

Ứng dụng mô hình Time-Series trong dự báo giá chứng khoán FPT

Thành viên:
Trần Văn Tuấn
Dương Chí Vinh
Trần Thị Ngọc Anh
Võ Đức Mẫn

Giáo viên hướng dẫn:
TS. Cao Văn Lợi



Bài toán kinh doanh

Tại sao phải dự báo giá chứng khoán?
Giá chứng khoán có dễ dự đoán không?



Thu thập dữ liệu - PP đánh giá

Phân tích khai phá dữ liệu

Mô hình và kết quả

Hướng tiếp theo

Demo

1. Thu thập dữ liệu - PP đánh giá

- Dữ liệu: mã cổ phiếu nguồn số liệu
- Metric đánh giá: MAPE

2. Phân tích khai phá dữ liệu

- Xem xét xu hướng và tính chu kì
- Tự tương quan, tự tương quan riêng phần
- Tính dừng - tính không dừng

3. Lựa chọn mô hình & kết quả

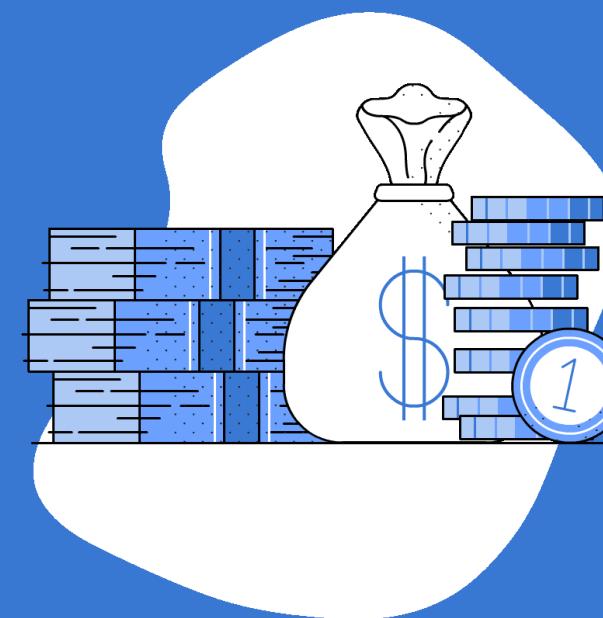
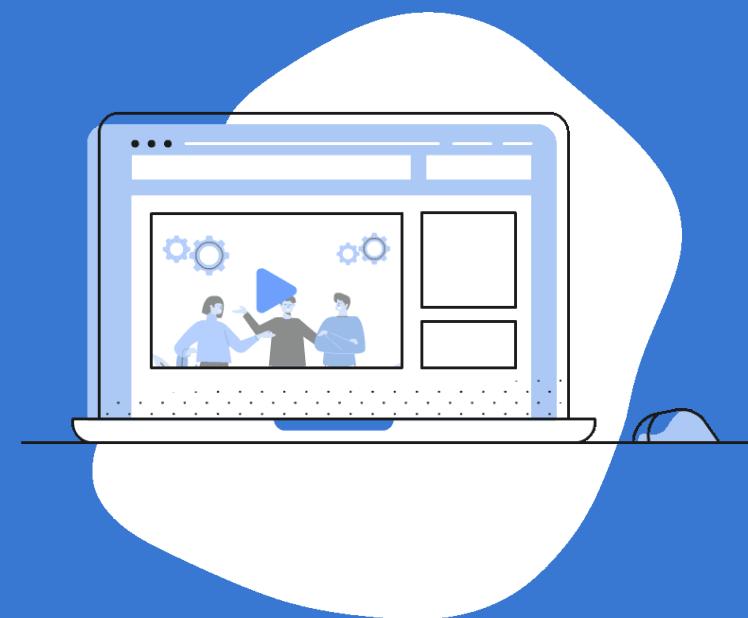
- ARIMA
- Ensemble learning
- Prophet
- LSTM

4. Hướng nghiên cứu tiếp theo

- Bổ sung biến, confidence intervals
- Điều chỉnh mô hình....

5. Demo sản phẩm

1.Thu thập dữ liệu & Phương pháp đánh giá



Mã cổ phiếu

Thuộc rổ VN30:

- Thanh khoản cao
- Vốn hóa lớn
- DN đầu ngành (bluechip)



Các trường dữ liệu

- Open
- Low
- High
- Close
- Volume

Time	Open	High	Low	Close	Volume
20201216	57.2	57.5	57.0	57.1	1995530
20201215	57.5	57.6	56.6	56.7	26666900
20201214	57.0	58.0	56.9	57.2	1443760
20201211	56.6	57.0	56.3	57.0	1750270
20201210	57.0	57.7	56.5	56.5	2112490

Nguồn dữ liệu

cophieu68.com



Mã cổ phiếu

Thuộc rổ VN30:

- Thanh khoản cao
- Vốn hóa lớn
- DN đầu ngành (bluechip)



Các trường dữ liệu

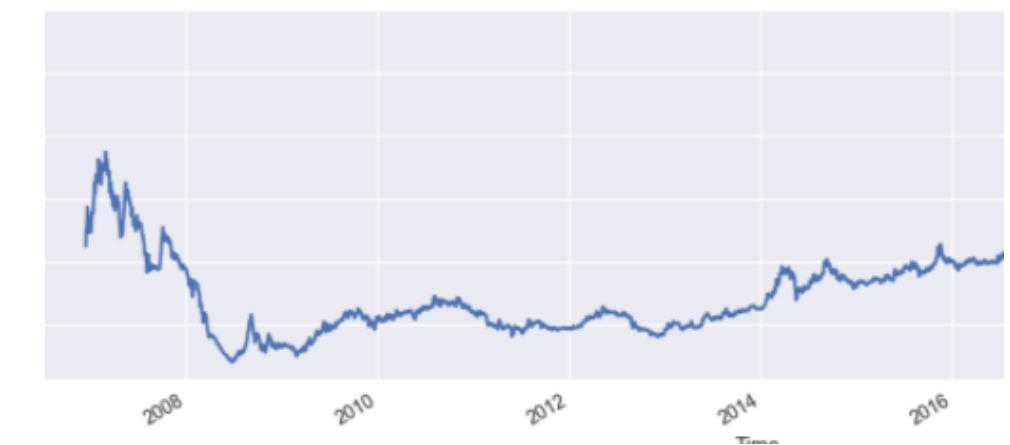
- Open
- Low
- High
- Close
- Volume

Time	Open	High	Low	Close	Volume
20201216	57.2	57.5	57.0	57.1	1995530
20201215	57.5	57.6	56.6	56.7	26666900
20201214	57.0	58.0	56.9	57.2	1443760
20201211	56.6	57.0	56.3	57.0	1750270
20201210	57.0	57.7	56.5	56.5	2112490

Thời gian

14 năm

- Start: 12/2006
- End: 12/2020



Phân chia tập dữ liệu

Chia theo trình tự thời gian

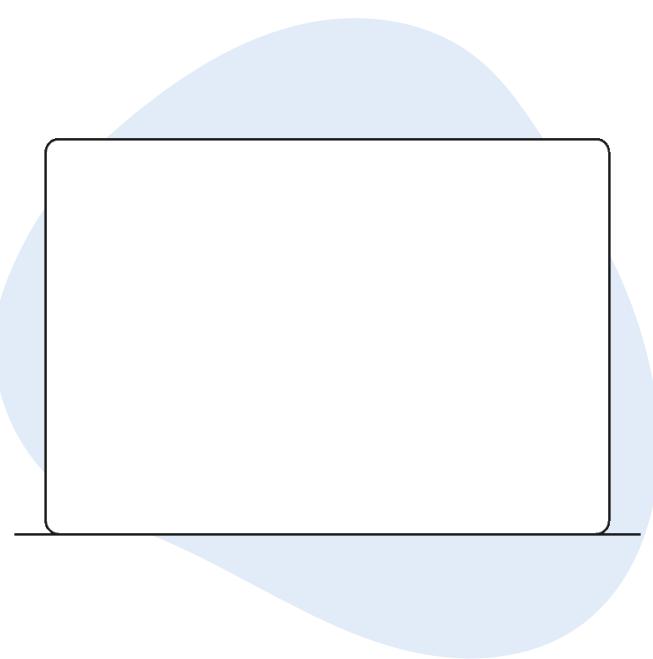
- Tập huấn luyện
- Tập kiểm định
- Tập kiểm tra

