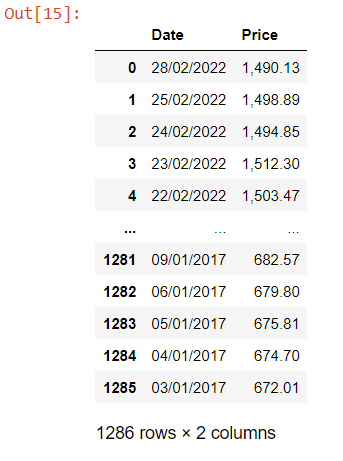
**Đề tài**

**ÁP DỤNG MÔ HÌNH ARIMA DỰ ĐOÁN VN-INDEX**

1. **Thu thập dữ liệu**

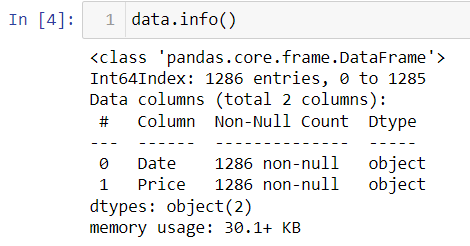
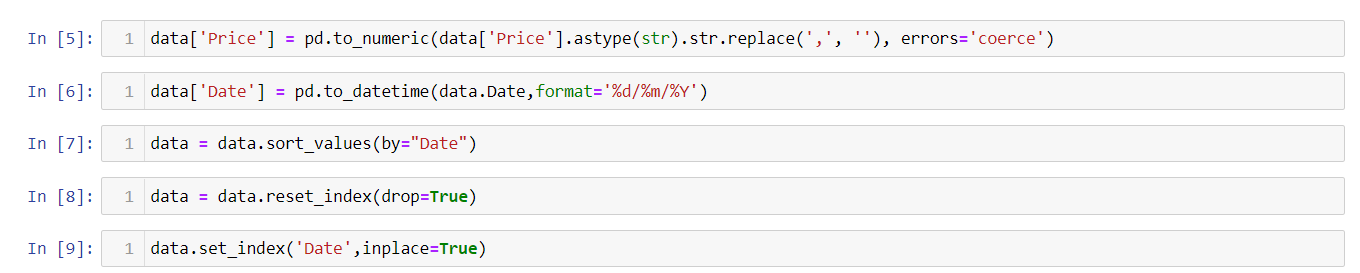
* Dữ liệu về VN-INDEX được thu thập từ nguồn https://vn.investing.com/indices/vn-historical-data
* Khoảng thời gian thu thập từ 01/01/2017 đến 28/02/2022.
* Sử dụng thư viện selenium để tiến hành thu thập dữ liệu.

Dữ liệu sau khi được thu về, tiến hành lấy ra những dữ liệu cần thiết cho việc xây dựng mô hình duwk báo, cụ thể là Date (Ngày giao dịch), và chỉ số VN-Index (được lấy theo giá đóng cửa của ngày hôm đó). Kết quả ta có dữ liệu thô sau:



1. **Xử lý dữ liệu**

Thông tin về dữ liệu ban đầu:

Dữ liệu ban đầu có 1286 dòng dữ liệu, không chứa giá trọ Null nhưng đang định dạng sai kiểu dữ liệu. Thực hiện chuyển đồi kiểu dữ liệu, đưa cột Price về kiểu Number và cột Date về kiểu Datatime

Kết quả sau khi xử lý dữ liệu



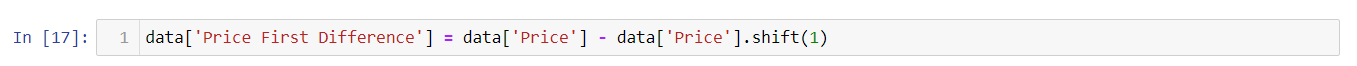
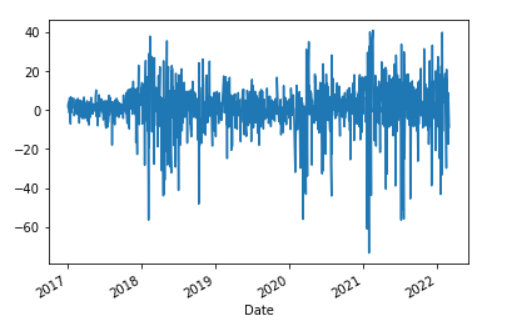
1. **Xây dựng mô hình**
   1. ***Kiểm định tính dừng của chuỗi dữ liệu***

