ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỆ THỐNG THÔNG TIN**

****

**PHÂN TÍCH DỮ LIỆU KINH DOANH**

**BÀI THỰC HÀNH SỐ 4**

Giảng viên hướng dẫn: **PGS.TS. Nguyễn Đình Thuân**

**KS. Nguyễn Minh Nhựt**

Lớp: **IS403.N21**

Nhóm sinh viên thực hiện: **Nhóm 4**

|  |  |  |
| --- | --- | --- |
| **STT** | **Họ và tên** | **MSSV** |
| 1. | Thái Minh Triết | 19522397 |
| 2. | Nguyễn Ngọc Hiền | 20520496 |
| 3. | Nguyễn Tô Đức Tài | 20520743 |
| 4. | Nguyễn Thị Kim Liên | 20520909 |
| 5. | Trần Ngọc Linh | 20521538 |

**TP. HỒ CHÍ MINH, 05/2023**

NHẬN XÉT CỦA GIẢNG VIÊN

MỤC LỤC

[Chương 1. Cài đặt thuật toán ARIMA và thuật toán Học máy trên bộ dữ liệu Gas & Electric 1](#_Toc135132190)

[1.1. Thống kê bộ dữ liệu 1](#_Toc135132191)

[1.1.1. Thống kê mô tả 1](#_Toc135132192)

[1.1.2. Vẽ biểu đồ chuỗi thời gian 2](#_Toc135132193)

[1.1.3. Kiểm tra tính dừng của chuỗi thời gian 3](#_Toc135132194)

[1.2. Cài đặt thuật toán ARIMA 4](#_Toc135132195)

[1.2.1. MS Excel 4](#_Toc135132196)

[1.2.2. Python 13](#_Toc135132197)

[1.3. Cài đặt thuật toán Học máy – Support Vector Regressor 20](#_Toc135132198)

[1.3.1. MS Excel 20](#_Toc135132199)

[1.3.2. Python 20](#_Toc135132200)

[Chương 2. Cài đặt thuật toán ARIMA và thuật toán Học máy trên bộ dữ liệu về Việt Nam 25](#_Toc135132201)

[2.1. Tổng quan bộ dữ liệu 25](#_Toc135132202)

[2.1.1. Thống kê mô tả 25](#_Toc135132203)

[2.1.2. Vẽ biểu đồ chuỗi thời gian 26](#_Toc135132204)

[2.1.3. Kiểm tra tính dừng của chuỗi thời gian 26](#_Toc135132205)

[2.2. Cài đặt thuật toán ARIMA 27](#_Toc135132206)

[2.2.1. MS Excel 27](#_Toc135132207)

[2.2.2. Python 29](#_Toc135132208)

[2.3. Cài đặt thuật toán Học máy – Support Vector Regressor 32](#_Toc135132209)

[2.3.1. MS Excel 32](#_Toc135132210)

[2.3.2. Python 32](#_Toc135132211)

# Cài đặt thuật toán ARIMA và thuật toán Học máy trên bộ dữ liệu Gas & Electric

## Thống kê bộ dữ liệu

### Thống kê mô tả

|  |  |  |
| --- | --- | --- |
|  | **Gas Use** | **Electric Use** |
| Count | 24.00 | 24.00 |
| Mean | 107.79 | 945.25 |
| Std | 86.54 | 255.46 |
| Min | 26.00 | 672.00 |
| 25% | 32.25 | 800.25 |
| 50% | 70.50 | 854.00 |
| 75% | 190.75 | 1006.75 |
| Max | 247.00 | 1445.00 |

A picture containing text, screenshot, diagram, plot

Description automatically generated

A picture containing text, screenshot, diagram, line

Description automatically generated

### Vẽ biểu đồ chuỗi thời gian

* Gas Use

A picture containing diagram, plot, line, text

Description automatically generated

* Electric Use

A picture containing diagram, plot, line, text

Description automatically generated

### Kiểm tra tính dừng của chuỗi thời gian

Để kiểm tra tính dừng của chuỗi thời gian, ta sử dụng phép kiểm định Augmented Dicky-Fuller:

* Phát biểu giả thuyết không và đối thuyết:
  + H0: Chuỗi thời gian không có tính dừng
  + H1: Chuỗi thời gian có tính dừng
* **Gas Use:**

Sử dụng hàm adfuller trong thư viện statmodels bằng ngôn ngữ Python ta thu được kết quả:

A screen shot of a computer code

Description automatically generated with low confidence

Với p-value = 0.982985 > 0.05, ta không thể bác bỏ H0

**Kết luận:** Chuỗi thời gian Gas Use không có tính dừng (Non-stationary).

* **Electric Use:**

Sử dụng hàm adfuller trong thư viện statmodels bằng ngôn ngữ Python ta thu được kết quả:

A picture containing text, screenshot, font

Description automatically generated

Với p-value = 0.005 < 0.05, bác bỏ H0

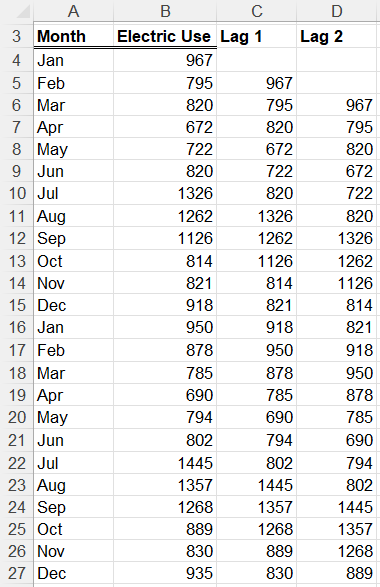
**Kết luận:** Chuỗi thời gian Electric Use có tính dừng (Stationary).

## Cài đặt thuật toán ARIMA

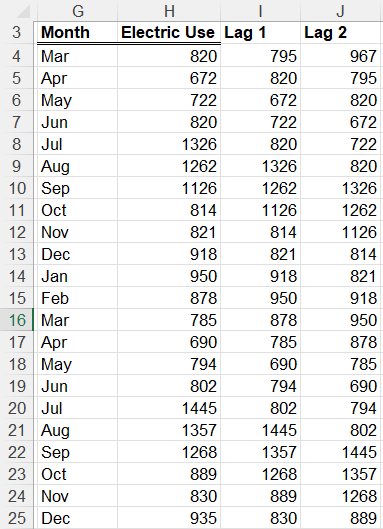
### MS Excel

- Dự đoán giá trị cột Electric Use:

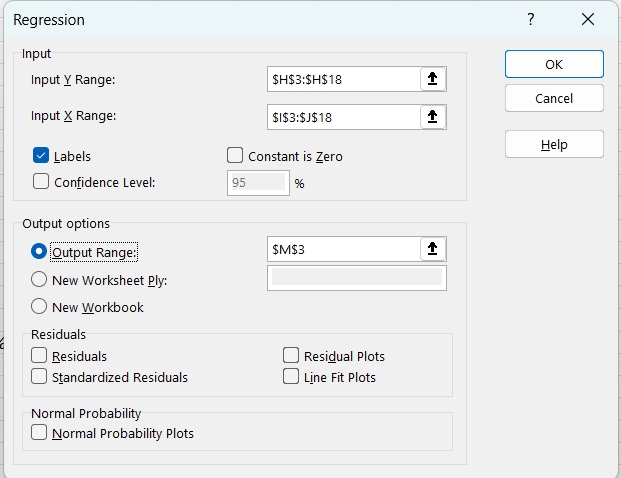
Lấy p=lag=2 tạo 2 cột lag1 đến lag2 tương ứng



Bỏ những dòng dữ liệu đầu tiên không đủ dữ liệu:



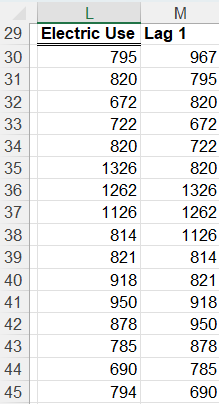
Sử dụng Data Analysis

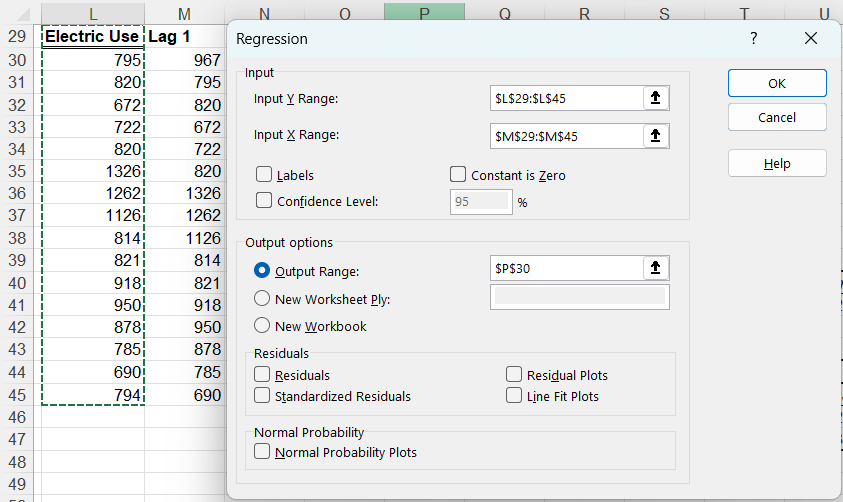


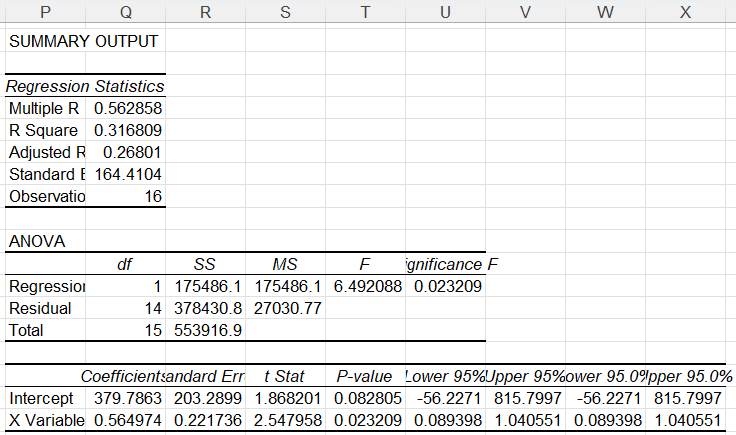
A screenshot of a spreadsheet

Description automatically generated with medium confidence

Vì giá trị p-value của lag2 lớn hơn 0.05 nên sẽ tính lại lag1

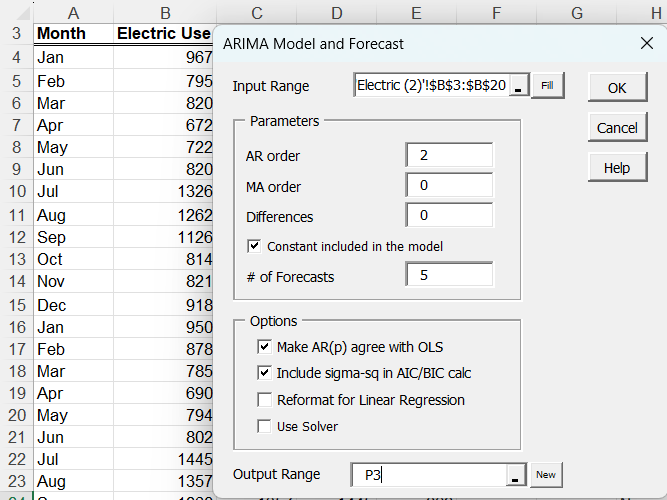


Sử dụng Data Analysis tính lại lag1



Vậy phương trình là:

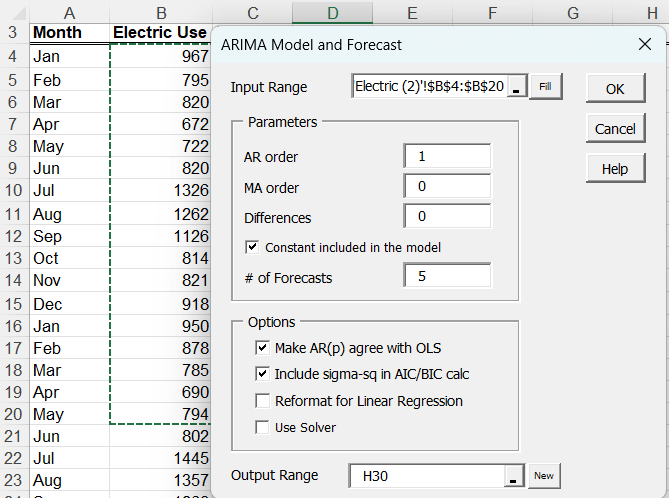
- Sử dụng Real Statistics:

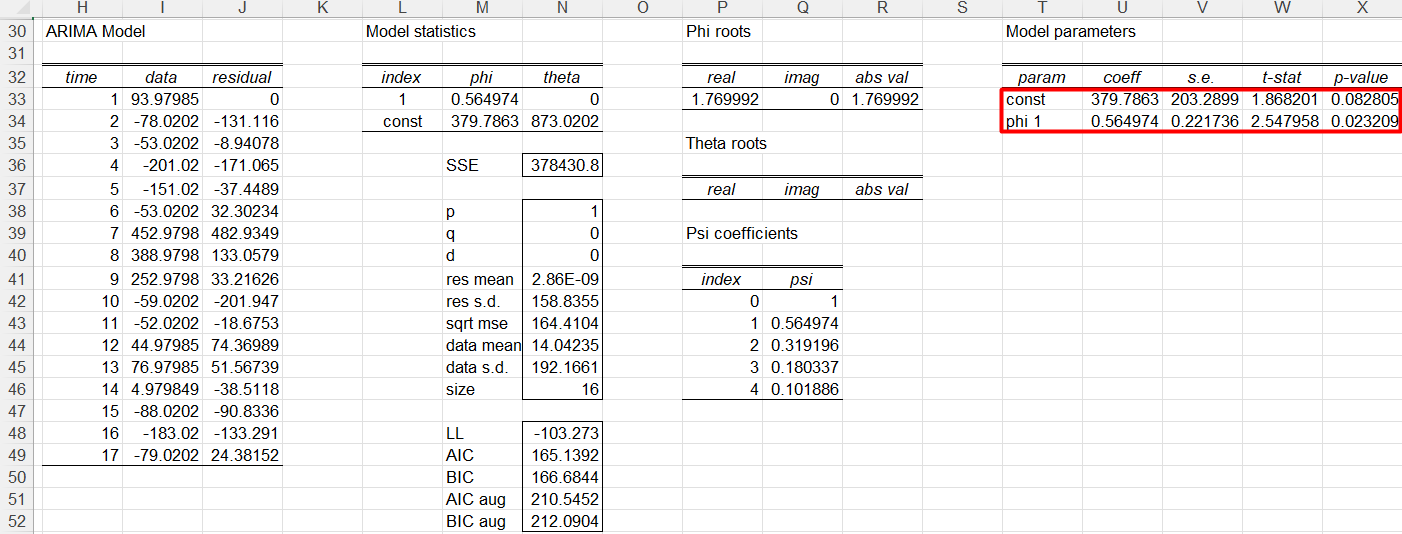


A screenshot of a computer

Description automatically generated with medium confidence

Kết quả cho ra giống với khi sử dụng Data Analysis giá trị p-value của phi 2 lớn hơn 0.05 nên sẽ tính lại lag