Chuẩn đánh giá: MAPE

- Dùng để đánh giá mô hình
- Thể hiện phần trăm sai số so với giá trị thực tế



2. Phân tích khai phá dữ liệu



2.1 Yếu tố xu hướng (Trend)

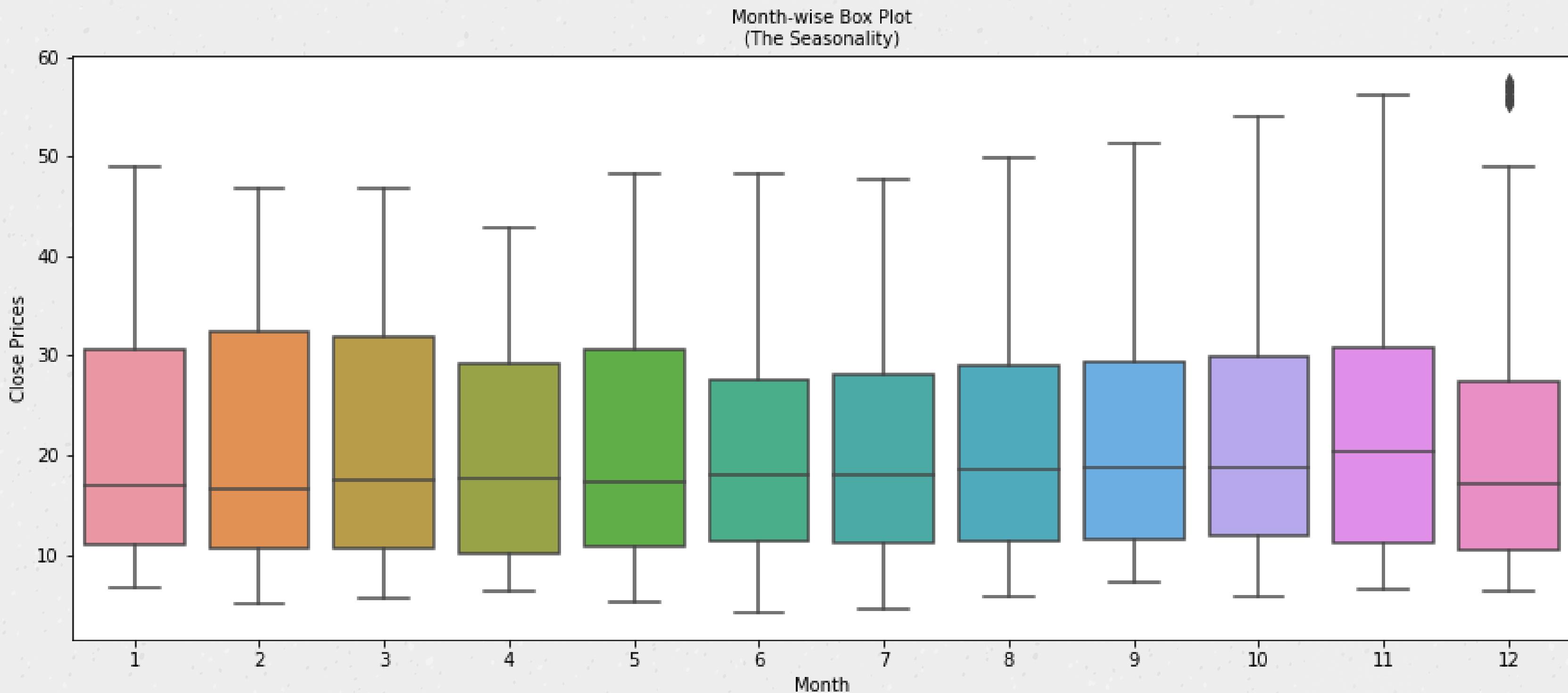


12/2016
- 06/2018

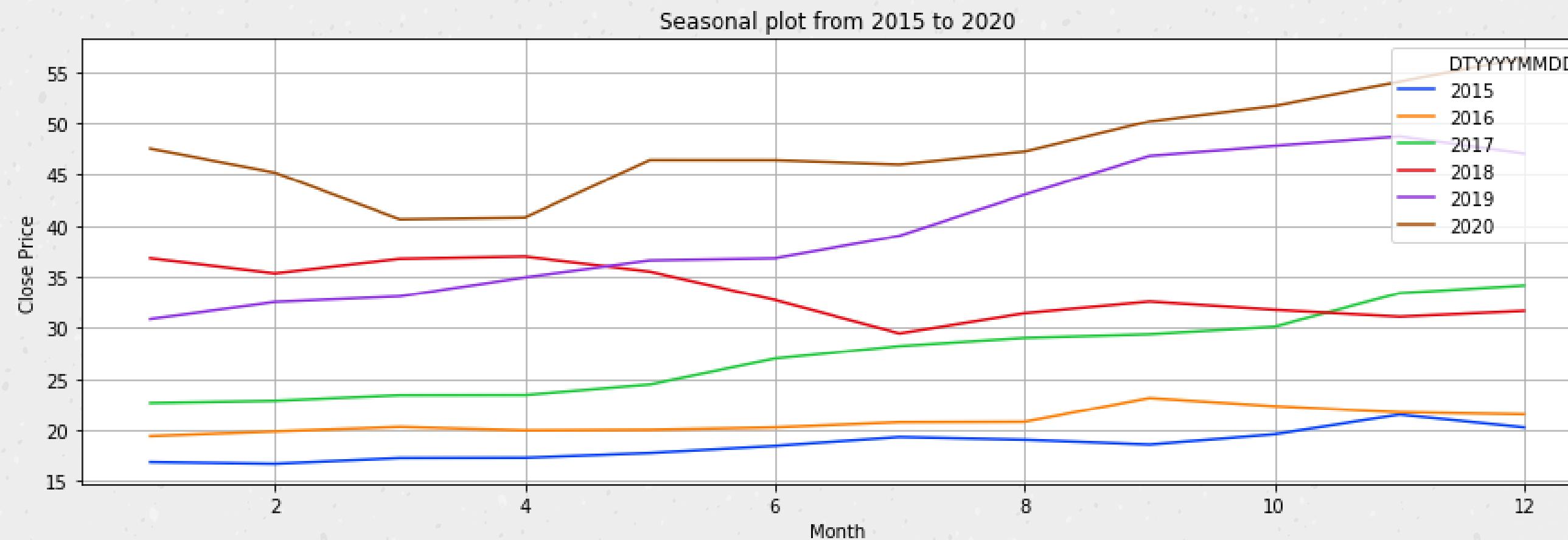
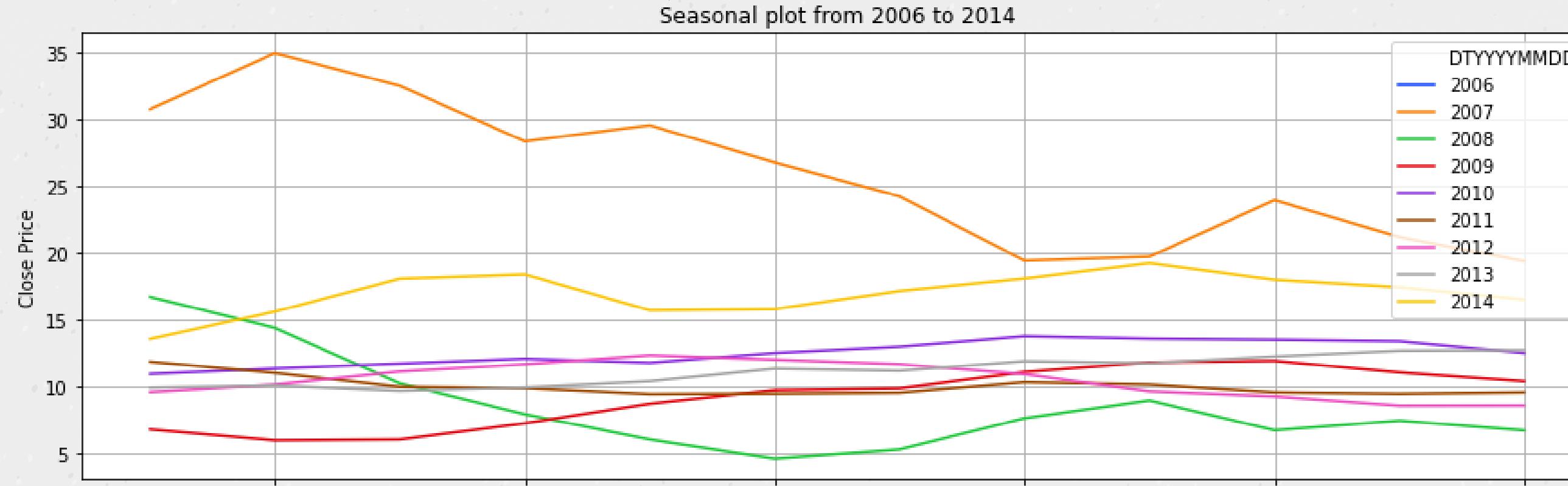
06/2008 - 12/2017

01/2018 -
12/2020

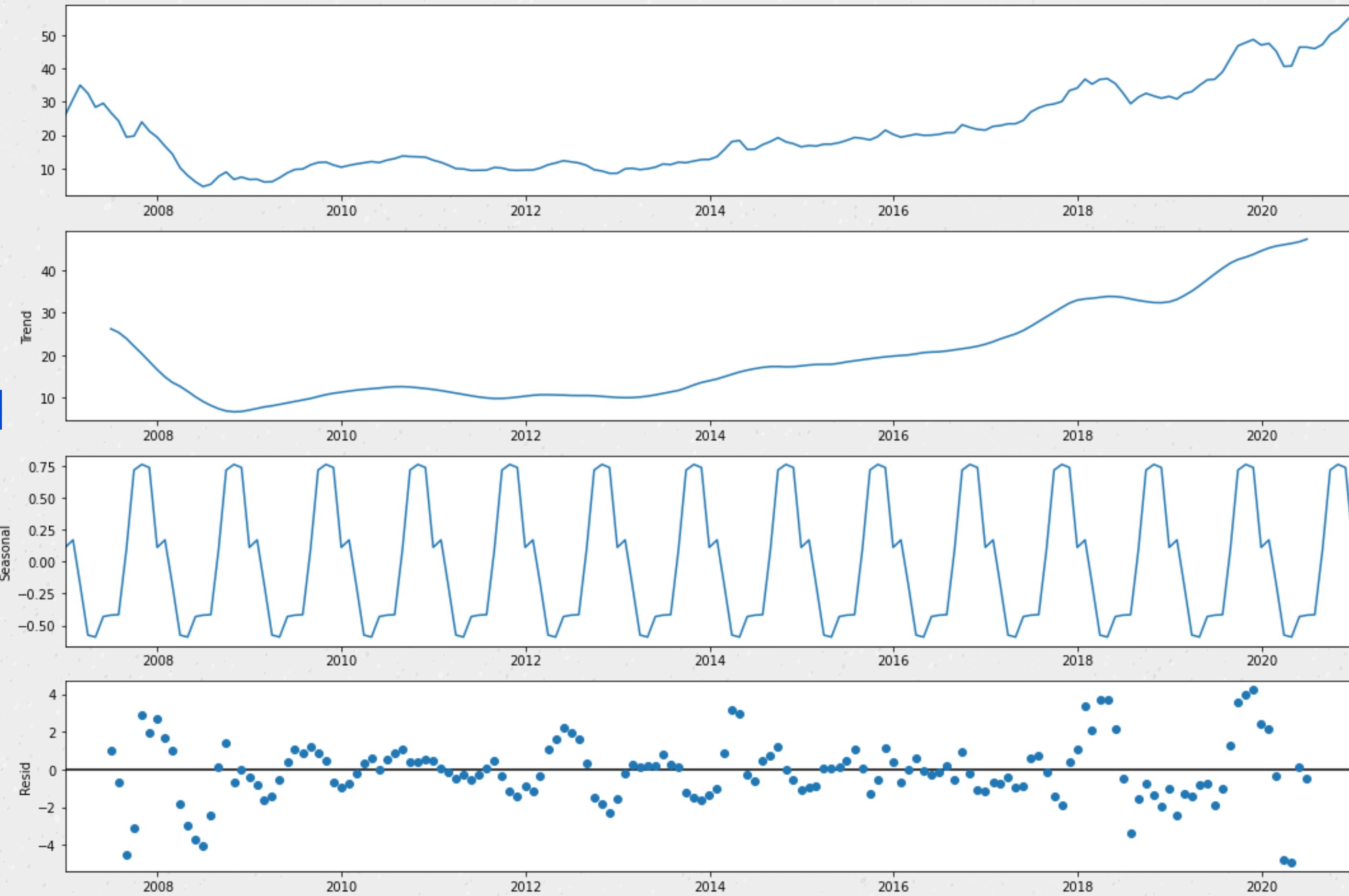
2.1 Yếu tố chu kỳ, thời vụ (Seasonal)