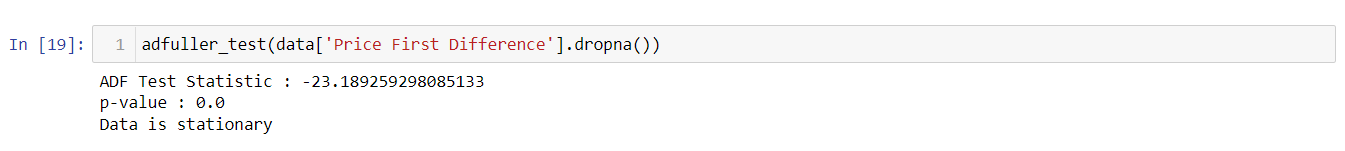
Sử dụng **ADF Test** để kiểm định tính dừng của chuỗi dữ liệu

Kết quả: chuổi dữ liệu là chuôi không dừng.

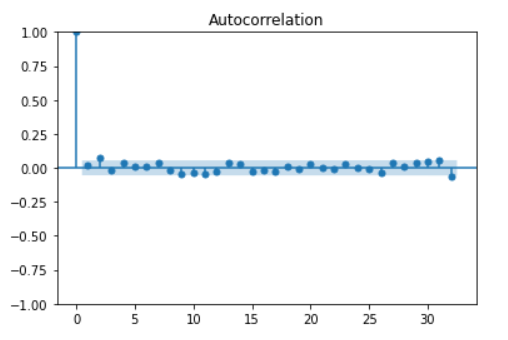
* 1. ***Kiểm định tính dừng của chuỗi sai phân bậc 1***

Lấy sai phân bậc 1 cho chuỗi dữ liệu

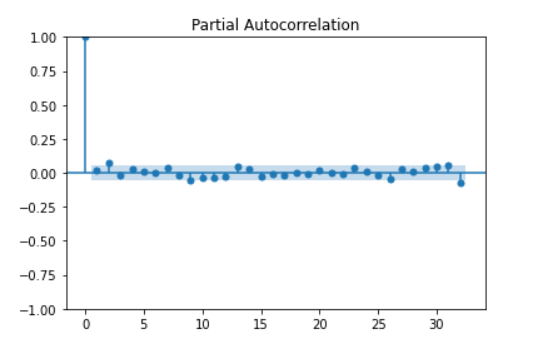
Biểu diễn chuỗi sai phân bậc 1

Kiểm định tính dừng của chuỗi sai phân bậc 1

Chuối sai phân bậc 1 là chuỗi dừng. Như vậy hệ số d trong mô hình ARIMA được xác định là 1.

* 1. ***Xác định hệ số p,q cho mô hình ARIMA***
* Sử dụng ACF để tìm hệ số q, độ trễ của quá trình trung bình trượt MA(q)

*Nhận xét:* Hệ số q được chọn là 2 vì tại độ trễ 2 độ dài đại diện cho giá trị của hệ số tự tương quan nằm ngoài khoảng tin cậy.

* Sử dụng PACF xác định hệ số p, hệ số bậc tự do p của quá trình tự hồi qui AR(p)

