháng).
- Tìm kiếm siêu tham số: 70% dữ liệu cho huấn luyện, 30% cho kiểm tra, với tập kiểm tra tương ứng với khoảng thời gian từ 31/08/2017 đến 31/12/2019 (29 tháng).

a) **Mô hình cơ bản**

Mô hình Prophet cơ bản được thiết lập để dự báo giá cổ phiếu, tận dụng khả năng mô hình hóa xu hướng và mùa vụ của Prophet. Cấu hình chi tiết bao gồm:

- **Tham số mô hình:**
  - *Xu hướng:* Mô hình sử dụng xu hướng tuyến tính với các điểm thay đổi tự động (automatic changepoints), được điều chỉnh bởi phân phối Laplace:

$$\delta_j \sim \text{Laplace}(0, \tau)$$

Tham số `changepoint_prior_scale` được giữ ở giá trị mặc định ($\tau = 0.05$), đảm bảo mô hình đủ linh hoạt để phát hiện các thay đổi xu hướng lớn mà không quá khớp.

  - *Mùa vụ:*
    * **Mùa vụ hàng năm:** Kích hoạt bằng tham số `yearly_seasonality=True`, sử dụng chuỗi Fourier với số lượng thành phần mặc định ($N = 10$) để mô hình hóa các mô hình chu kỳ dài hạn:

$$s(t) = \sum_{n=1}^{10} \left(a_n \cos\left(\frac{2\pi nt}{365.25}\right) + b_n \sin\left(\frac{2\pi nt}{365.25}\right) \right)$$

Điều này phù hợp với dữ liệu hàng tháng, nơi các mô hình mùa vụ ngắn hạn (hàng tuần, hàng ngày) ít có ý nghĩa.

    * **Mùa vụ ngắn hạn:** Tắt mùa vụ hàng tuần (`weekly_seasonality=False`) và hàng ngày (`daily_seasonality=False`), vì dữ liệu có tần suất thấp (hàng tháng) và không có bằng chứng về chu kỳ ngắn hạn trong giá cổ phiếu.
    * **Ngày lễ:** Không sử dụng hiệu ứng ngày lễ (`holidays=None`), vì không có thông tin về các sự kiện tài chính cụ thể (như báo cáo thu nhập, khủng

hoảng thị trường). Do đó, thành phần ngày lễ được đặt bằng 0:

$$h(t) = 0$$

– *Chế độ mùa vụ:* Sử dụng chế độ cộng tính (`seasonality_mode='additive'`), nghĩa là mùa vụ được thêm trực tiếp vào xu hướng:

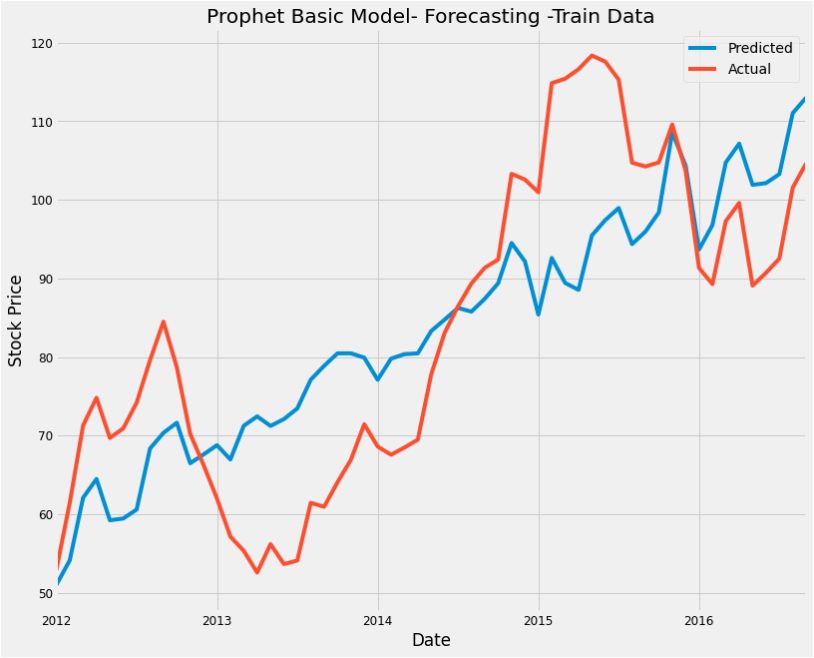
$$y(t) = g(t) + s(t) + \epsilon_t$$

Chế độ này giả định biên độ mùa vụ ổn định, không phụ thuộc vào mức giá cổ phiếu.