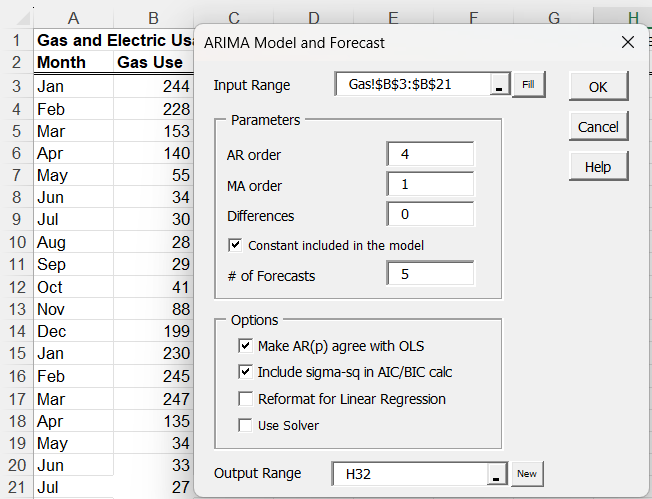


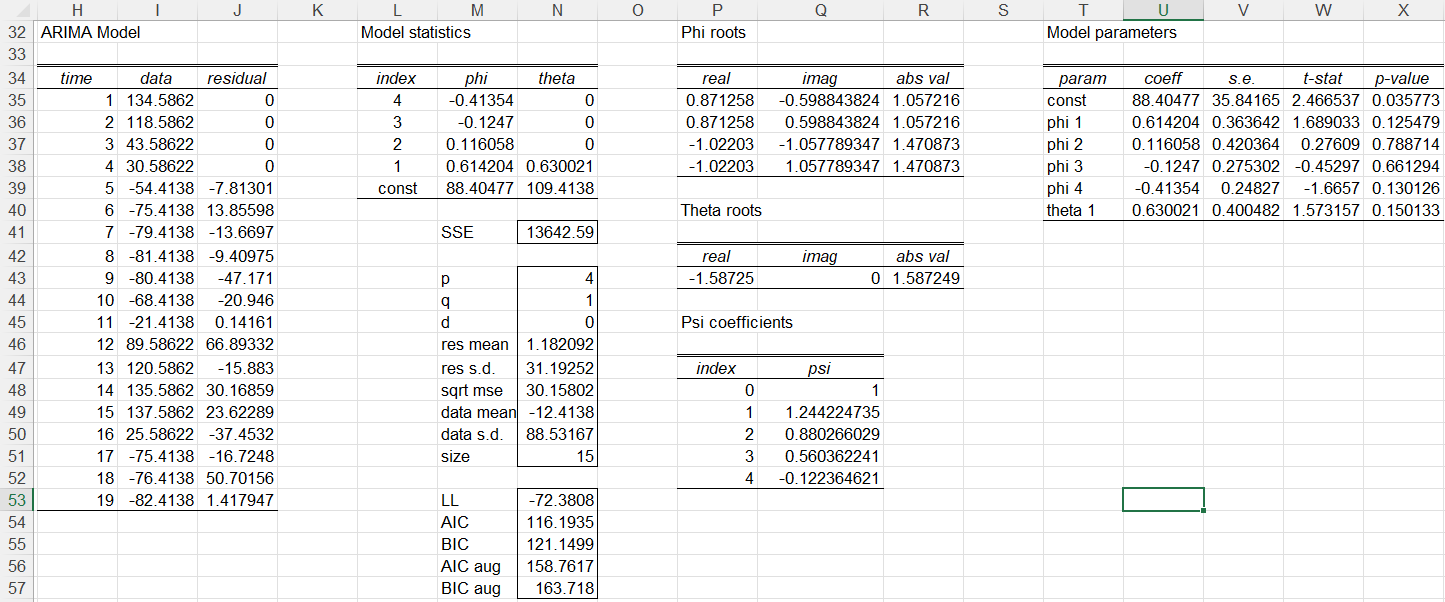


Kết quả cho ra giống với khi sử dụng Data Analysis

- Dự đoán giá trị cột Gas Use:

Sử dụng Real Statistics:





Vậy phương trình có dạng:

### Python

Dự đoán giá trị Gas Use:

- Import dữ liệu từ file csv sử dụng thư viện pandas.

A screenshot of a computer

Description automatically generated

- Chia dữ liệu thành tập train và tập test với tỉ lệ 8:2.

A picture containing text, screenshot, software, font

Description automatically generated

- Tìm mô hình ARIMA tốt nhất sử dụng auto\_arima.

* Import auto\_arima từ pmdarima.arima và tiến hành tìm mô hình ARIMA tốt nhất:

A screenshot of a computer

Description automatically generated

* Từ kết quả thu được, ta có p = 4, d = 0 và q = 1.

A screen shot of a computer

Description automatically generated with medium confidence

- Sử dụng câu lệnh model.summary() để theo dõi các tham số rút ra từ mô hình tối ưu.

A screenshot of a computer

Description automatically generated with medium confidence

- Tiến hành kiểm thử dữ liệu.

A picture containing text, font, screenshot

Description automatically generated

- Tạo index dự báo 24 tháng tiếp theo.

A screenshot of a computer program

Description automatically generated with medium confidence

- Đánh giá độ chính xác của tập test.

A screen shot of a computer code

Description automatically generated with low confidence

- Trực quan hóa giá trị dự báo của Gas Use trong 24 tháng tiếp theo.

A picture containing text, screenshot, font, software

Description automatically generated

A picture containing text, plot, line, diagram

Description automatically generated

Dự báo giá trị Electric Use:

- Tương tự như ở trên, ta cũng tiến hành import dữ liệu từ file csv bằng thư viện pandas.

A picture containing text, screenshot, font

Description automatically generated

- Chia dữ liệu thành tập train và tập test theo tỉ lệ 7:3.