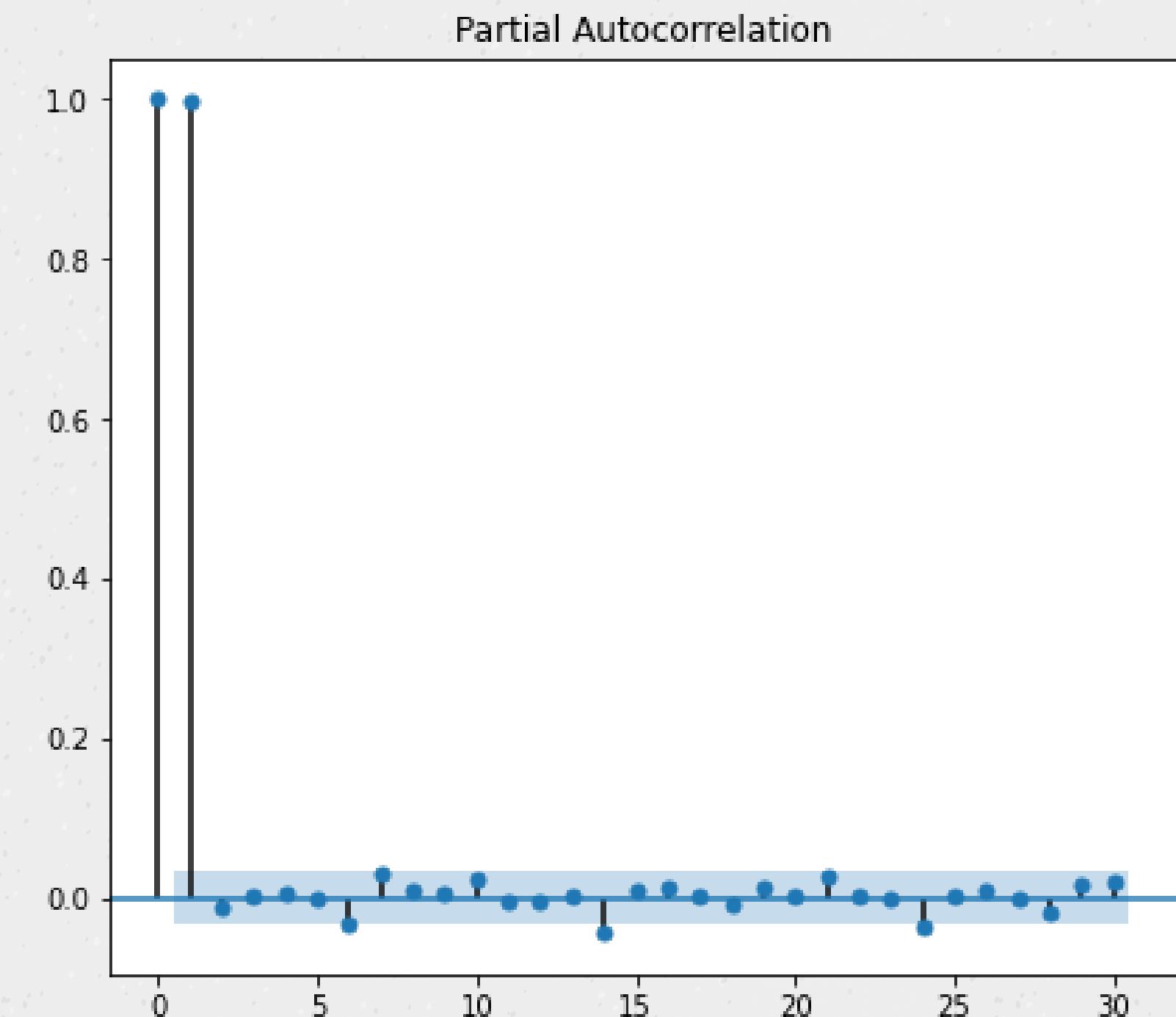
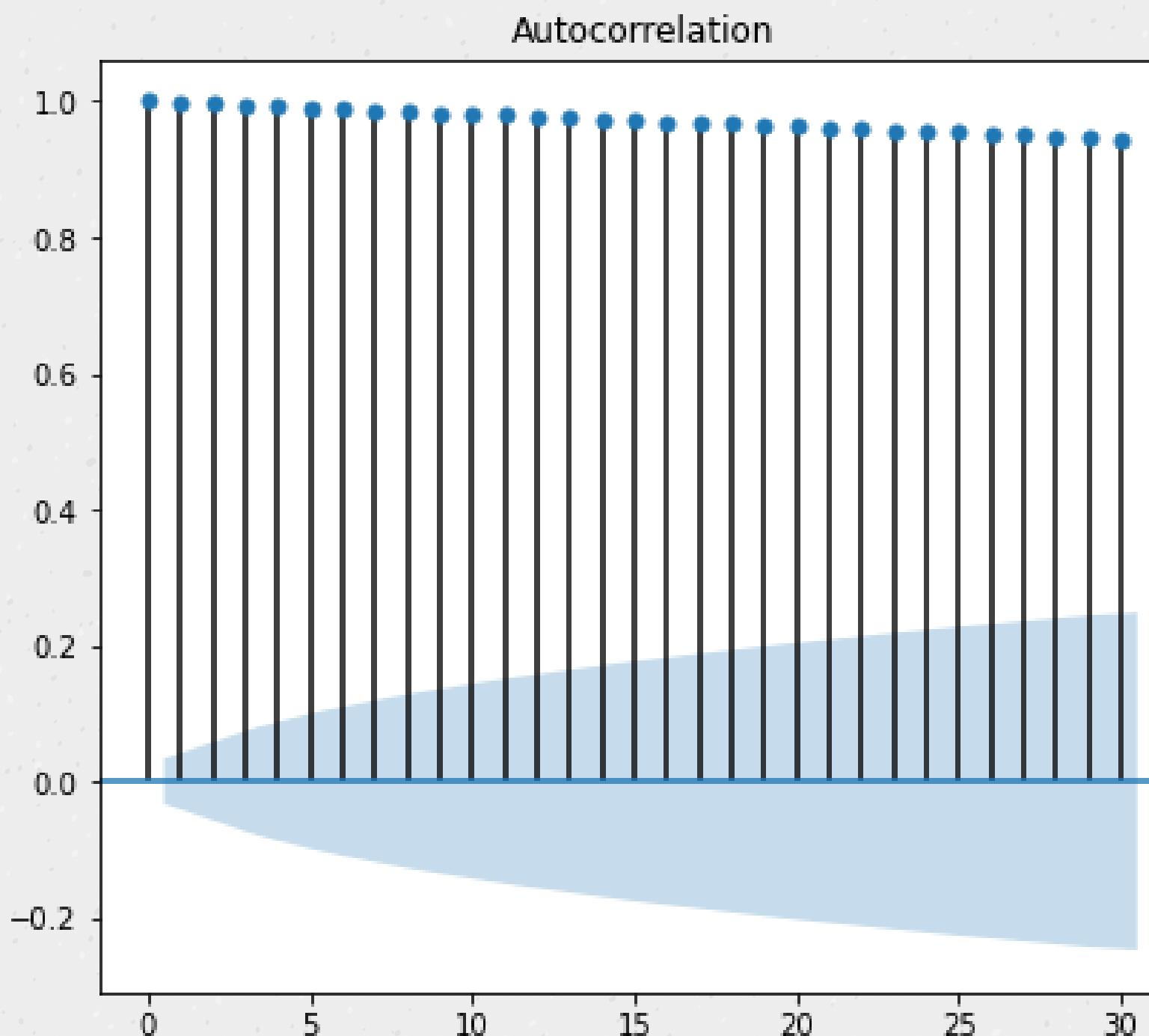
2.1 Yếu tố chu kỳ, thời vụ (Seasonal)



Phân tách (Trend - Seasonal decomposition)



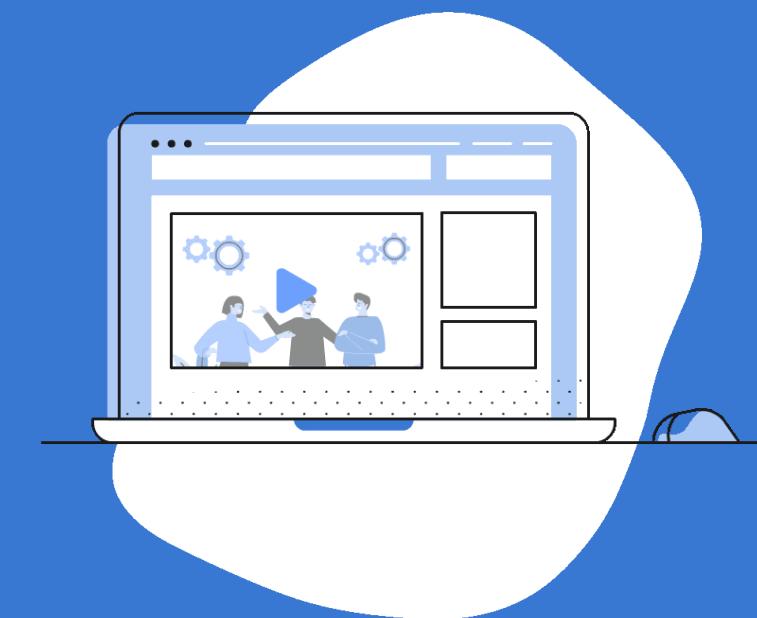
2.2 Tự tương quan (ACF) Tự tương quan riêng phần (PACF)



2.3 Tính dừng (Stationary) Tính không dừng (Non-stationary)

Kiểm định	Kiểm định Dickey và Fuller mở rộng (ADF)	Kiểm định Kwiatkowski, Phillips, Schmidt and Shin (KPSS)
Giả thuyết - Đối thuyết	<ul style="list-style-type: none">• H0: non-stationary• H1: stationary	<ul style="list-style-type: none">• H0: stationary• H1: non-stationary
Kết quả	ADF Statistic: 0.763469 p-value: 0.991 >> 0.05 Critical Values: 1%: -3.432 5%: -2.862 10%: -2.567	KPSS Statistic: 1.9823461 P-Value: 0.01 < 0.05 Critical Values: 5%, 0.146 2.5%, 0.176 1%, 0.216
Kết luận	Dữ liệu không có tính dừng	Dữ liệu không có tính dừng

3. Mô hình & Kết quả



3.1 ARIMA (p,d,q)

$$\hat{y}' = c + \phi_1 \hat{y}'_{t-1} + \dots + \phi_p \hat{y}'_{t-p} + \theta_1 \epsilon_{t-1} + \dots + \theta_q \epsilon_{t-q} + \epsilon_t$$

ARIMA hoạt động dựa trên giả thuyết
chuỗi dừng và phương sai sai số không đổi

=> Cần biến đổi dữ liệu về dạng có tính
dừng



AR: Autoregression

p: bậc của phần Autoregression



MA: Moving Average

q: bậc của phần Moving Average



I: Integrated

d: bậc sai phân [số lần lấy sai
phân]

Lựa chọn tham số cho ARIMA

Chọn giá trị d : Vì dữ liệu ban đầu không có tính dừng, ta thử biến đổi dữ liệu bằng cách dùng logarit để kiểm tra tính dừng.