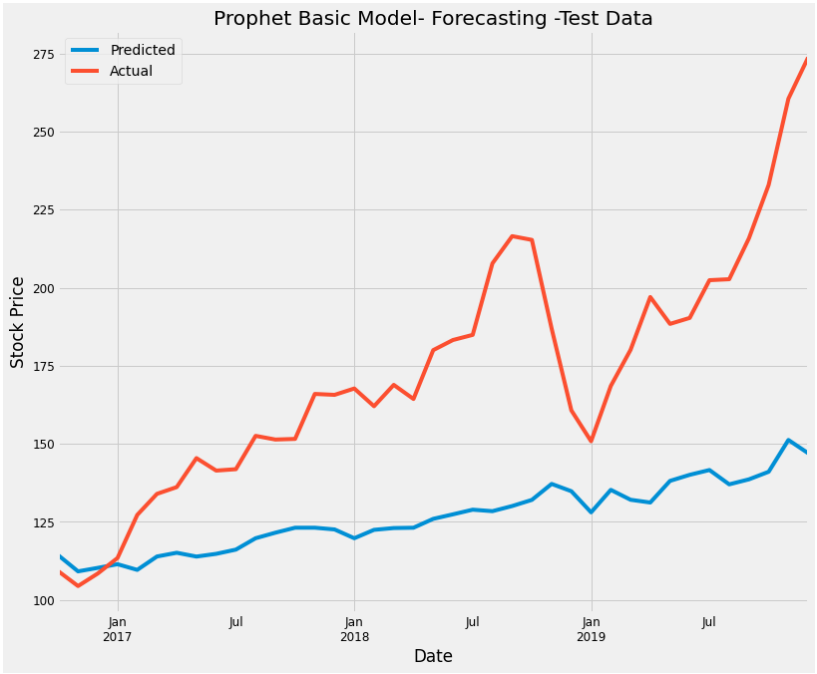
- **Dự báo:**
  - Tạo một khung dữ liệu bao gồm toàn bộ khoảng thời gian (57 tháng huấn luyện + 39 tháng kiểm tra).
  - Mô hình dự báo giá trị $\hat{y}(t)$ cho mỗi tháng:

$$\hat{y}(t) = g(t) + s(t)$$

  - Kết quả được căn chỉnh với chỉ số thời gian của dữ liệu gốc để so sánh trực tiếp với giá trị thực tế y.
  - Dự báo được hình ảnh hóa bằng biểu đồ so sánh giá trị dự báo ($\hat{y}$) và thực tế (y) trên cả tập huấn luyện và kiểm tra



Hình 18: Biểu đồ so sánh giá trị dự báo và thực tế trên tập huấn luyện



Hình 19: Biểu đồ so sánh giá trị dự báo và thực tế trên tập kiểm tra

Metric	**Train Data**	**Test Data**
R^2 Score	0.598	-0.973
Mean Squared Error (MSE)	157.648	2936.426
Mean Absolute Error (MAE)	10.794	46.175
Mean Absolute Percentage Error (MAPE)	13.785%	24.535%
Accuracy (100 - MAPE)	86.0%	75.0%

