ĐẠI HỌC QUỐC GIA – THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KINH TẾ LUẬT

TIỂU LUẬN

PYTHON ỨNG DỤNG TRONG TÀI CHÍNH 2

ĐỀ TÀI:

Dự báo giá cổ phiếu ngân hàng TMCP Sài Gòn Thương Tín với mô hình ARIMA

Giảng viên: ThS. NGÔ PHÚ THANH

Sinh viên thực hiện: NGUYỄN VĂN THÀNH

MSSV: K204140640

Lớp: K20414C

Thành phố Hồ Chí Minh, tháng 01 năm 2022

Mục lục

I. Tổng quan

1.1. Tóm tắt đề tài

Thực tiễn trong hai năm qua, thị trường chứng khoán Việt Nam bùng nổ với lượng lớn nhà đầu tư (NĐT) mới tham gia, nâng tổng số tài khoản giao dịch chứng khoán vượt ngưỡng 6,3 triệu. Đặc điểm này được giới chuyên gia đánh giá là động lực giúp thị trường chứng khoán bùng nổ trong vòng 3-5 năm tới. Điều này cũng có nghĩa rằng thị trường dễ nhận những cú sốc do việc nhà đầu tư tư nhân có xu hướng đầu tư lướt sóng, đầu cơ áp đảo về khối lượng giao dịch trên thị trường. Các cổ phiếu được SHS phân tích sẽ bùng nổ trong năm 2023 là những nhóm cổ phiếu trong quá trình đẩy mạnh và chuyển đổi số. Nắm bắt được việc dự báo giá trong thị trường chứng khoán là một vấn đề quan trọng, Đó là lý do lựa chọn đề tài: "Dự báo giá cổ phiếu ngân hàng TMCP Sài Gòn Thương Tín với mô hình ARIMA".

1.2. Lý do chọn đề tài

Hoạt động kinh doanh cốt lõi của ngân hàng TMCP Sài Gòn Thương Tín trong quý III tiếp tục ghi nhận kết quả tích cực với tỷ lệ NIM tăng cao nhất trong các quý gần đây (khoảng 4,43%). Cả hoạt động cho vay và các hoạt động khác đều ghi nhận mức tăng trưởng vượt trội, với 7 nghìn tỷ đồng tổng thu nhập hoạt động trong quý III năm 2022 (tăng 68% so với cùng kỳ). Các chỉ số khác của hoạt động cho vay cũng được duy trì ổn định, với tỷ lệ nợ xấu thấp và tỷ lệ LLR cao, lần lượt là 0,9% và 154%.

Trong quý VI/2022, SSI dự báo STB sẽ đạt lợi nhuận trước thuế là 1,9 nghìn tỷ đồng (tăng 63,5%), giúp lợi nhuận trước thuế cả năm đạt 6,3 nghìn tỷ đồng (tăng 43,7%). Năm 2023, nhóm chuyên gia kỳ vọng lợi nhuận trước thuế sẽ tăng trưởng mạnh hơn, đạt 11,5 nghìn tỷ đồng (tăng 83%) do kỳ vọng phần lớn trái phiếu VAMC sẽ được trích lập dự phòng vào năm 2023. Ngoài ra, STB là một trong những ngân hàng cuối cùng đang phải xử lý nợ xấu còn lại của chu kỳ tín dụng trước. Trong những năm qua, STB đã nâng cao chất lượng tài sản một cách ấn tượng và không cho vay trái phiếu doanh nghiệp cũng như chỉ duy trì dư nợ cho vay bất động sản ở mức thấp. "Dự báo giá cổ phiếu ngân hàng TMCP

Sài Gòn Thương Tín với mô hình ARIMA" sẽ góp một phần nào đó giúp cho các nhà đầu tư có thể có những quyết định đầu tư an toàn hơn cho cổ phiếu STB trong năm 2023.

II. Phương pháp nghiên cứu

2.1. Lý cho chọn mô hình

Một trong những mô hình quan trọng trong thống kê và Manchine Learning là dự báo chuỗi thời gian. Một mô hình chuỗi thời gian thường dự báo dựa trên giả định rằng các qui luật trong quá khứ sẽ lặp lại ở tương lai. Do đó xây dựng mô hình chuỗi thời gian là đang mô hình hóa mối quan hệ trong quá khứ giữa biến độc lập (biến đầu vào) và biến phụ thuộc (biến mục tiêu). Dựa vào mối quan hệ này để dự đoán giá trị trong tương lai của biến phụ thuộc.

Các dự báo chuỗi thời gian có tính ứng dụng cao và được sử dụng rất nhiều lĩnh vực như tài chính ngân hàng, chứng khoán, bảo hiểm, thương mại điện tử, marketing, quản lý chính sách. Trong đó, mô hình dự báo Arima là một trong những mô hình phổ biến dùng để dự báo giá chứng khoán và các chuỗi lợi suất danh mục để quản trị danh mục đầu tư.

2.2. Tổng quan về phương pháp nghiên cứu

2.2.1. Lý thuyết mô hình Arima

ARIMA model là viết tắt của cụm từ Autoregressive Intergrated Moving Average. Mô hình sẽ biểu diễn phương trình hồi qui tuyến tính đa biến (multiple linear regression) của các biến đầu vào (còn gọi là biến phụ thuộc trong thống kê) là 2 thành phần chính:

Thứ nhất Auto regression: Kí hiệu là AR. Đây là thành phần tự hồi qui bao gồm tợp hợp các độ trễ của biến hiện tại. Độ trễ bậc p chính là giá trị lùi về quá khứ p bước thời gian của chuỗi. Độ trễ dài hoặc ngắn trong quá trình AR phụ thuộc vào tham số trễ p. Cụ thể, quá trình AR(p) của chuỗi x_t được biểu diễn như bên dưới:

$$\mathrm{AR}(p) = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p}$$

Moving average: Qúa trình trung bình trượt được hiểu là quá trình dịch chuyển hoặc thay đổi giá trị trung bình của chuỗi theo thời gian. Do chuỗi của chúng ta được giả định là dừng nên quá trình thay đổi trung bình dường như là một chuỗi nhiễu trắng. Qúa trình moving average sẽ tìm mối liên hệ về mặt tuyến tính giữa các phần tử ngẫu nhiên ϵ_t (stochastic term). Chuỗi này phải là một chuỗi nhiễu trắng thỏa mãn các tính chất:

$$\begin{cases} \mathrm{E}(\epsilon_t) & = 0 & (1) \\ \sigma(\epsilon_t) & = \alpha & (2) \\ \rho(\epsilon_t, \epsilon_{t-s}) & = 0, \forall s <= t & (3) \end{cases}$$