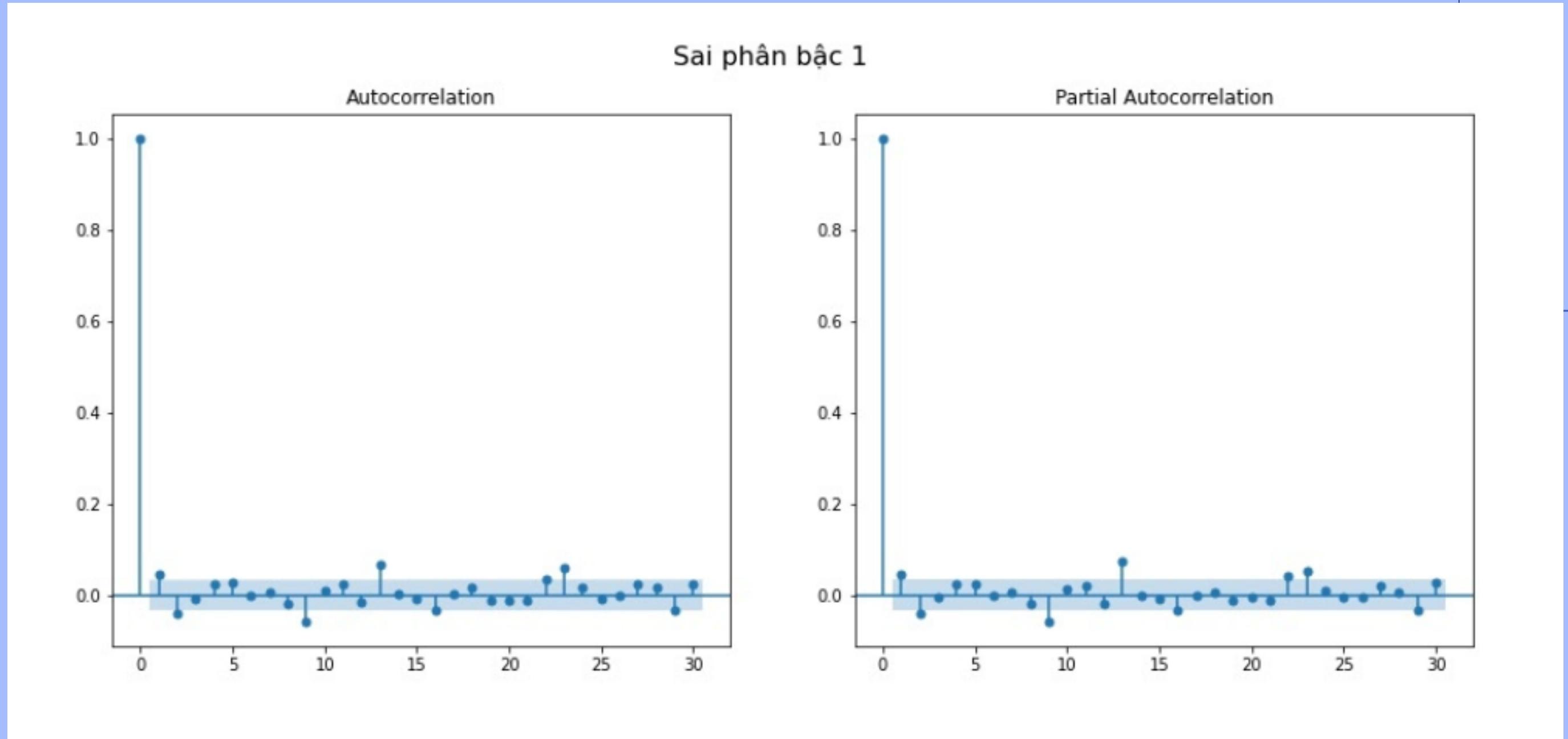
Biến đổi	Dùng hàm logarit	Sai phân bậc 1	Sai phân bậc 2
Kiểm tra tính dừng	ADF Statistic: -2.080686 p-value: 0.252351 > 0.05 Critical Values: 1%: -3.432 5%: -2.862 10%: -2.567	ADF Statistic: -12.054887 p-value: 0.0000 < 0.05 Critical Values: 1%: -3.432 5%: -2.862 10%: -2.567	ADF Statistic: -19.637767 p-value: 0.0000 < 0.05 Critical Values: 1%: -3.432 5%: -2.862 10%: -2.567
Kết luận	Dữ liệu không có tính dừng, chuyển sang dùng sai phân	Dữ liệu có tính dừng, chọn $d = 1$	Dữ liệu có tính dừng, chọn $d = 2$

Lựa chọn tham số cho ARIMA

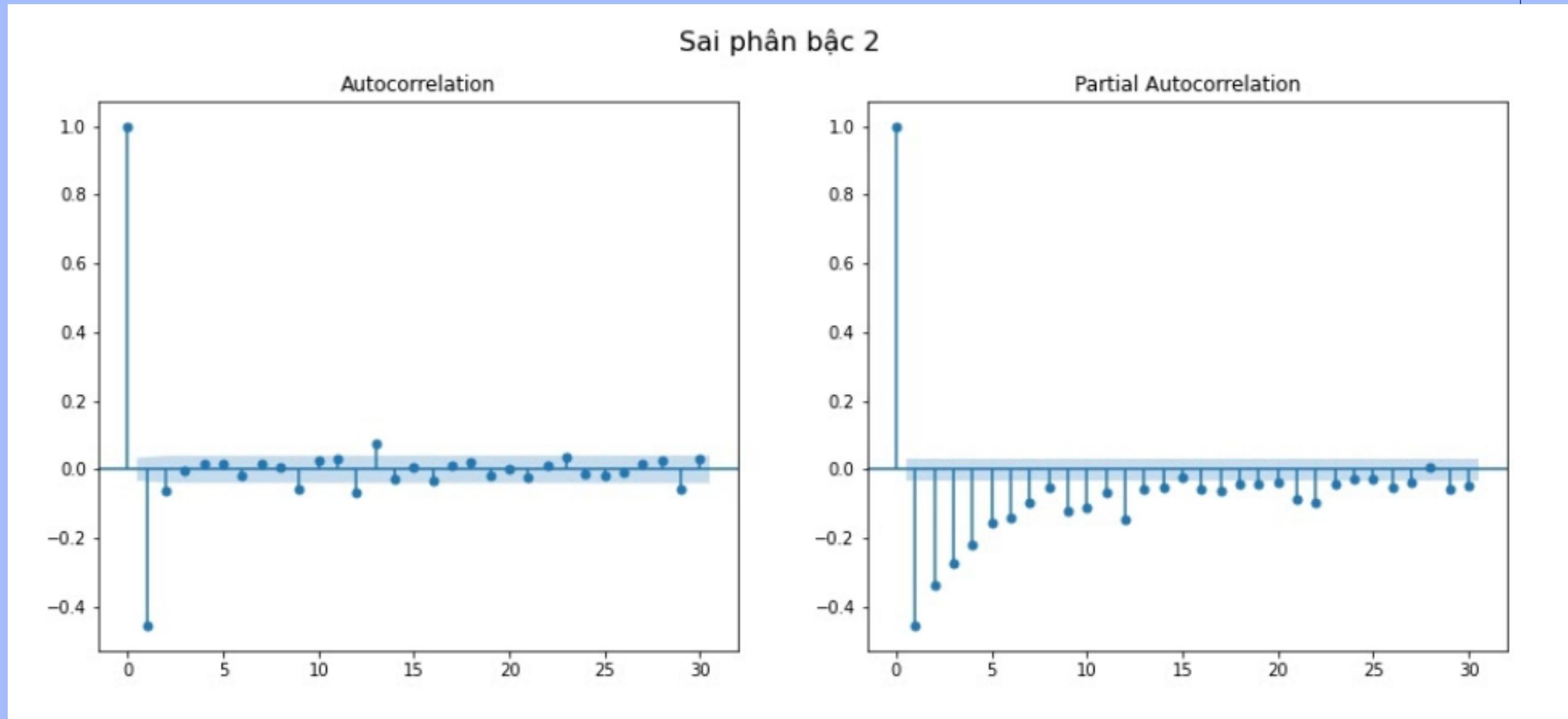
Chọn giá trị p: Từ biểu đồ Partial Autocorrelation của các sai phân, chọn giá trị bậc p của quá trình tự hồi quy AR

Chọn giá trị q: Từ biểu đồ Autocorrelation của các sai phân, chọn giá trị bậc q của quá trình MA

Lựa chọn tham số cho ARIMA



Lựa chọn tham số cho ARIMA



KL: $(p, d, q) = (1, 1, 1), (1, 1, 0), (1, 2, 1), (1, 2, 0)$

AIC

Tiêu chí lựa chọn mô hình ARIMA phù hợp nhất

Công thức:

$$AIC = T \log\left(\frac{SSE}{T}\right) + 2(k + 2)$$

Công cụ phạt lỗi dự báo
(SSE) và phạt số lượng
tham số của mô hình

T là số lượng quan sát

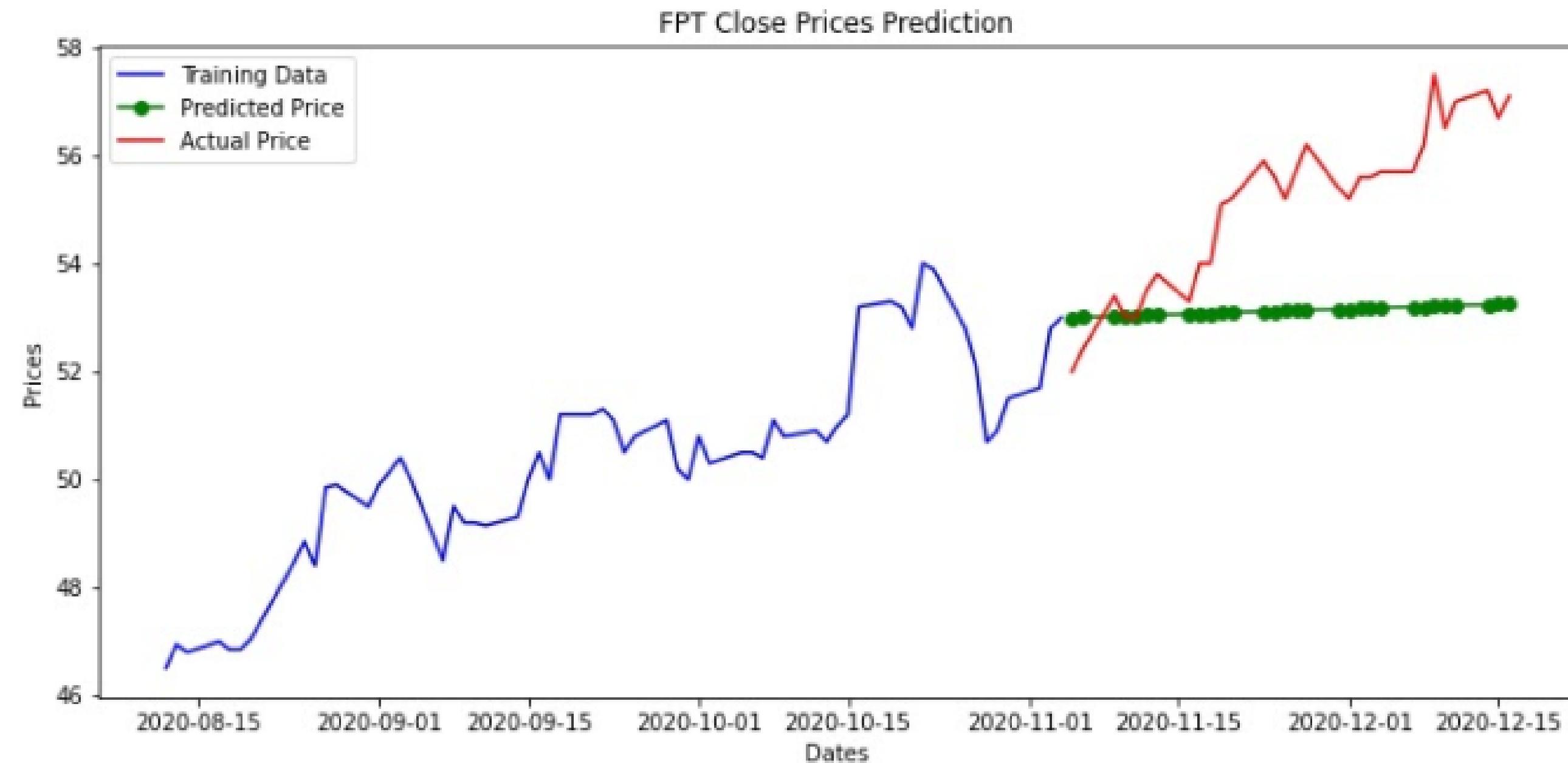
k là số lượng tham số

AIC càng nhỏ càng phù
hợp

Kết quả: Chọn mô hình
có tham số tốt nhất
(1,1,1)

Kết quả mô hình ARIMA

MAPE: 3.7088 %



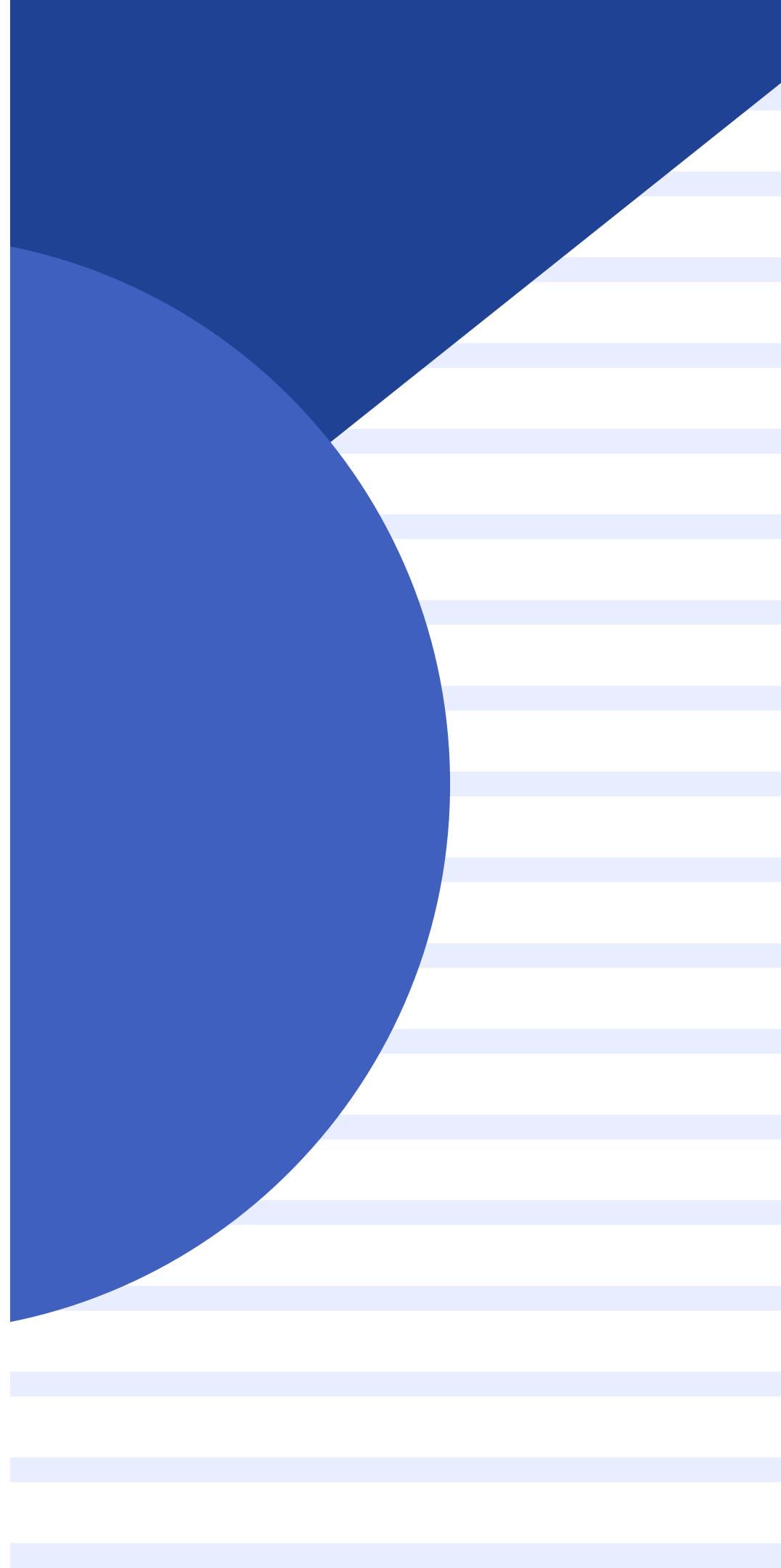
3.2 Ensemble learning

Khi kết hợp nhiều mô hình độc lập lại với nhau ta được một ensemble.

Dự đoán tốt hơn các mô hình đơn nhờ việc triệt tiêu các lỗi ngẫu nhiên
(random errors)

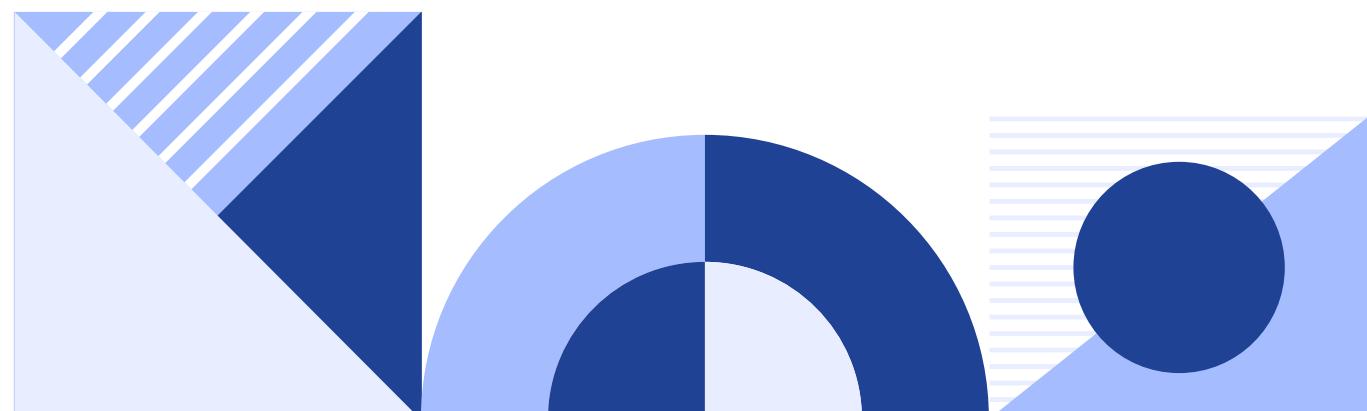
$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i},$$

Kết quả của phép dự báo sẽ là một trung bình cộng có trọng số từ kết quả dự báo của các mô hình trong ensemble



Ensemble learning

Một số mô hình được đưa vào Ensemble

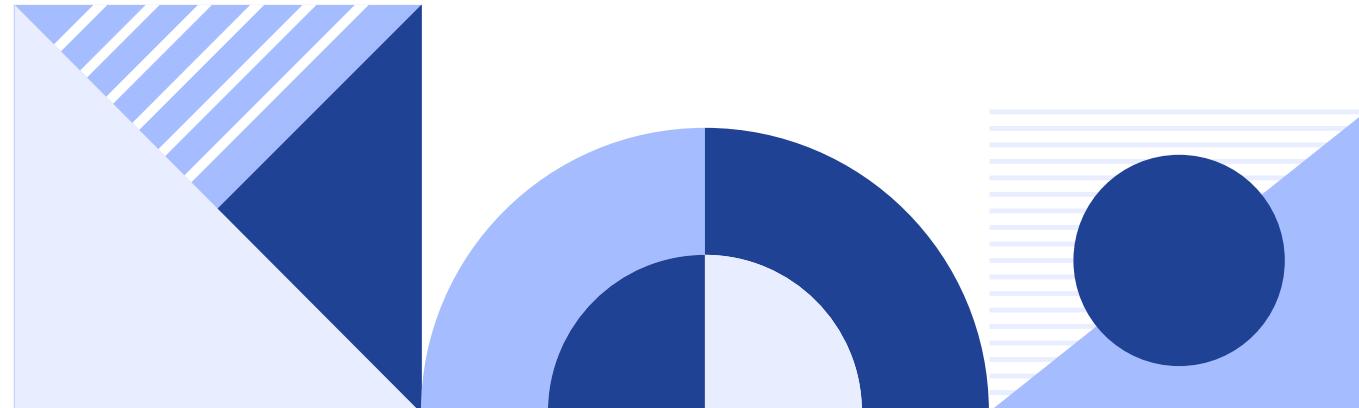


1) Theta model

- Mô hình có hiệu suất cao,
- Ý tưởng thay đổi độ cong của một chuỗi thời gian thông qua hệ số theta.
- Thu được các đường Theta-lines bằng cách lấy đạo hàm cấp hai của dữ liệu, đại diện cho các tính chất dài hạn/ngắn hạn (tùy thuộc vào giá trị)

Ensemble learning

Một số mô hình được đưa vào Ensemble



2) Exponential Smoothing (Holt-Winters)

Giá trị của một điểm cần dự đoán sẽ được tính bằng trung bình cộng có trọng số của các quan sát trước đó, các trọng số này sẽ giảm dần ngược chiều thời gian và tuân theo hàm mũ.