Bảng 8: Các độ đo đánh giá mô hình Prophet trên tập train và tập test

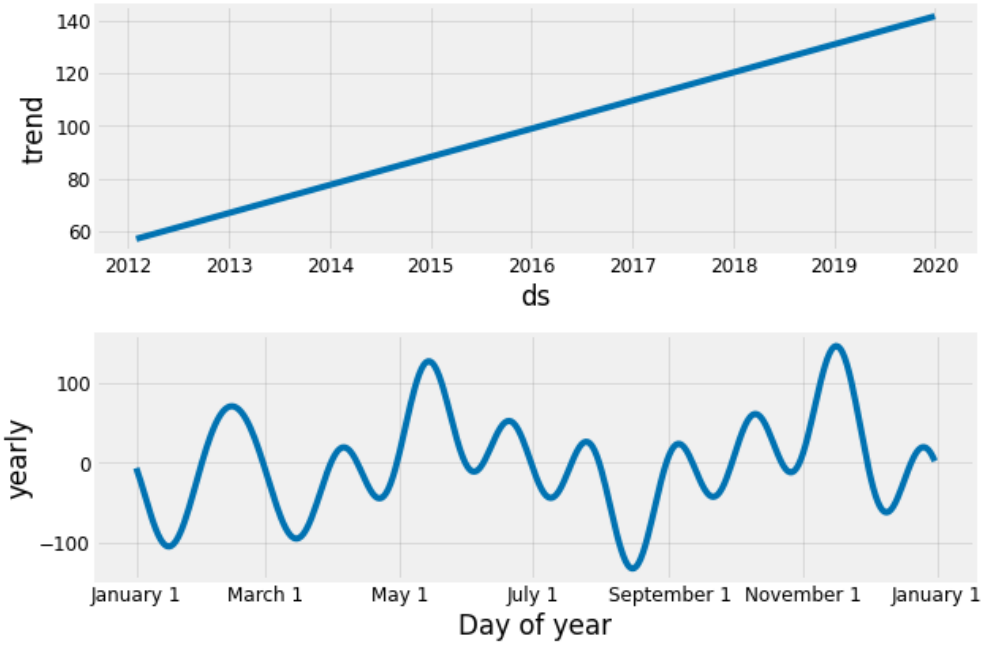
Ta nhận thấy Mô hình Prophet cơ bản với các tham số mặc định bị overfit, điều này thể hiện qua sự chênh lệch độ chính xác giữa tập huấn luyện và tập kiểm tra.

- Xu hướng:

  Giá cổ phiếu có xu hướng tăng đều qua thời gian, từ 60 (2012) lên 140 (2020), tương ứng với tốc độ tăng trung bình khoảng 10 đơn vị mỗi năm.

  Đường xu hướng không hiển thị các điểm thay đổi rõ ràng (*changepoints*), tức là những thời điểm mà tốc độ tăng trưởng thay đổi (ví dụ: từ tăng sang giảm, hoặc chuyển sang tăng nhanh hơn).

  Điều này có thể do tham số `changepoint_prior_scale` mặc định ($\tau = 0.05$) quá nhỏ, khiến mô hình ưu tiên một đường xu hướng mượt mà thay vì phát hiện các thay đổi đột ngột.



Hình 20: Biểu đồ xu hướng và mùa vụ hàng năm

Hệ quả: Mô hình có thể bỏ sót các sự kiện lớn ảnh hưởng đến giá cổ phiếu, chẳng hạn như khủng hoảng tài chính (ví dụ: giai đoạn 2015–2016) hoặc các giai đoạn tăng trưởng đột biến.

- Mùa vụ hàng năm:

  Biến động mạnh quanh tháng 5 và tháng 11: Dễ thấy có những đỉnh cao (tăng mạnh) vào khoảng đầu tháng 5 và tháng 11, trong khi các đáy (giảm mạnh) xuất hiện vào khoảng giữa tháng 9 và đầu tháng 1.

  Điều này có thể phản ánh các đặc điểm đặc thù của thị trường chứng khoán, như quý báo cáo tài chính, sự kiện kinh tế, hoặc thói quen giao dịch theo mùa.

b) **Tìm kiếm siêu tham số**

Để cải thiện hiệu suất của mô hình Prophet cơ bản, chúng em tiến hành tối ưu hóa siêu tham số thông qua phương pháp tìm kiếm lưới (grid search). Quá trình này nhằm tìm ra cấu hình tối ưu giúp giảm sai số dự báo, đặc biệt trên tập kiểm tra, đồng thời đảm bảo mô hình đủ linh hoạt để nắm bắt các đặc điểm phức tạp của giá cổ phiếu (Adj Close).