ARIMA là mô hình kết hợp của 2 quá trình tự hồi qui và trung bình trượt. Dữ liệu trong quá khứ sẽ được sử dụng để dự báo dữ liệu trong tương lai. Trước khi huấn luyện mô hình, cần chuyển hóa chuỗi sang chuỗi dừng bằng cách lấy sai phân bậc 1 hoặc logarit. Ngoài ra mô hình cũng cần tuân thủ điều kiện ngặt về sai số không có hiện tượng tự tương quan và phần dư là nhiễu trắng. Đó là lý thuyết của kinh tế lượng. Còn theo trường phái machine learning thì tôi chỉ cần quan tâm đến làm sao để lựa chọn một mô hình có sai số dự báo là nhỏ nhất.

2.2.2. Quy trình dự báo giá cổ phiếu STB

Thứ nhất - Phân rã dữ liệu: Phân rã dữ liệu chuỗi thời gian là việc phân tách từng yếu tố thành phần của dữ liệu chuỗi thời gian ra thành các yếu tố khác nhau. Từ đó, chúng ta có thể phân tích từng yếu tố này nhằm hiểu rõ hơn về bản chất của dữ liệu mà nếu không phân tách riêng các yếu tố này thì việc phân tích sẽ trở nên khó khăn hơn. Phân rã dữ liệu chuỗi thời gian sau đó vẽ biểu đồ để dễ dàng quan sát.

Thứ hai - Kiểm định tính dừng của dữ liệu: Một chuỗi thời gian được xem là dừng khi có giá trị trung bình, phương sai và hiệp phương sai không đổi tại mọi thời điểm. Dùng kiểm định Dickey-Fuller mở rộng (Augmented Dickey-Fuller Test), hay còn gọi là ADF để kiểm định tính dừng cho bộ dữ liệu.

Thứ ba – Sửa lỗi tính dừng cho dữ liệu: Đây là cách biến đổi dữ liệu sao cho chuỗi dữ liệu sau khi được biến đổi thỏa mãn các điều kiện của tính dừng. Các cách sửa lỗi tính dừng phổ biến cho dữ liệu là sử dụng sai phân.

Thứ tư – Kiểm định tự tương quan: Tự tương quan là hiện tượng các giá trị trong dữ liệu chuỗi thời gian có quan hệ tương quan lẫn nhau. Đối với dữ liệu chuỗi thời gian trong lĩnh vực tài chính thì vấn đề tự tương quan rất phổ biến. Điều này dẫn đến việc vi phạm giả định cho phần dư của đa số các mô hình hồi quy thường sử dụng cho dạng dữ liệu này. Biểu đồ ACF (Autocorrelation Function)/PACF (Partial Autocorrelation Function) là phương pháp phổ biến để xác định hiện tượng tự tương quan trong dữ liệu chuỗi. Trong đó ACF thể hiện tự tương quan tổng thể giữa các độ trễ của dữ liệu; còn PACF thể hiện tương quan trực tiếp giữa các độ trễ của dữ liệu.

Thứ năm – Phân tách các nhân tố mô hình ARIMA(p,d,q):

+ AR (Autoregressive) mô hình tự hồi quy: mô hình hóa các mối quan hệ giữa dữ liệu và độ trễ của chính nó. Bậc của mô hình tự hồi quy được thể hiện bởi nhân tố "p" trong mô hình ARIMA.

+ I (Integrated) tích hợp: trong trường hợp này là việc sử dụng sai phân nhằm loại bỏ tính xu hướng trong dữ liệu. Bậc sai phân sử dụng được thể hiện bằng nhân tố "d" trong mô hình ARIMA

+ MA(Moving Average) mô hình trung bình trượt: mô hình hóa mối quan hệ giữa dữ liệu và sai số so với trung bình của các độ trễ khác. Bậc của mô hình trung bình trượt được thể hiện bằng nhân tố "q" trong mô hình ARIMA.

Thứ sáu – Kiểm định tính thích hợp của mô hình, các tiêu chí để đánh giá mô hình tốt như sau:

+ Phần dư của mô hình dự báo phải là nhiễu trắng.

+ Các hệ số hồi quy có ý nghĩa trong thống kê.

+ Hệ số R2 lớn.

+ Giá trị dự báo càng gần với giá trị thực tế càng tốt.

Thứ bảy – Thực hiện dự báo với mô hình ARIMA.

2.3. Ngôn ngữ lập trình

Trong bài báo cáo sẽ sử dụng ngôn ngữ lập trình là Python.

Các thư viện sử dụng trong mô hình:

```
import pandas as pd
import numpy as np
import datetime
import matplotlib.pyplot as plt
import seaborn as sns
from vnstock import listing_companies,ticker_overview,stock_historical_data
from sklearn.metrics import mean_squared_error
from plotly.subplots import make_subplots
import plotly.graph_objects as go
from scipy.signal import argrelextrema
from collections import deque
from matplotlib.lines import Line2D
from statsmodels.tsa.stattools import adfuller
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.tsa.arima_model import ARIMA
import pmdarima as pm
import scipy.stats as scs
import statsmodels.api as sm
from datetime import timedelta
import warnings
```

III. Thu thập dữ liệu và trực quan hóa

3.1. Nguồn dữ liệu

Dữ liệu được thu thập thông qua gói thư viện "vnstock", sáng tạo bởi TCBS và SSI, một trong những công ty chứng khoáng cung cấp dữ liệu phân tích đầy đủ và trực quan nhất cho các nhà đầu tư Fo. Dữ liệu được truyền qua api và được đọc bằng pandas một cách nhanh chóng.