- Mô hình hóa yếu tố Trend

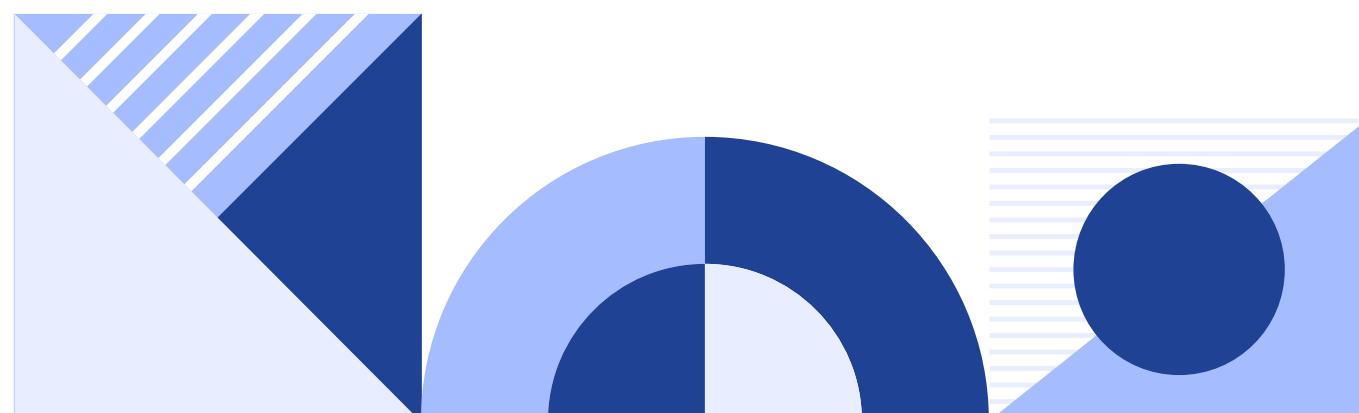
Forecast equation	$\hat{y}_{t+h t} = \ell_t + hb_t$
Level equation	$\ell_t = \alpha y_t + (1 - \alpha)(\ell_{t-1} + b_{t-1})$
Trend equation	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1},$

- Mô hình hóa yếu tố Seasonality

$$\begin{aligned}\hat{y}_{t+h|t} &= \ell_t + hb_t + s_{t+h-m(k+1)} \\ \ell_t &= \alpha(y_t - s_{t-m}) + (1 - \alpha)(\ell_{t-1} + b_{t-1}) \\ b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1 - \beta^*)b_{t-1} \\ s_t &= \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1 - \gamma)s_{t-m},\end{aligned}$$

Ensemble learning

Một số mô hình được đưa vào Ensemble



3) Auto ARIMA (SARIMAX)

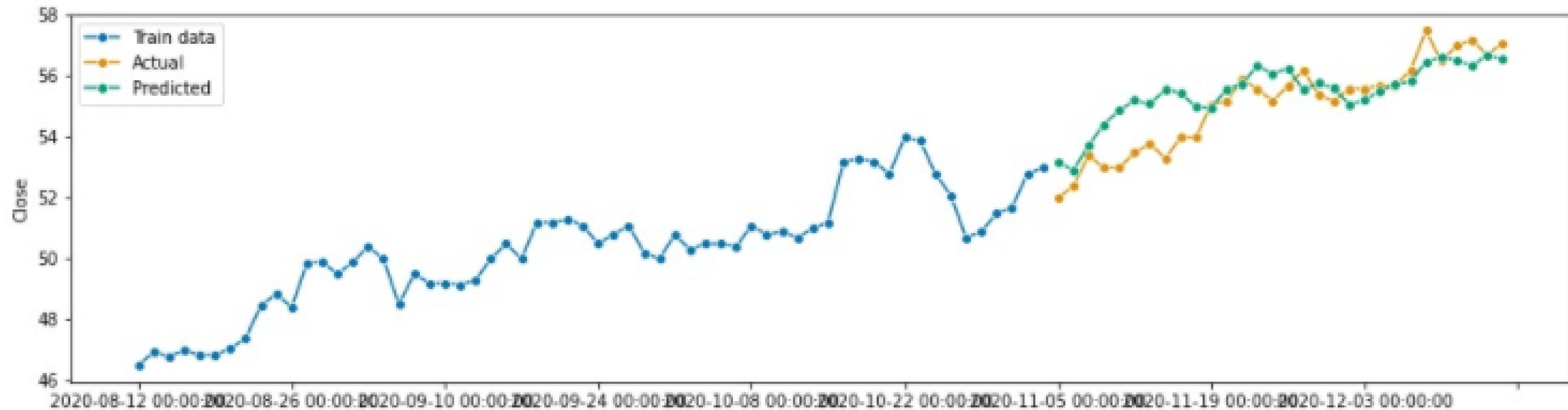
- ARIMAX: là ARIMA mở rộng, có thêm yếu tố tự tương quan được biểu diễn trong phần dư của mình, và được xem như một mô hình hồi quy động.
- SARIMA: là ARIMA được điều chỉnh đặc biệt cho những chuỗi thời gian có yếu tố mùa vụ, giúp tìm ra chu kỳ và quy luật của yếu tố mùa vụ, loại bỏ nó ra khỏi chuỗi.

=> Các bậc của mô hình được tìm kiếm vét cạn và chọn lấy mô hình có AIC tương ứng thấp nhất.

Ensemble learning

MAPE: 1.3429%

Mô hình	Trọng số
Theta model	0.9012
Exponential Smoothing	0.7696
Auto ARIMA	0.7933



3.3 PROPHET

Thư viện mở do Core data science team của Facebook xây dựng

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

- $g(t)$: hàm mô hình hóa Trend của dữ liệu theo thời gian
- $s(t)$: hàm mô hình hóa tính seasonal của dữ liệu theo thời gian
- $h(t)$: hàm mô hình hóa tính holiday của dữ liệu theo thời gian
- ϵ_t : thành phần đặc trưng của dữ liệu theo thời gian.

3.3 PROPHET

1. Mô hình hóa trend của dữ liệu theo thời gian

Prophet có 2 tùy chọn hàm để mô tả trend của dữ liệu: Linear & Logistic.

Prophet thực hiện việc hiệu chỉnh hàm trend tại nhiều
điểm có sự thay đổi trend lớn (được gọi là changepoints)

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

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

3.3 PROPHET

2. Mô hình hóa Seasonal của dữ liệu theo thời gian

Dữ liệu seasonal của dữ liệu sẽ được phân tách vào một hàm số theo thời gian sử dụng kiến thức về chuỗi Fourier.

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

3.3 PROPHET

3. Mô hình hóa Holiday của dữ liệu theo thời gian

$$h(t) = Z(t)\kappa.$$

Trong đó K được rút ra từ phân phối chuẩn $N(0, \text{gama})$. Nhưng trong bài toán chứng khoán, vì các ngày holiday các sàn không giao dịch nên ta sẽ bỏ qua mô hình này

3.3 PROPHET

Hiệu chỉnh mô hình sử dụng Grid Search

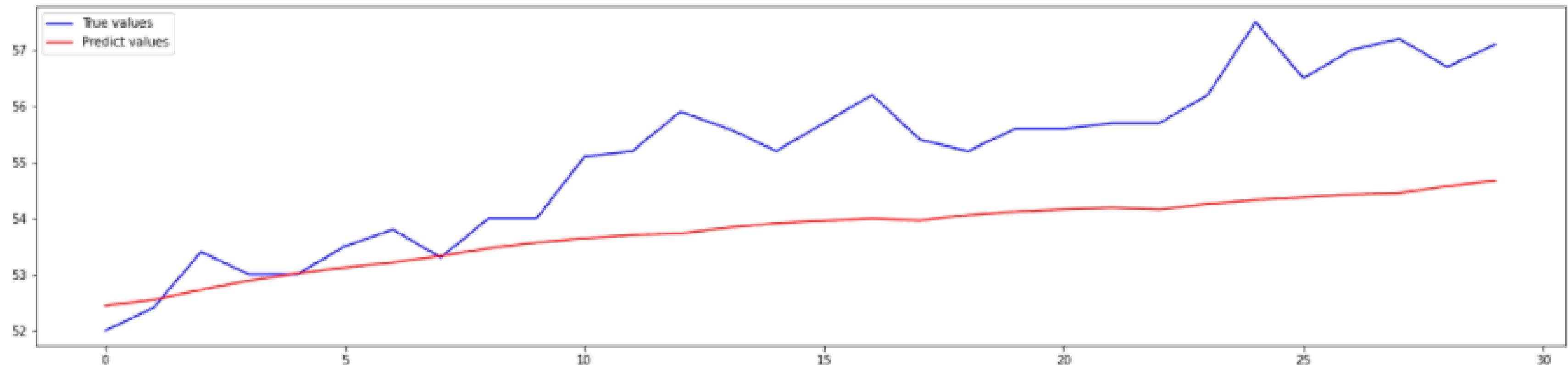
Ta sẽ tiến hành hiệu chỉnh khoảng đặt các change-points, T trong phân phối Laplace của trend, Sigma trong phân phối chuẩn của seasonal.Các thành phần này trong Prophet có các tên tương ứng sau:

- `changepoint_range`: khoảng hiệu chỉnh [0.8, 0.99]
- `changepoint_prior_scale`: khoảng hiệu chỉnh [0.05, 0.8]
- `seasonality_prior_scale`: khoảng hiệu chỉnh [0.01, 10]

3.3 PROPHET

Kết quả mô hình

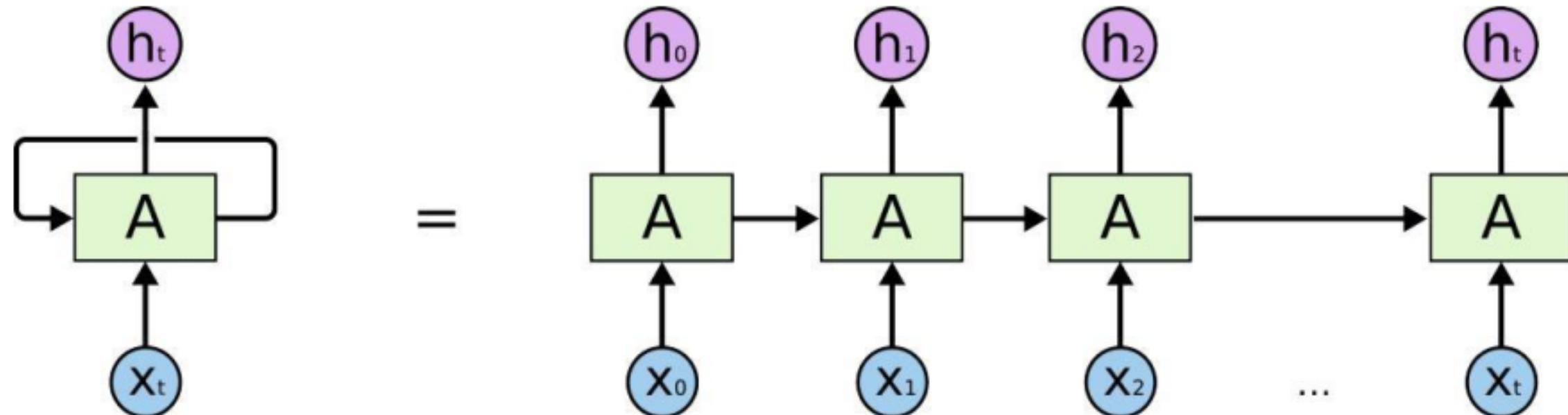
changepoint_range	changepoint_prior_scale	seasonality_prior_scale	MAPE
0.99	0.2	10	2.4516 %