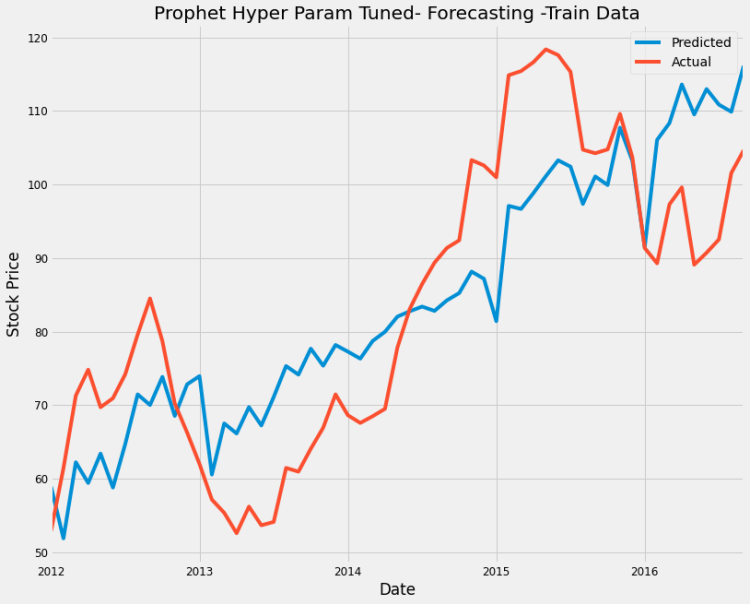
Chúng em khởi tạo lưới tham số bao gồm 16 tổ hợp, tương ứng với 16 mô hình Prophet khác nhau được thử nghiệm. Siêu tham số tối ưu: Sau khi thử nghiệm toàn bộ 16 tổ hợp trong lưới tham số, mô hình Prophet với hiệu suất dự báo tốt nhất có cấu hình như sau:

- **changepoint_prior_scale** = 0.4

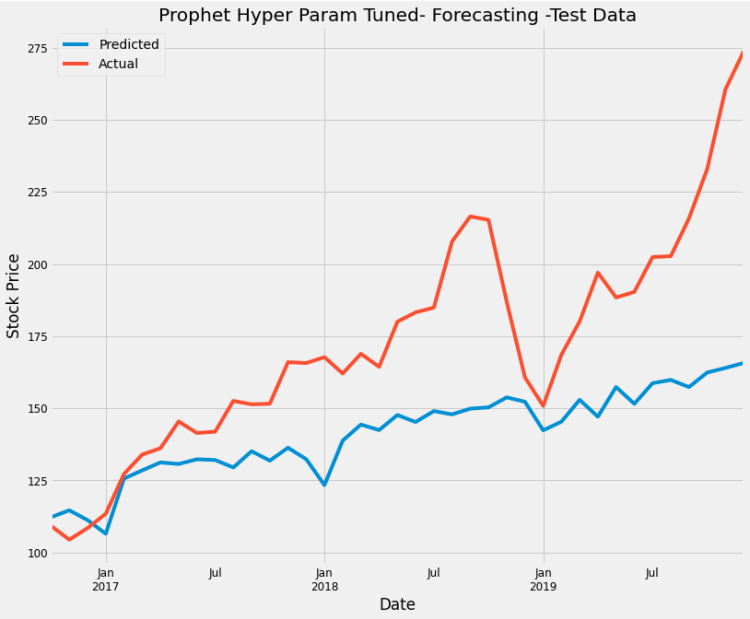
- **holidays_prior_scale** = 0.4
- **n_changepoints** = 50
- **seasonality_mode** = `"additive"`

c) **Mô hình tối ưu:**

Thực thi mô hình Prophet với các siêu tham số mới ở mục (b) được kết quả như sau:



Hình 21: Biểu đồ so sánh giá trị dự báo và thực tế trên tập huấn luyện



Hình 22: Biểu đồ so sánh giá trị dự báo và thực tế trên tập kiểm tra

Nhận xét:

- Biểu đồ tập huấn luyện:

  Đường màu xanh dương (Predicted) khớp tương đối tốt với đường màu đỏ (Actual), đặc biệt là từ năm 2013 đến 2015.