3.2. Tải và đọc dữ liệu

Cài đặt thư viện "vnstock": **pip install vnstock.**

Đọc dữ liệu cổ phiếu của ngân hàng TP Bank thông qua câu lệnh sau:

```
df =  stock_historical_data(symbol='STB', start_date="2021-01-01", end_date='2022-12-31')
```

Dữ liệu dùng để dự báo trong bài báo cáo được sử dụng từ ngày 01/01/2021 đến ngày 31/12/2022.

3.3. Trực quan hóa dữ liệu

Dữ liệu thu được có 6 cột quan sát.

Hình 3.1: 5 dòng đầu của dữ liệu.

	Open	High	Low	Close	Volume	TradingDate
0	17050.0	17700.0	16900.0	17550.0	46809600	2021-01-04
1	17550.0	17950.0	17450.0	17700.0	23399500	2021-01-05
2	17900.0	18350.0	17700.0	18000.0	31193600	2021-01-06
3	18050.0	18450.0	17850.0	18450.0	39683500	2021-01-07
4	18800.0	19500.0	18650.0	19400.0	34889200	2021-01-08

Đưa cột quan sát TradingDate làm index.

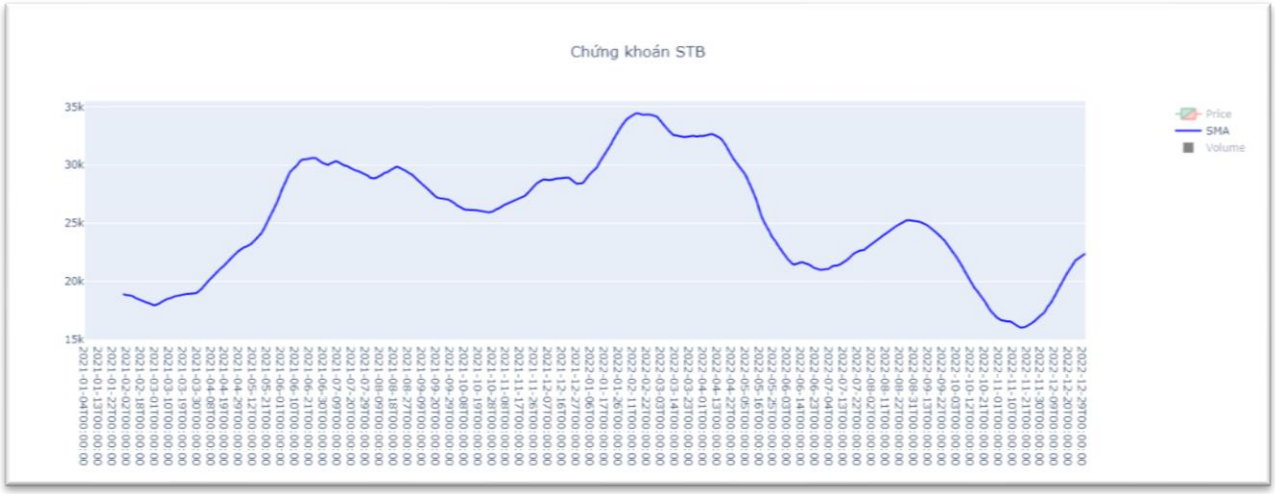
Hình 3.2: Code và đưa TradingDate làm index.

```
df['TradingDate'] = pd.to_datetime(df['TradingDate'])
df = df.set_index('TradingDate')
df.head()
```

TradingDate	Open	High	Low	Close	Volume
2021-01-04	17050.0	17700.0	16900.0	17550.0	46809600
2021-01-05	17550.0	17950.0	17450.0	17700.0	23399500
2021-01-06	17900.0	18350.0	17700.0	18000.0	31193600
2021-01-07	18050.0	18450.0	17850.0	18450.0	39683500
2021-01-08	18800.0	19500.0	18650.0	19400.0	34889200

Vẽ đồ đồ thị biểu diễn các quan sát trong bộ dữ liệu (Biểu đồ có thể zoom và quan sát kỹ hơn trong kết quả của file code.)

Biểu đồ 3.1: Đường SMA



Biểu đồ 3.2: Biểu đồ nến

Nhận xét: Dựa vào biểu đồ 3.1,3.2 có thể thấy rằng giá cổ phiếu của STB đang có xu hướng tăng trong những ngày cuối năm 2022.

Vì mô hình chỉ sử dụng để dự báo giá đóng cửa nên loại bỏ các cột quan sát không cần thiết.

Hình 3.3: Code và kết quả loại bỏ các cột quan sát không cần thiết

```
df_arima =  stock_historical_data(symbol='STB', start_date="2021-01-01", end_date='2022-12-31')
df_arima = df_arima[['TradingDate','Close']]
```

```
df_arima['TradingDate'] = pd.to_datetime(df_arima['TradingDate'])
df_arima = df_arima.set_index('TradingDate')
df_arima.head()
```

TradingDate	Close
2021-01-04	17550.0
2021-01-05	17700.0
2021-01-06	18000.0
2021-01-07	18450.0
2021-01-08	19400.0

IV. Kiểm định mô hình dự báo ARIMA

4.1. Phân rã dữ liệu chuỗi thời gian

Chuyển dữ liệu theo tuần.