3.4 LSTM

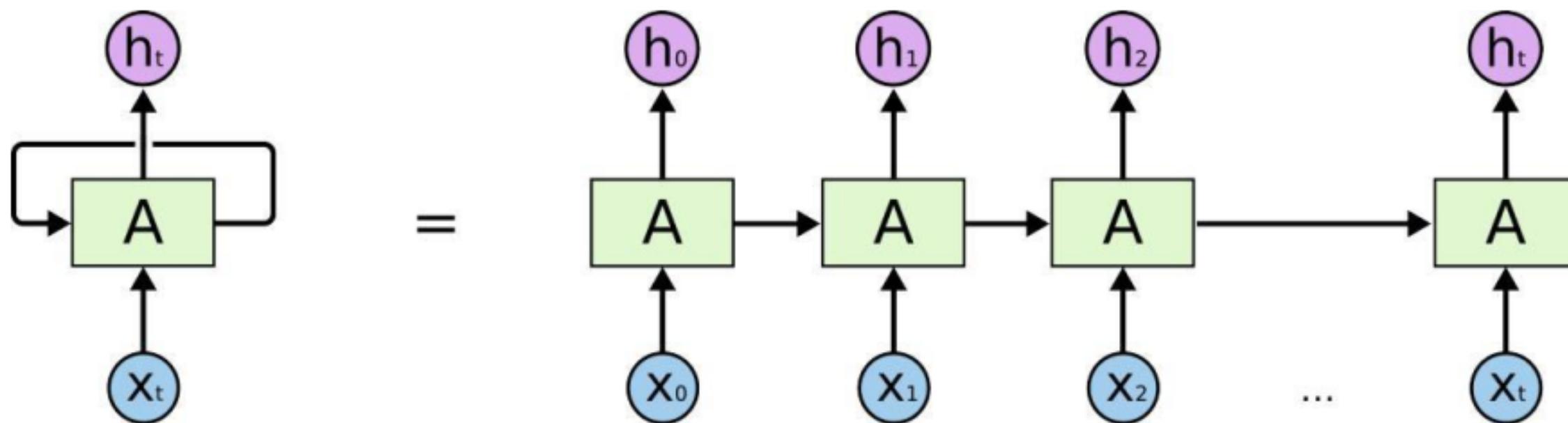
1) Recurrent Neural Network

Mô hình RNN được sinh ra để giải quyết các vấn đề sử dụng thông tin đã học được trước đó để cho ra kết quả mô hình



3.4 LSTM

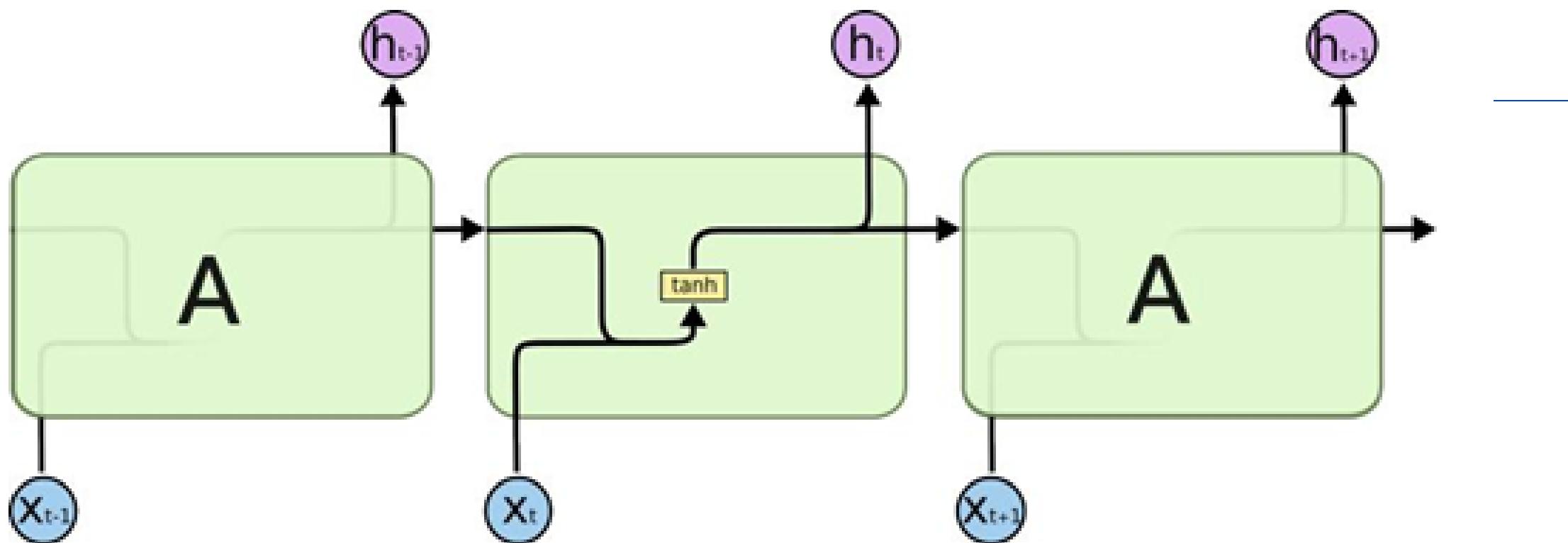
1) Recurrent Neural Network



Tuy mô hình RNN sử dụng được các thông tin đã được học trước đó để cho ra kết quả hiện tại, nhưng nếu các thông tin đã được học quá lâu sẽ dẫn đến không nhớ được

3.4 LSTM

2) LSTM



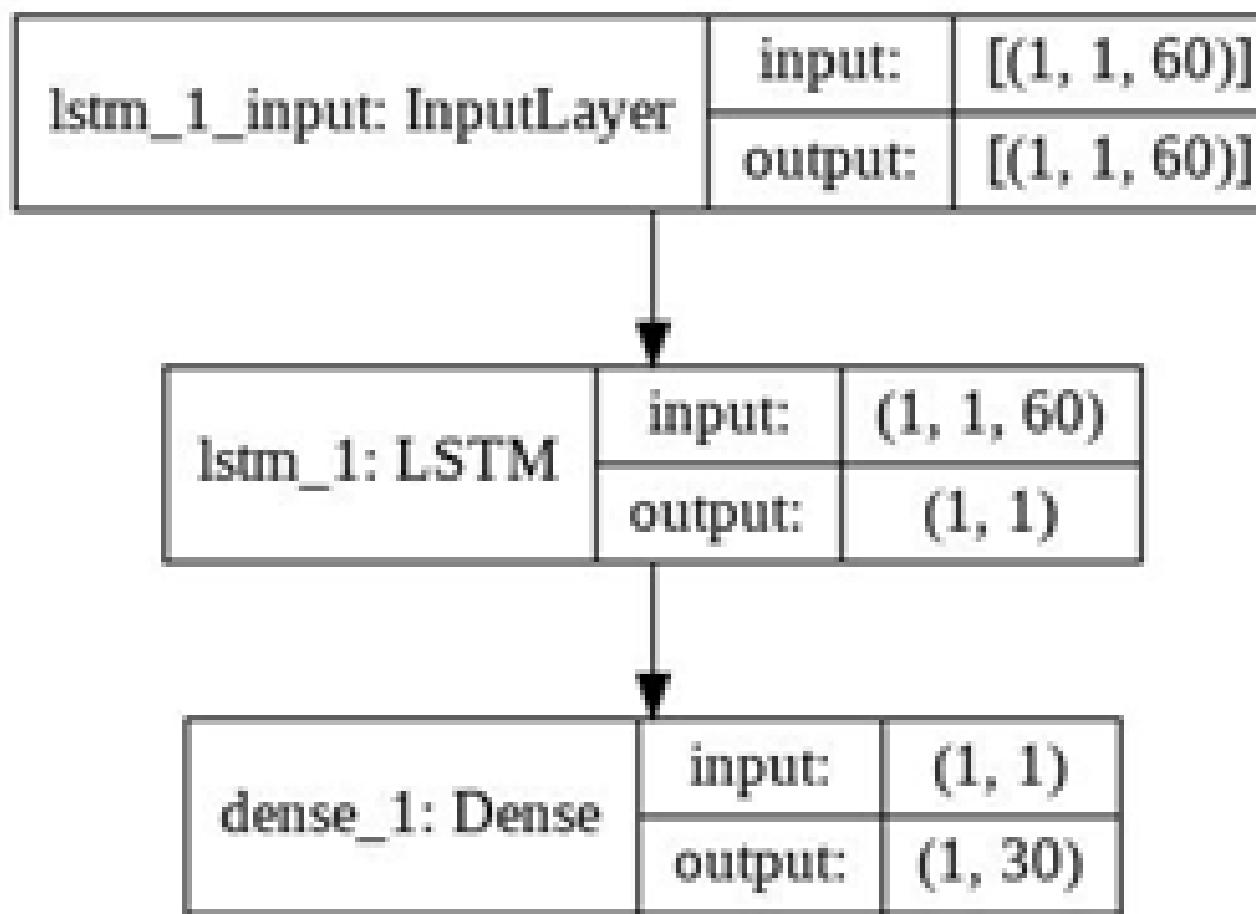
LSTM được thiết kế để tránh được vấn đề phụ thuộc xa (long-term dependency). Việc nhớ thông tin trong suốt thời gian dài là đặc tính mặc định của mô hình, không cần phải huấn luyện nó để có thể nhớ được