  Tuy nhiên, một số thời điểm mô hình vẫn chưa nắm bắt được các đỉnh/đáy đột ngột, cho thấy khả năng bắt trend ngắn hạn còn hạn chế.

- Biểu đồ tập kiểm tra:

  Mô hình không theo kịp sự tăng mạnh của giá thực tế từ giữa năm 2018 đến 2020, dẫn đến độ lệch đáng kể giữa giá dự báo và giá thực tế.

  Điều này cho thấy mô hình Prophet, mặc dù đã được tinh chỉnh, vẫn có xu hướng dưới dự báo (underestimate) trong giai đoạn thị trường tăng mạnh.

Metric	**Train Data**	**Test Data**
R^2 Score	0.784	-1.31
Mean Squared Error (MSE)	129.543	2187.789
Mean Absolute Error (MAE)	9.986	40.501
Mean Absolute Percentage Error (MAPE)	12.076%	20.338%
Accuracy (100 - MAPE)	88.0%	80.0%

Bảng 9: Các độ đo đánh giá mô hình Prophet tối ưu trên tập train và tập test

Nhận xét:

- Mô hình Prophet tối ưu bằng Grid Search đã fit tốt trên tập huấn luyện nhưng lại overfitting dẫn đến hiệu suất kém trên tập kiểm tra.
- Điều này cho thấy Prophet có thể không phù hợp để mô hình hóa các chuỗi thời gian với biến động mạnh và xu hướng phi tuyến tính dài hạn như giá cổ phiếu.
- Mặc dù đã tinh chỉnh, Prophet vẫn hạn chế trong việc nắm bắt xu hướng tăng đột ngột hoặc các pha biến động mạnh của thị trường.

Biểu thức toán học của mô hình này là:

$$y(t) = \left(\left(k + \sum_{j:t \geq s_j} \delta_j\right) t + \left(m + \sum_{j:t \geq s_j} \gamma_j\right)\right) + \sum_{n=1}^{10} \left(a_n \cos\left(\frac{2\pi n t}{365.25}\right) + b_n \sin\left(\frac{2\pi n t}{365.25}\right)\right) + \epsilon_t$$

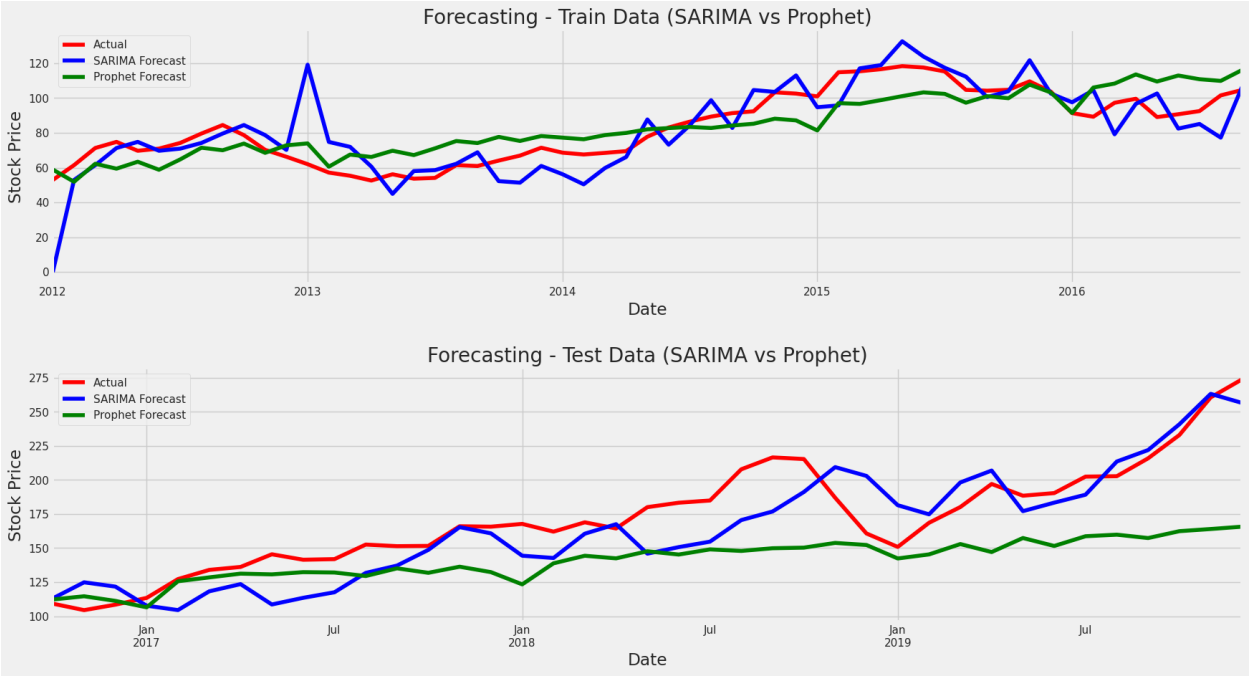
Trong đó:

- $y(t)$: Giá trị dự báo tại thời điểm t (giá cổ phiếu `Adj Close`).
- k: Tốc độ tăng trưởng cơ bản.
- m: Độ lệch.
- s_j: Các điểm thay đổi (changepoints).
- δ_j: Điều chỉnh tốc độ tăng trưởng tại điểm thay đổi j.
- γ_j: Điều chỉnh độ lệch tại điểm thay đổi j.
- a_n, b_n: Hệ số Fourier của thành phần mùa vụ hàng năm.
- ϵ_t: Thành phần lỗi, giả định phân phối chuẩn $\epsilon_t \sim \mathrm{N}(0, \sigma_\epsilon^2)$.

Các đặc trưng cơ bản:

- Kỳ vọng: 107.1827
- Phương sai: 1058.0478
- Hiệp phương sai: 1606.9842
- Hệ số tương quan: 0.9397

6.3 So sánh kết quả của hai mô hình SARIMA và Prophet



Hình 23: Biểu đồ so sánh dự báo giá cổ phiếu bằng mô hình SARIMA và Prophet trên tập train và tập test

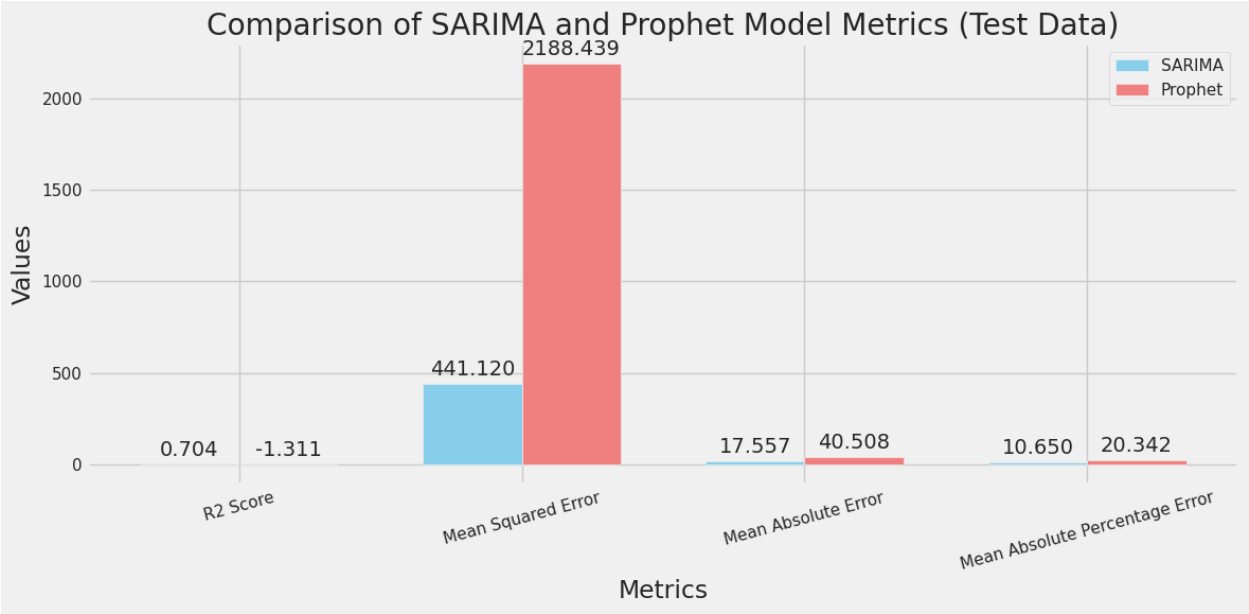
Phân tích hiệu suất dự báo của hai mô hình SARIMA và Prophet dựa trên biểu đồ cho thấy các đặc điểm sau:
Trên dữ liệu huấn luyện (2012-2016):