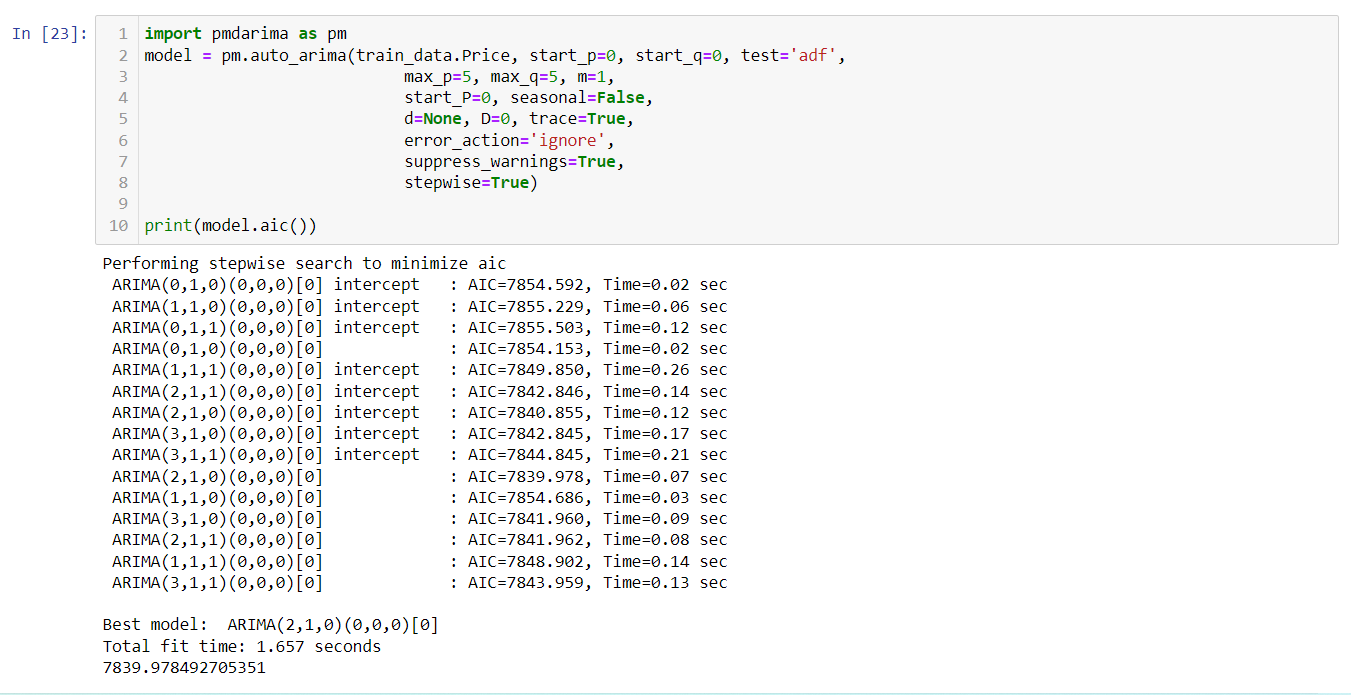
*Nhận xét*: Hệ số p được chọn là 2 vì tại độ trễ 2, độ dài đại diện cho giá trị của hệ số tự tương quan nằm ngoài khoảng tin cậy.

***Kết hợp giữa bậc của p và q và giá trị của d=1 ta có kịch bản ARIMA(2, 1, 2)***

* 1. ***Sử dụng Auto ARIMA tìm mô hình hồi quy tối ưu***

Auto ARIMA là thư viện dùng để tự động đề xuất ra mô hình ARIMA phù hợp nhất cho chuỗi dữ liệu đầu vào với hệ số AIC nhỏ nhất.

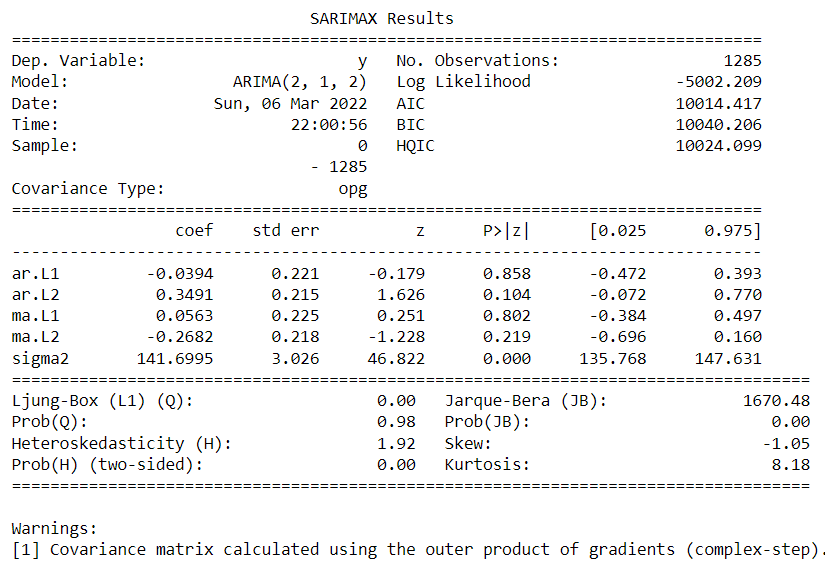
Kết quả sau khi xử lý bằng Auto ARIMA như sau:

Kết quả mô hình tốt nhất thu được là ARIMA(2,1,0) với chỉ số AIC = 7839.978

1. **Dự báo**

Dữ liệu ban đầu được chia ra 2 phần, 80% dùng làm dữ liệu train và 20% dùng làm dữ liệu test

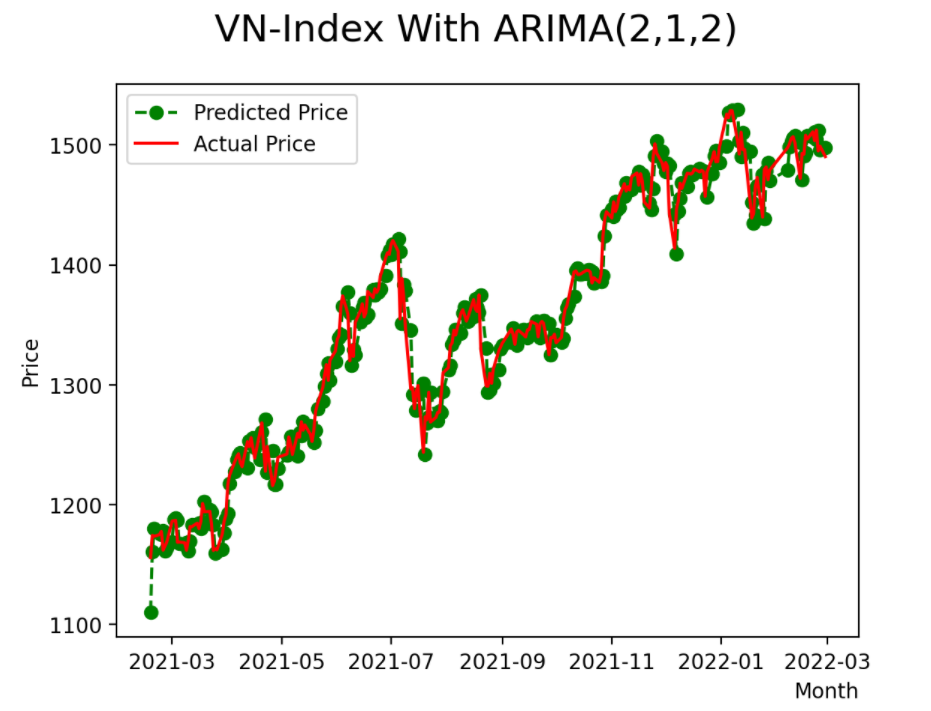
Như vậy ta có hai mô hình ARIMA(2,1,2) và ARIMA(2,1,0), tiến hành dự báo VN-Index bằng hao mô hình ARIMA trên và tìm ra mô hình tối ưu thông qua các chỉ số AIC, MSE, RMSE và MAE.