3.4 LSTM

2) LSTM

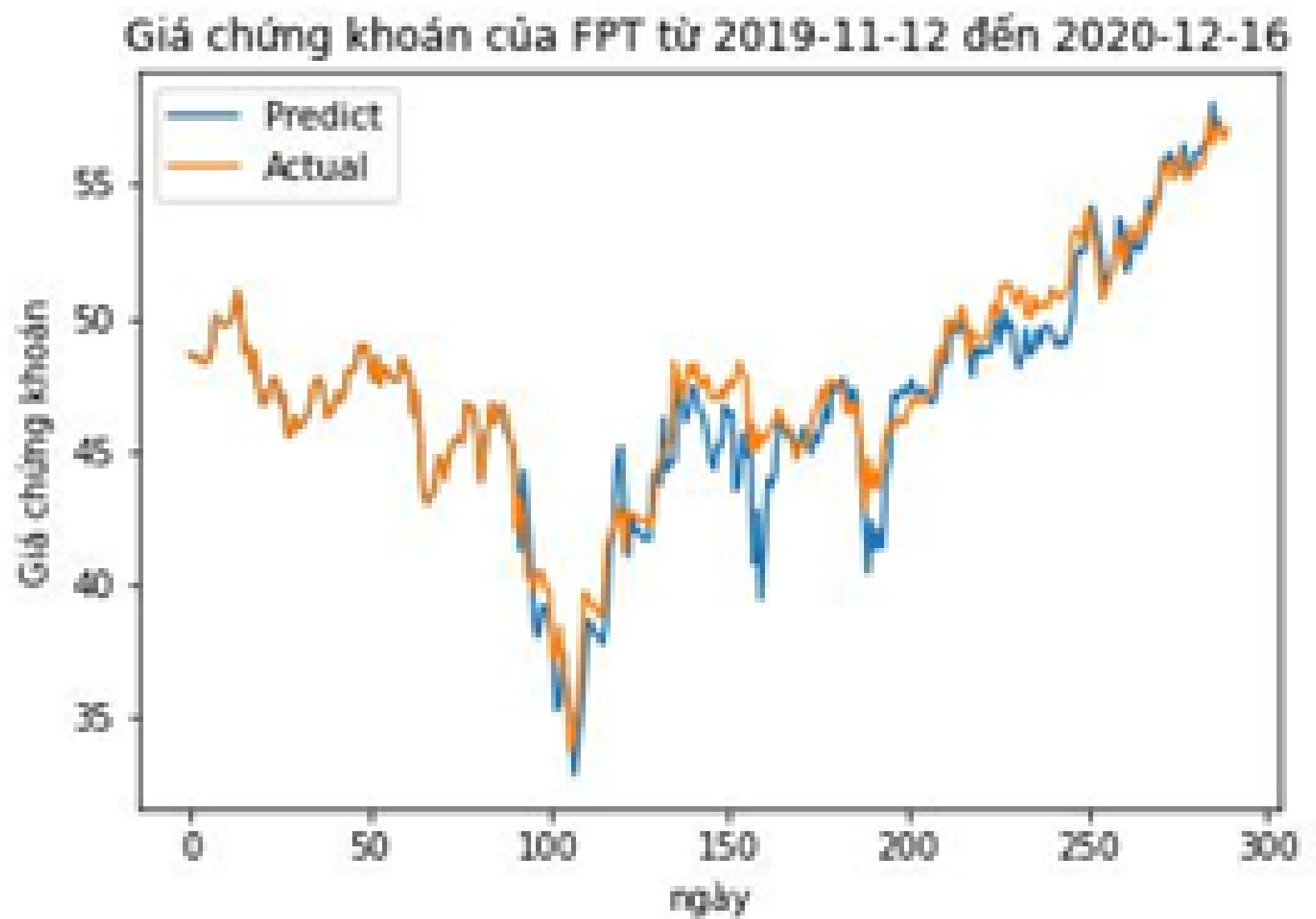
Nhóm sẽ dùng mô hình LSTM để dữ liệu đầu vào là các ngày trước đó và đầu ra sẽ là n-ngày sau các ngày trước đó.

Ví dụ: Lấy 21 ngày để dự đoán 7 ngày tiếp theo



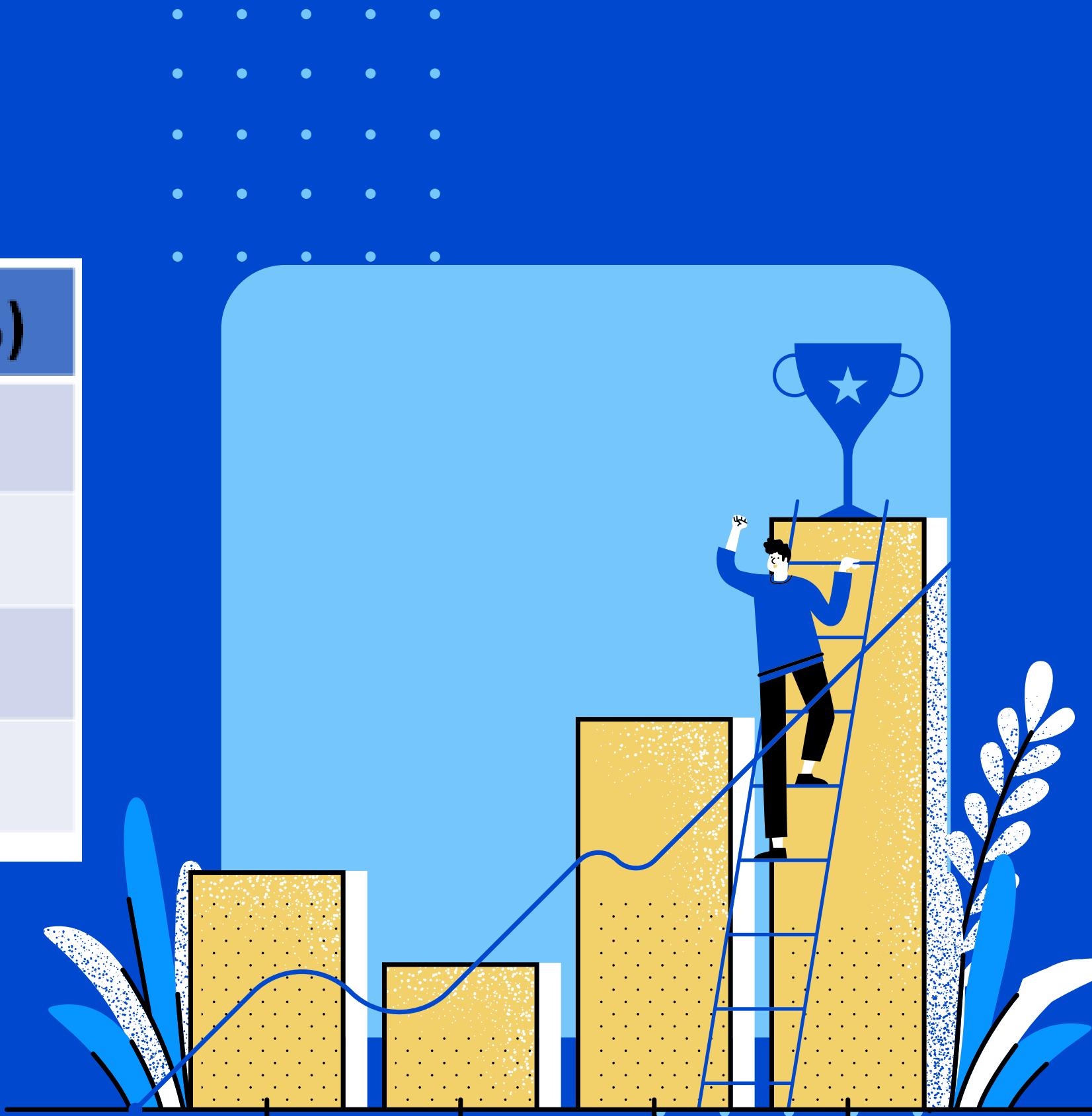
3.4 LSTM

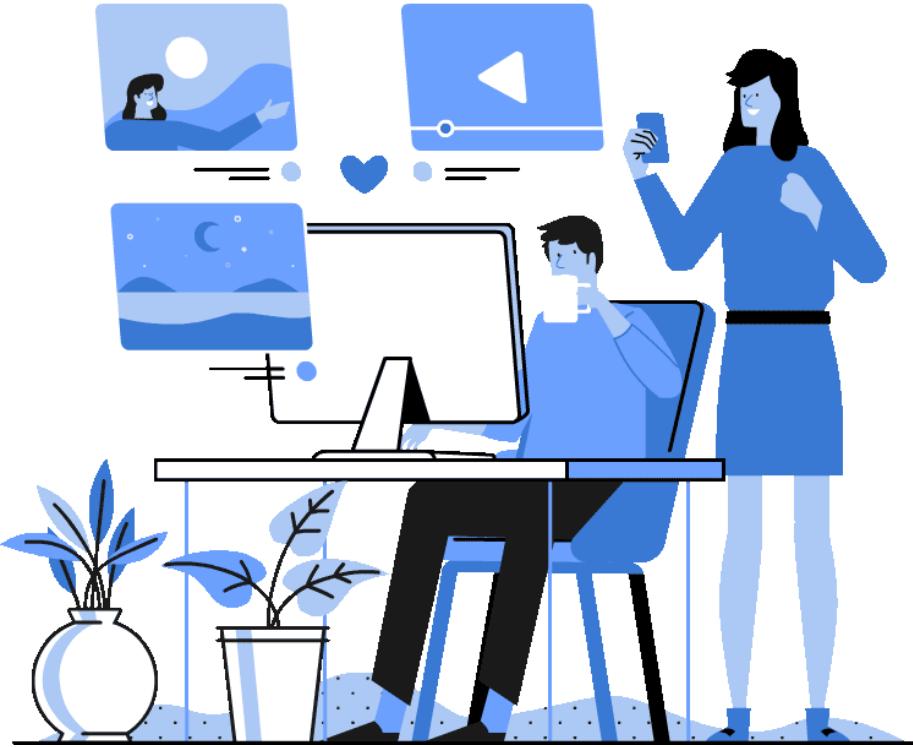
MAPE: 8.7472%



TỔNG KẾT

Model	Performance (MAPE) (%)
ARIMA	3.7088
Ensemble model	1.3429
Prophet	2.4516
LSTM	8.7472





4. Hướng nghiên cứu tiếp theo

01

Hiệu chỉnh các siêu tham số của mô hình thống kê bằng AutoML

02

Bổ sung các dữ liệu về Volume, EPS... để sử dụng mô hình dự báo đa biến

03

Nên kết hợp với các yếu tố constraints & loại trừ lạm phát

04

Bổ sung thêm
Confidence intervals

05

Dùng thêm các mô hình thường dùng cho Sequence: Attention, Transformers hoặc Dilated CNN

06

phát triển hệ thống tự động crawl dữ liệu, mô hình hóa, và đưa ra dự báo nhanh hỗ trợ các nhà đầu tư

5. DEMO



Xin cảm ơn sự lắng nghe của thầy cô và các bạn!



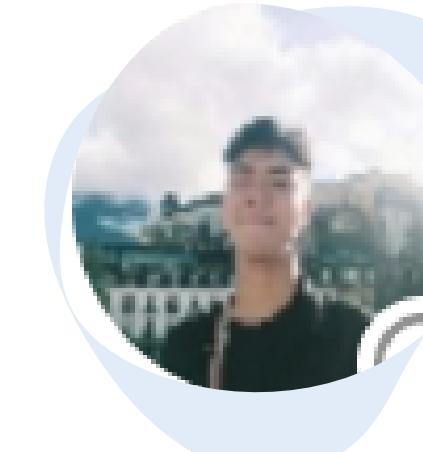
Trần Văn Tuấn
Trưởng nhóm



Trần Thị Ngọc Anh



Võ Đức Mẫn



Dương Chí Vinh