- Mô hình SARIMA thể hiện độ biến động cao hơn, với các đỉnh và đáy dự báo đạt mức cực đoan, phản ánh sự nhạy cảm với các biến động ngắn hạn trong dữ liệu.
- Ngược lại, Prophet tạo ra đường dự báo mượt mà và ổn định hơn, ít bị ảnh hưởng bởi các dao động ngắn hạn, duy trì tính liên tục trong xu hướng.
- Mặc dù SARIMA có khả năng bám sát các đỉnh giá thực tế tốt hơn, nhưng ở một số thời điểm, mô hình này dự báo thái quá, dẫn đến sai lệch cục bộ.

Trên dữ liệu kiểm tra (2017-2019):

- Prophet liên tục dự báo thấp hơn so với giá trị thực tế, đặc biệt khi giá cổ phiếu có xu hướng tăng mạnh, cho thấy mô hình không bắt kịp được các biến động lớn.
- SARIMA, trong khi đó, bám sát xu hướng thực tế tốt hơn, nhưng vẫn tồn tại độ trễ nhất định, khiến dự báo chưa hoàn toàn khớp với dữ liệu thực.
- Khoảng cách giữa giá trị dự báo của Prophet và giá trị thực tế ngày càng gia tăng về cuối giai đoạn kiểm tra, đặc biệt khi xu hướng tăng mạnh, làm nổi bật hạn chế của mô hình trong việc dự báo dài hạn.

Mô hình SARIMA thể hiện hiệu suất vượt trội hơn trong việc nắm bắt xu hướng giá tăng, đặc biệt trên tập kiểm tra, nhờ khả năng phản ánh các biến động lớn của dữ liệu. Ngược lại, Prophet tuy có ưu điểm về tính ổn định trong dự báo, nhưng lại không thể hiện tốt trong việc dự báo mức độ tăng giá mạnh, dẫn đến sai lệch lớn trên tập kiểm tra. Nhìn chung, SARIMA tỏ ra hiệu quả hơn trong trường hợp này, đặc biệt khi dự báo xu hướng tăng giá mạnh, dù có phần nhiễu hơn trên tập huấn luyện.



Hình 24: Biểu đồ so sánh các độ đo của 2 mô hình SARIMA và Prophet

Hệ số xác định (R^2) của mô hình SARIMA đạt giá trị 0.704, biểu thị khả năng giải thích khoảng 70.4% biến thiên trong dữ liệu kiểm tra. Ngược lại, mô hình Prophet thể hiện R^2 âm (-1.311), cho thấy mô hình này thậm chí còn kém hiệu quả hơn so với việc sử dụng giá trị trung bình đơn giản làm dự báo.

Các độ đo sai số cũng nhất quán ủng hộ hiệu suất vượt trội của SARIMA. Mean Squared Error của SARIMA (441.120) thấp hơn đáng kể so với Prophet (2188.439), chỉ bằng khoảng 20% MSE của Prophet. Tương tự, Mean Absolute Error của SARIMA (17.557) thấp hơn gần 57% so với Prophet (40.508).

Đặc biệt quan trọng đối với các ứng dụng thực tế, MAPE của SARIMA (10.650%) thấp hơn đáng kể so với Prophet (20.342%), cho thấy sai số tương đối trung bình của SARIMA thấp hơn gần 48%.

Kết luận

Kết quả nghiên cứu cho thấy mô hình SARIMA là lựa chọn phù hợp hơn đáng kể so với Prophet trong trường hợp dự báo bộ dữ liệu chuỗi thời gian này. Các độ đo hiệu suất đều nhất quán ủng hộ SARIMA với sự chênh lệch đáng kể trên tất cả các thước đo được áp dụng.