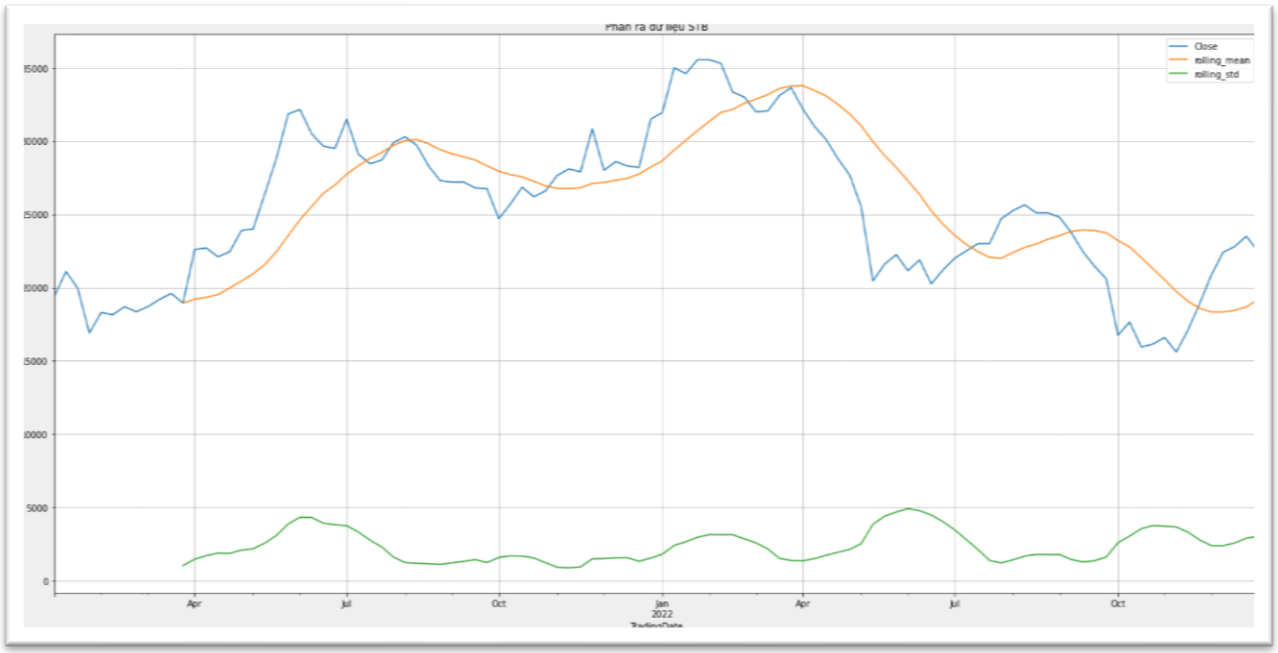
```
df_arima = df_arima.resample('W-Fri').ffill()
```

Chỉ rõ giá trị rolling là 12 kỳ giao dịch.

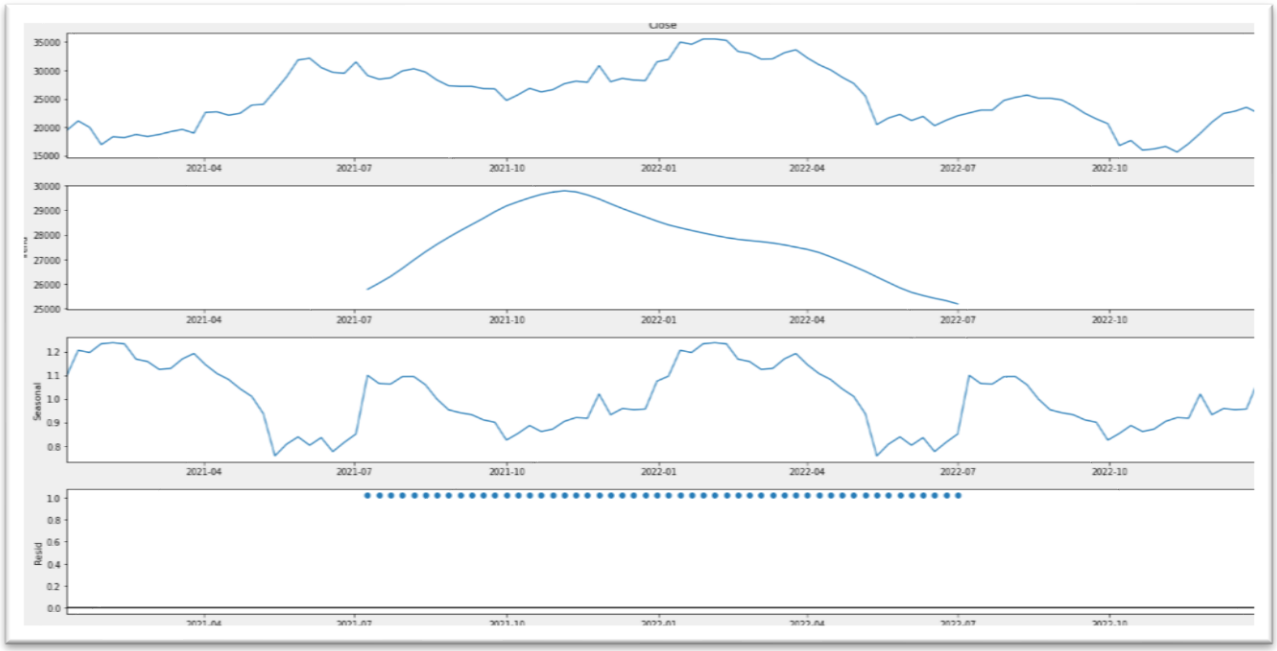
```
window_size = 12
```

Tạo cột mới chứa giá trị trung bình và độ lệnh chuẩn trong của 12 kỳ, sau đó vẽ đồ thị để biểu diễn các giá trị trên.

Biểu đồ 4.1: Phân rã dữ liệu chuỗi thời gian của STB



Biểu đồ 4.2: Phân rã dữ liệu chuỗi thời gian của STB



Nhận xét: Dựa vào biểu đồ 4.2 và 4.3, có thể thấy rằng dữ liệu có tính lặp lại rõ ràng theo năm 2021 và 2022. Giá cổ phiếu năm 2021 và 2022 có xu hướng giảm.

4.2. Tính dừng của dữ liệu

4.2.1. Kiểm định tính dừng

Như phương pháp nghiên cứu đã nói ở trên, phương pháp kiểm định tính dừng được sử dụng trong bộ dữ liệu là kiểm định Dickey-Fuller (ADF).

Hình 4.1: Kết quả kiểm định tính dừng

```
Test Statistic                  -1.941927
P-value                          0.312601
Numbers of Lags Used             2.000000
Numbers of Observations Used   101.000000
Critical Value (1%)             -3.496818
Critical Value (5%)             -2.890611
Critical Value (10%)            -2.582277
dtype: float64
```

Nhận xét: Dựa vào hình 4.1 có thể thấy rằng kết quả cho thấy p-value có giá trị là 0.31 và các giá trị Test statistic có giạ trị lớn hơn Critical Value ở cả 3 mức độ tin cậy. Kết luận chuỗi dữ liệu không dừng vì vậy phải khắc phục tính dừng cho dữ liệu theo phương pháp đã nêu ở trên là sử dụng sai phân.