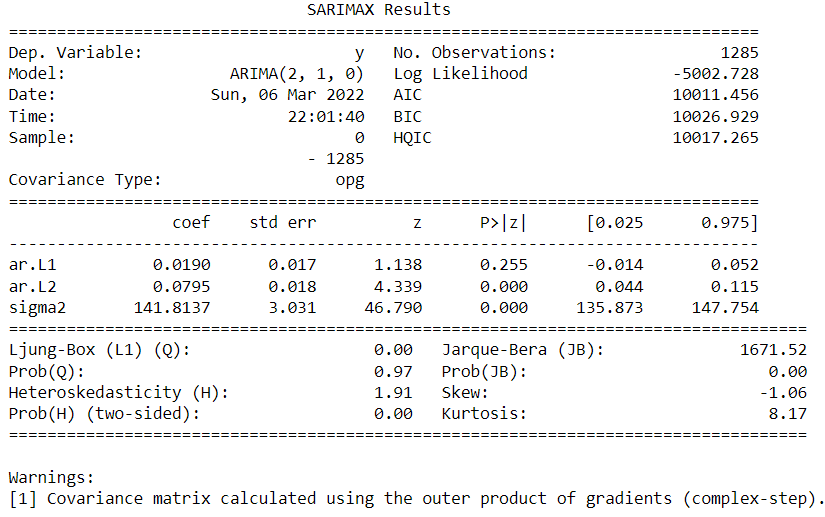
* 1. ***Kết quả dự báo bằng mô hình ARIMA(2,1,2)***

Testing RMSE: 15.134

Testing MSE: 229.038

MAE = 10.817975054919756

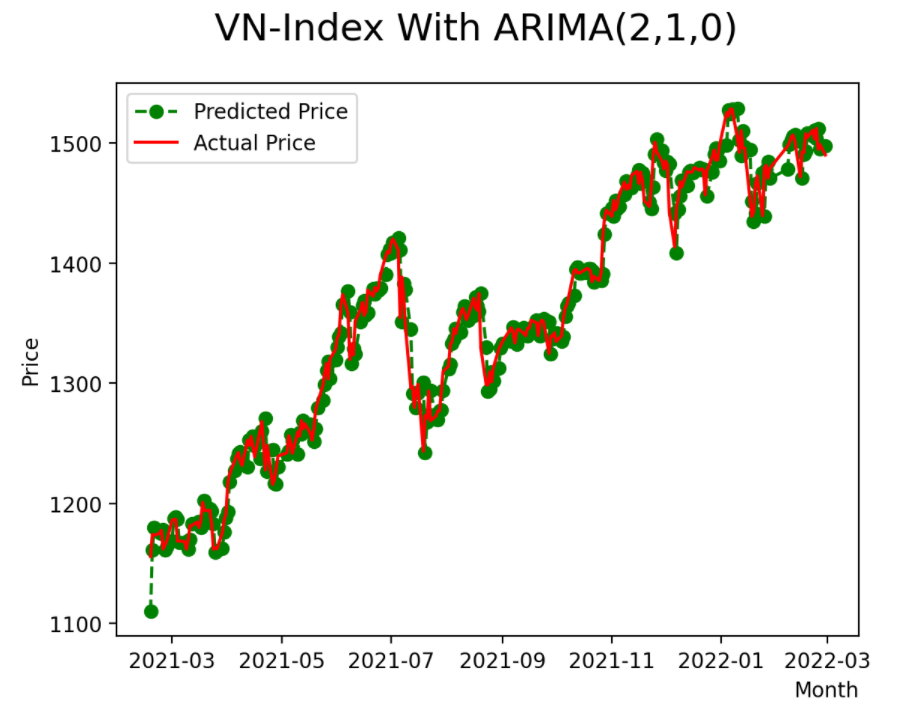


* 1. ***Kết quả dự báo bằng mô hình ARIMA(2,1,0)***

Testing RMSE: 15.105

Testing MSE: 228.150

MAE: 10.806



* 1. ***So sánh kết quả hai mô hình***

|  |  |  |
| --- | --- | --- |
| Chỉ số | ARIMA(2,1,2) | ARIMA(2,1,0) |
| AIC | 10014.417 | 10011.456 |
| RMSE | 15.134 | 15.105 |
| MSE | 229.038 | 228.1500 |
| MAE | 10.818 | 10.806 |

Các chỉ số AIC, MAE, RMSE, MSE của mô hình ARIMA(2,1,0) tốt hơn so với mô hình ARIMA(2,1,2) nên chọn mô hình ARIMA(2,1,0) cho dự báo VN-Index