A picture containing text, font, software, multimedia software

Description automatically generated

Tìm mô hình ARIMA tốt nhất sử dụng auto\_arima.

* Import auto\_arima từ pmdarima.arima và tiến hành tìm mô hình ARIMA tốt nhất:

A screenshot of a computer program

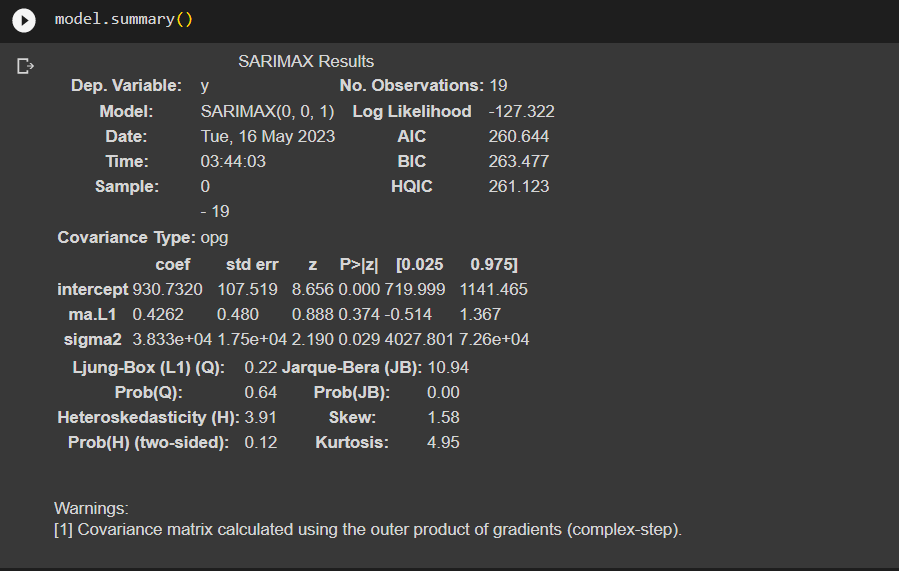
Description automatically generated with medium confidence

* Từ kết quả tìm được, ta có p = 2, d = 0 và q = 0.

A picture containing text, screenshot, font

Description automatically generated

- Sử dụng câu lệnh model.summary() để theo dõi các tham số rút ra từ mô hình tối ưu.



- Tiến hành kiểm thử dữ liệu.

A picture containing text, screenshot, font

Description automatically generated

- Tạo index dự báo 24 tháng tiếp theo.

A screenshot of a computer program

Description automatically generated with medium confidence

- Đánh giá độ chính xác của tập test.

A screen shot of a computer code

Description automatically generated with low confidence

- Trực quan hóa giá trị dự báo của Electric Use trong 24 tháng tiếp theo.

A picture containing text, screenshot, font, software

Description automatically generated

A picture containing text, plot, screenshot, diagram

Description automatically generated

## Cài đặt thuật toán Học máy – Support Vector Regressor

### Python

* Dự báo giá trị Gas Use
* Đọc dữ liệu từ file .csv sử dụng thư viện pandas

A screenshot of a computer

Description automatically generated

* Chia dữ liệu thành tập train và tập test với tỉ lệ 8:2

A picture containing text, screenshot, font, display

Description automatically generated

* Chuẩn hóa dữ liệu sử dụng MinmaxScaler

A picture containing text, screenshot, font, multimedia software

Description automatically generated

* Khởi tạo và huấn luyện mô hình Support Vector Regressor

A picture containing text, font, screenshot

Description automatically generated

* Dự báo kết quả trên tập test và đánh giá trên độ đo RMSE

A screen shot of a computer code

Description automatically generated with low confidence

* Dự báo 24 tháng tiếp theo và trực quan hóa giá trị dự báo

A picture containing text, screenshot, software, multimedia software

Description automatically generated

A picture containing text, plot, diagram, line

Description automatically generated

* Dự báo giá trị Electric Use
* Đọc dữ liệu từ file .csv sử dụng thư viện pandas

A screenshot of a computer

Description automatically generated

* Chia dữ liệu thành tập train và tập test với tỉ lệ 8:2

A screenshot of a computer program

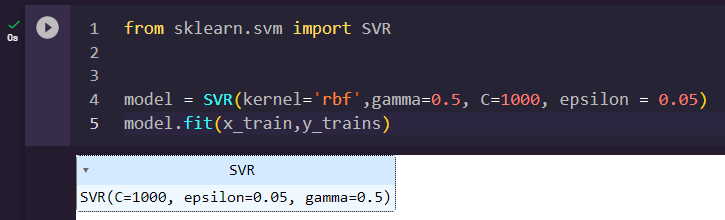
Description automatically generated with low confidence