4.2.2. Khắc phục tính dừng và kiểm định tự tương quan cho dữ liệu

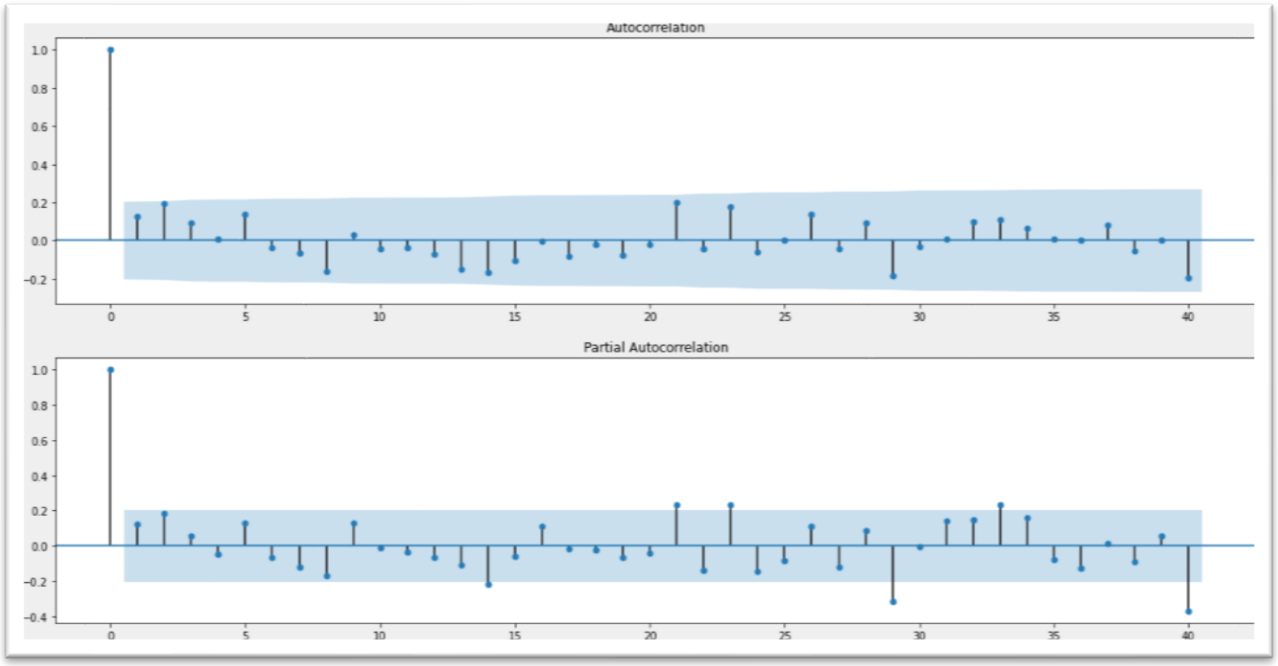
Tạo thêm cột quan sát chứa sai phân bậc 1, sau đó xóa các dữ liệu khuyết và tính lại các chỉ số kiểm định tính dừng.

Hình 4.2: Kết quả khắc phục tính dừng của dữ liệu

```
Test Statistic                  -3.400225
P-value                          0.010951
Numbers of Lags Used             8.000000
Numbers of Observations Used    84.000000
Critical Value (1%)             -3.510712
Critical Value (5%)             -2.896616
Critical Value (10%)            -2.585482
dtype: float64
```

Vẽ biểu đồ ACF và PACF của STB.

Biểu đồ 4.3: Biểu đồ ACF và PACF cho dữ liệu sai phân bậc 1 của STB



Nhận xét: Dữ liệu đã đảm bảo các yêu cầu của tính dừng. Có thể kết luận rằng tham số "d" là 1, các nút giao động từ 1 đến 2 nút và có một vài nút ở 3. Các hệ số có "p" và "q" có thể sử dụng từ 0 đến 2 và "d" là 1. Quyết định sử dụng ARIMA(1,1,0) vì hệ số thấp nhất.

4.3. Số liệu thống kê mô hình ARIMA cho STB

Hình 4.3: Các số liệu thống kê ARIMA(1,1,0)

SARIMAX Results

Dep. Variable:	Close	No. Observations:	93
Model:	ARIMA(1, 1, 0)	Log Likelihood	-797.273
Date:	Tue, 10 Jan 2023	AIC	1598.546
Time:	10:43:37	BIC	1603.589
Sample:	03-26-2021 - 12-30-2022	HQIC	1600.582
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2056	0.073	2.816	0.005	0.062	0.349
sigma2	1.971e+06	2.43e+05	8.118	0.000	1.5e+06	2.45e+06

Ljung-Box (L1) (Q):	0.96	Jarque-Bera (JB):	4.91
Prob(Q):	0.33	Prob(JB):	0.09
Heteroskedasticity (H):	0.99	Skew:	-0.34
Prob(H) (two-sided):	0.97	Kurtosis:	3.91

Nhận xét: Dựa vào hình 4.3 có thể thấy rằng “P>|z|”, có nghĩa rằng các hệ số thống kê đều có ý nghĩa, mô hình có thể sử dụng được. Để có thể chắc chắn hơn, sử dụng phương pháp kiểm định AIC để có thể lựa chọn ra mô hình thấp nhất.