Tuy nhiên, cần lưu ý rằng kết quả này có thể đặc thù cho bộ dữ liệu cụ thể được phân tích. Trong các ứng dụng thực tế, việc lựa chọn giữa SARIMA và Prophet nên dựa trên đặc điểm cụ thể của dữ liệu, yêu cầu dự báo và tài nguyên tính toán sẵn có. Các nghiên cứu trong tương lai có thể mở rộng phân tích này để xem xét hiệu suất của các mô hình này trên nhiều loại chuỗi thời gian khác nhau và trong các khoảng thời gian dự báo dài hơn.

7 Tài Liệu Tham Khảo

Tài liệu

[1] Dalal, M., Li, A. C., & Taori, R. (2019). *Autoregressive Models: What Are They Good For?* Accepted for the Information Theory and Machine Learning workshop at NeurIPS 2019. arXiv preprint arXiv:1910.07737. Available at: https://doi.org/10.48550/arXiv.1910.07737

[2] Marek, T. (2005). *On Invertibility of a Random Coefficient Moving Average Model.* Kybernetika, 41(6), 743–756. Available at: https://www.kybernetika.cz/content/2005/6/743/paper.pdf

[3] Hasan, M., Wathodkar, G., & Muia, M. (2023). *ARMA Model Development and Analysis for Global Temperature Uncertainty.* Department of Mathematics, University of Mississippi, Oxford, MS, USA. Available at: https://arxiv.org/pdf/2303.02070

[4] Mbaye, A., Ndiaye, M., Ndione, D. M., Diaw, M., Traoré, V., Ndiaye, A., Sylla, M. C., Aidara, M. C., Diaw, V., Traoré, A., et al. (n.d.). *ARMA model for short-term forecasting of solar potential: Application to a horizontal surface on Dakar site.* Available at: https://hal.science/hal-02151290/document

[5] Siami-Namini, S., & Siami Namin, A. (2018). *Forecasting Economics and Financial Time Series: ARIMA vs. LSTM.* arXiv preprint arXiv:1803.06386. Available at: 10.48550/arXiv.1803.06386

[6] Wang, S., Li, C., & Lim, A. (2021). *Why Are the ARIMA and SARIMA not Sufficient.* arXiv preprint arXiv:1904.07632. Available at: 10.48550/arXiv.1904.07632

[7] Noor, T. H., Almars, A. M., Alwateer, M., Almaliki, M., Gad, I., & Atlam, E. (2022). *SARIMA: A Seasonal Autoregressive Integrated Moving Average Model for Crime Analysis in Saudi Arabia. Electronics*, 11(23), 3986. Available at: 10.3390/electronics11233986

[8] Diebold, F. X. (2006). *Forecasting and Time Series Analysis.* 2nd Edition. Thomson South-Western. Available at: https://www.researchgate.net/publication/261307350_SARIMA_Seasonal_ARIMA_implementation_on_time_series_to_forecast_the_number_of_Malaria_incidence

[9] Žunić, E., Korjenić, K., Hodžić, K., & Đonko, D. (2022). *Application of Facebook's Prophet Algorithm for Successful Sales Forecasting Based on Real-World Data.* Info Studio d.o.o. Sarajevo, Bosnia and Herzegovina; Faculty of Electrical Engineering, University of Sarajevo, Bosnia and Herzegovina. Available at: https://arxiv.org/pdf/2005.07575

[10] Sharma, K., Bhalla, R., & Ganesan, G. (2022). *Time Series Forecasting Using FB-Prophet*. Lovely Professional University, Punjab, India; Jain (Deemed-to-be University), Bengaluru, India. Available at: https://ceur-ws.org/Vol-3445/PAPER_07.pdf

[11] Gnanasekaran, H., Prabakar, D., Köse, U. (2023). *Time-series Forecasting of Web Traffic Using Prophet Machine Learning Model*.Dhinakaran Prabakar's Lab. Available at: https://www.researchgate.net/publication/376596105_Time-series_Forecasting_of_Web_Traffic_Using_Prophet_Machine_Learning_Model

[12] Taylor, S. J., & Letham, B. (2018). *Forecasting at Scale*. Facebook, Menlo Park, California, United States. Available at: https://peerj.com/preprints/3190.pdf