Hình 4.4: Kết quả kiểm định AIC

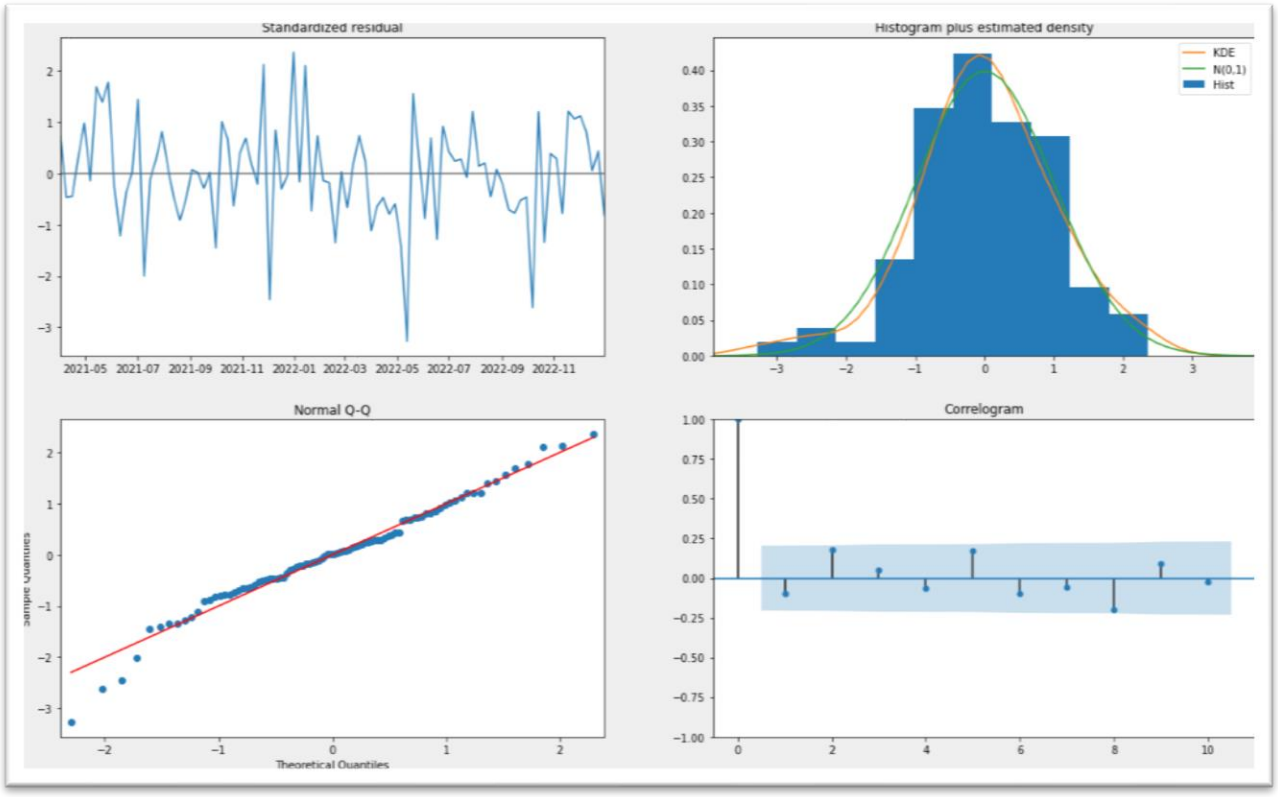
```
Performing stepwise search to minimize aic
 ARIMA(2,1,2)(0,0,0)[0] intercept   : AIC=inf, Time=0.46 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=1605.828, Time=0.03 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=1600.549, Time=0.03 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=1601.759, Time=0.05 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=1603.897, Time=0.03 sec
 ARIMA(2,1,0)(0,0,0)[0] intercept   : AIC=1601.457, Time=0.09 sec
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=1601.607, Time=0.13 sec
 ARIMA(2,1,1)(0,0,0)[0] intercept   : AIC=1603.430, Time=0.20 sec
 ARIMA(1,1,0)(0,0,0)[0]             : AIC=1598.546, Time=0.05 sec
 ARIMA(2,1,0)(0,0,0)[0]             : AIC=1599.464, Time=0.05 sec
 ARIMA(1,1,1)(0,0,0)[0]             : AIC=1599.610, Time=0.09 sec
 ARIMA(0,1,1)(0,0,0)[0]             : AIC=1599.762, Time=0.05 sec
 ARIMA(2,1,1)(0,0,0)[0]             : AIC=1601.429, Time=0.14 sec

Best model:  ARIMA(1,1,0)(0,0,0)[0]
Total fit time: 1.447 seconds
```

Nhận xét: Dựa vào hình 4.4 có thể thấy rằng mô hình ARIMA(1,1,0) là tối ưu nhất.

4.4. Kiểm định phần dư

Biểu đồ 4.4: Đồ thị phân phối phần dư của mô hình ARIMA



Nhận xét: Dựa vào biểu 4.4 có thể đánh giá phần dư ổn định, mô hình phù hợp vì các giá trị trong "Normal Q-Q" các chấm gần như trùng với đường màu đỏ, phần dư được phân phối bình thường. Đường KDE có cùng xu hướng với N(0,1).

V. Dự báo giá cổ phiếu STB với mô hình ARIMA

5.1. Chuẩn bị dữ liệu

Chia dữ liệu thành 2 tập train và test với tập train từ đầu năm 2021 đến 30/11/2022 và tập test là phần còn lại. Sau đó huấn luyện mô hình với mô hình ARIMA.

Hình 5.1: Tập test

```
test
```

```
TradingDate
2022-12-02    20850.0
2022-12-09    22400.0
2022-12-16    22800.0
2022-12-23    23500.0
2022-12-30    22500.0
Freq: W-FRI, Name: Close, dtype: float64
```

Hình 5.2: Tập train

```
train
```

```
TradingDate
2021-03-26    18950.0
2021-04-02    22600.0
2021-04-09    22700.0
2021-04-16    22100.0
2021-04-23    22450.0
               ...
2022-10-28    16150.0
2022-11-04    16600.0
2022-11-11    15600.0
2022-11-18    17100.0
2022-11-25    18900.0
Freq: W-FRI, Name: Close, Length: 88, dtype: float64
```

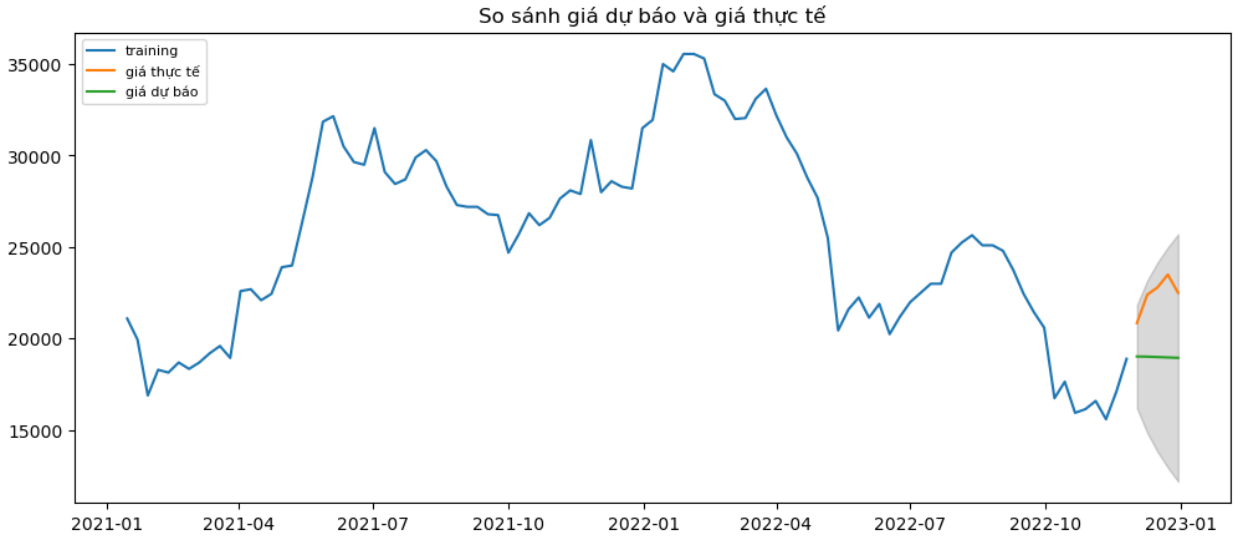
5.2 Dự báo

Thực hiện dự báo với mô hình ARIMA sau khi chuẩn bị dữ liệu cho ra được kết quả như sau:

Hình 5.3: So sánh giá thực tế và giá dự báo

TradingDate	Giá thực tết	giá dự báo
2022-12-02	20850.0	22482.873044
2022-12-09	22400.0	22482.579712
2022-12-16	22800.0	22482.574688
2022-12-23	23500.0	22482.574602
2022-12-30	22500.0	22482.574600

Biểu đồ 5.1: Kết quả dự báo giá STB với mô hình ARIMA



Nhận xét: dựa vào kết quả nhận được, có thể thấy rằng mô hình dự báo giá cho ra kết quả nằm trong vùng dự báo, tuy nhiên giá dự báo lại chênh lệch thấp hơn so với giá thực tế và các chỉ số giá được dự báo không chênh lệch nhau nhiều.

Mặc dù các hệ số thống kê của bộ dữ liệu phù hợp với mô hình ARIMA, nhưng kết quả dự báo mô hình đưa ra lại không sử dụng được vì có thể thấy từ trong biểu đồ thì giá dự báo chênh lệch lớn,. Gần như là giá dự báo không thay đổi giữa các giá trị dự báo.

Kết luận: Mô hình ARIMA không phù hợp trong trường hợp dự báo giá cổ phiếu STB, nên sử dụng các phương pháp khác.

VI. Kết luận

Thị trường chứng khoán, cổ phiếu những năm qua đã chứng kiến những biến động lớn cả về điểm số và thanh khoản khi mà phải chịu những áp lực từ thị trường quốc tế vì các nước lớn có xu hướng thắt chặt tiền tệ để kiềm chế lạm phát và nguy cơ suy giảm kinh tế toàn cầu. Song thị trường chứng khoán Việt Nam trong năm 2023 vẫn có tiềm năng và được kỳ vọng sẽ hồi phục và tăng trưởng bền vững. Việt Nam vẫn đang duy trì việc tăng trưởng kinh tế ở mức độ cao so với thế giới, các doanh nghiệp đang cho thấy sức chống chịu tốt và khả quan trong thời gian biến động, thị trường Việt Nam vẫn là một thị trường hấp dẫn đối với các nhà đầu tư.

Việc thực hiện dự báo giá chứng khoán và cổ phiếu dự trên các mô hình hồi quy và mô hình máy học đang dần trở nên phổ biến rộng rãi vì những ưu thế và độ chính xác mà nó mang lại. Các nhà đầu tư có thể dùng những công cụ này để dự báo và đưa ra những quyết định đầu tư đúng đắn hơn. Tuy nhiên việc sử dụng mô hình hợp lí cho việc dự báo giá cần phải xem xét nhiều yếu tố để có thể sử dụng hợp lí. Ngoài việc sử dụng các dữ liệu để dự báo, các nhà đầu tư cần phải quan tâm nhiều đến các yếu tố thực tế, bối cảnh thị trường và tình hình kinh tế thế giới để có thể đưa ra những quyết định sáng suốt trong việc đầu tư.

VII. Code

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from vnstock import listing_companies, ticker_overview, stock_historical_data
from plotly.subplots import make_subplots
import plotly.graph_objects as go
from scipy.signal import argrelextrema
from collections import deque
from matplotlib.lines import Line2D
from statsmodels.tsa.stattools import adfuller
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
```

```
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.tsa.arima_model import ARIMA
import pmdarima as pm
import scipy.stats as scs
import statsmodels.api as sm
from datetime import timedelta
import warnings
import seaborn as sns
from numpy import log
from sklearn.metrics import mean_squared_error


# Lấy dữ liệu

df =  stock_historical_data(symbol='STB', start_date="2021-01-01", end_date='2022-12-31')

df.head()

# Chuyển đổi index

df['TradingDate'] = pd.to_datetime(df['TradingDate'])

df = df.set_index('TradingDate')

df.head()

# Vẽ biểu đồ

fig3 = make_subplots(specs=[[{"secondary_y": True}]])

fig3.add_trace(go.Candlestick(x=df.index,

                              open=df['Open'],

                              high=df['High'],

                              low=df['Low'],

                              close=df['Close'],

                              name='Price'))


df['diff'] = df['Close'] - df['Open']

df.loc[df['diff']>=0, 'color'] = 'green'

df.loc[df['diff']<0, 'color'] = 'red'
```

```
fig3.add_trace(go.Scatter(x=df.index,y=df['Close'].rolling(window=20).mean(),marker_color='blue',name='SMA'))

fig3.add_trace(go.Bar(x=df.index, y=df['Volume'], name='Volume',
marker={'color':df['color']}),secondary_y=True)

fig3.update_yaxes(range=[0,700000000],secondary_y=True)

fig3.update_yaxes(visible=False, secondary_y=True)

fig3.update_layout(xaxis_rangeslider_visible=False)  #hide range slider

fig3.update_layout(title={'text':'Chứng khoán STB', 'x':0.5})


fig3.layout.xaxis.type = 'category'


fig3.show()


# Dữ liệu cho mô hình

df_arima =  stock_historical_data(symbol='STB', start_date="2021-01-01",
end_date='2022-12-31')

df_arima = df_arima[['TradingDate','Close']]


# Loại bỏ dữ liệu không cần thiết

df_arima['TradingDate'] = pd.to_datetime(df_arima['TradingDate'])

df_arima = df_arima.set_index('TradingDate')

df_arima.head()


# Chuyển dữ liệu

df_arima = df_arima.resample('W-Fri').ffill()

window_size = 12

# Phân rã
```

```
WINDOW_SIZE = 12
df_arima['rolling_mean'] = df_arima.Close.rolling(window=WINDOW_SIZE).mean()
df_arima['rolling_std'] = df_arima.Close.rolling(window=WINDOW_SIZE).std()
df_arima.plot(title='Phân rã dữ liệu STB')
plt.tight_layout()
plt.grid(True)
plt.tight_layout()
plt.show()

plt.rcParams['figure.figsize'] = (20, 10)
decompose_results = seasonal_decompose(df_arima['Close'],model='multiplicative')
decompose_results.plot()

plt.show()

# Kiểm định
#tạo function adf_test
def stationary_test(a):

    stas_index = ['Test Statistic', 'P-value', 'Numbers of Lags Used', 'Numbers of
Observations Used']

    test_adf = adfuller(a, autolag='AIC')
    results = pd.Series(test_adf[0:4], index = stas_index)

    for key, value in test_adf[4].items():
        results[f'Critical Value ({key})'] = value
```

```
    return results
stationary_test(df_arima['Close'])

df_arima['diff'] = df_arima['Close'].diff(1)
df_arima

df_arima.dropna(inplace=True)
df_arima

stationary_test(df_arima['diff'])

fig,ax=plt.subplots(2,figsize=(20,10))
plot_acf(df_arima['diff'],ax=ax[0],lags=40,alpha=0.05)
plot_pacf(df_arima['diff'],ax=ax[1],lags=40,alpha=0.05)


df_arima= df_arima.dropna()

arima = sm.tsa.arima.ARIMA(df_arima.Close, order = (1, 1, 0)).fit()
print(arima.summary())

auto_arima = pm.auto_arima(df_arima['Close'], trace = 1,
                           error_action = 'ignore',
                           suppress_warnings = True,
                           seasonal = False,
                           stepwise = True,
```

```
                    approximation = False,
                    n_jobs = -1,
                    seasonal_test = False)

auto_arima.plot_diagnostics(figsize=(20, 12))
plt.show()

# Dự báo
# Chia tệp
train = df_arima['Close'][:'2022-11-30']
test = df_arima['Close']['2022-11-30':]

print(train.shape, test.shape)

# Huấn luyên mô hình
model = ARIMA(train, order=(1, 1, 0))
fitted = model.fit()

fc,se,conf= fitted.forecast(test.shape[0], alpha=0.05)

arima_pred = arima.forecast(n_forecasts,freq='W')

arima_pred = pd.DataFrame(arima_pred)
arima_pred.columns = ['Predict']

df_fc = pd.DataFrame({'Giá thực tết': list(test),
                    'giá dự báo': list(arima_pred.Predict)},
```

```
                    index= test.index)

fc_series = pd.Series(fc, index=test.index)
lower_series = pd.Series(conf[:, 0], index=test.index)
upper_series = pd.Series(conf[:, 1], index=test.index)

# Biểu đồ so sánh giá
plt.figure(figsize=(12,5), dpi=100)
plt.plot(train, label='training')
plt.plot(test, label='giá thực tế')
plt.plot(fc_series, label='giá dự báo')
plt.fill_between(lower_series.index, lower_series, upper_series,
                 color='k', alpha=.15)
plt.title('So sánh giá dự báo và giá thực tế')
plt.legend(loc='upper left', fontsize=8)
plt.show()
```

Tài liệu tham khảo:

https://phamdinhkhanh.github.io/2019/12/12/ARIMAmodel.html#1-gi%E1%BB%9Bi-thi%E1%BB%87u-v%E1%BB%81-chu%E1%BB%97i-th%E1%BB%9Di-gian

http://www1.vnua.edu.vn/tapchi/Upload/1452012-Tap%20chi%20so%202-22.pdf

https://laodong.vn/kinh-doanh/nhung-nhom-co-phieu-tiem-nang-mo-ra-co-hoi-dau-tu-nam-2023-1134555.ldo

https://thinhvu.com/2022/09/22/vnstock-api-tai-du-lieu-chung-khoan-python/

https://www.tinnhanhchungkhoan.vn/ssi-research-bat-mi-5-ma-tiem-nang-mo-man-nam-2023-post313244.html

https://baochinhphu.vn/thi-truong-chung-khoan-viet-nam-co-tiem-nang-lon-trong-dai-han-102220927154512519.htm

https://phamdinhkhanh.github.io/2019/12/12/ARIMAmodel.html

https://tapchitaichinh.vn/nam-2023-thi-truong-chung-khoan-viet-nam-hoi-phuc-va-tang-truong-ben-vung-hon.html

Sách tham khảo: "Ứng dụng Python trong tài chính" NXB Đại học quốc gia Tp.HCM, Nguyễn Anh Phong, NCT, Phan Huy Tâm, Ngô Phú Thanh biên soạn năm 2020.