* Chuẩn hóa dữ liệu sử dụng MinmaxScaler

A picture containing text, screenshot, font, multimedia software

Description automatically generated

* Khởi tạo và huấn luyện mô hình Support Vector Regressor



* Dự báo kết quả trên tập test và đánh giá trên độ đo RMSE

A picture containing text, screenshot, font

Description automatically generated

* Dự báo 24 tháng tiếp theo và trực quan hóa giá trị dự báo

A picture containing text, screenshot, software, multimedia software

Description automatically generated

A picture containing text, screenshot, plot, diagram

Description automatically generated

# Cài đặt thuật toán ARIMA và thuật toán Học máy trên bộ dữ liệu về Việt Nam

## Tổng quan bộ dữ liệu

* Tên bộ dữ liệu: Giá cố phiếu Tập đoàn FPT.
* Nguồn dữ liệu: https://www.investing.com/equities/fpt-corp-historical-data.
* Thời gian: từ 01/2007 tới 05/2023 (thống kê theo tháng).
* Đơn vị: VND.

### A picture containing text, screenshot, diagram, plot Description automatically generatedA picture containing text, screenshot, diagram, plot Description automatically generatedThống kê mô tả

|  |  |
| --- | --- |
|  | **Price** |
| Count | 197.00 |
| Mean | 23764.95 |
| Std | 23444.74 |
| Min | 3104.50 |
| 25% | 8263.70 |
| 50% | 14207.00 |
| 75% | 25909.80 |
| Max | 90530.60 |

### Vẽ biểu đồ chuỗi thời gian

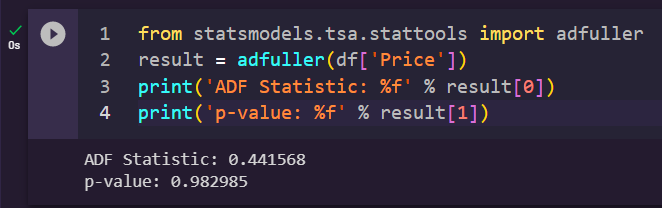
A picture containing text, diagram, plot, line

Description automatically generated

### Kiểm tra tính dừng của chuỗi thời gian

Để kiểm tra tính dừng của chuỗi thời gian, ta sử dụng phép kiểm định Augmented Dicky-Fuller:

* Phát biểu giả thuyết không và đối thuyết:
  + H0: Chuỗi thời gian không có tính dừng
  + H1: Chuỗi thời gian có tính dừng
* Sử dụng hàm adfuller trong thư viện statmodels bằng ngôn ngữ Python ta thu được kết quả:



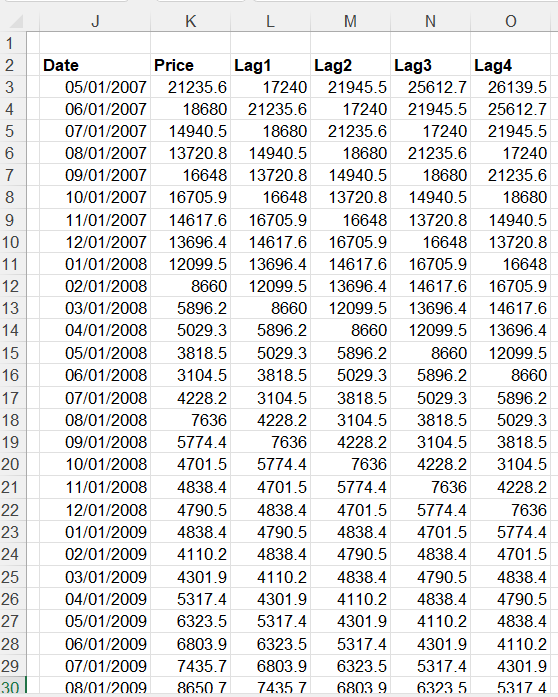
Với p-value = 0.982985 > 0.05, ta không thể bác bỏ H0

**Kết luận:** Chuỗi thời gian không có tính dừng (Non-stationary).

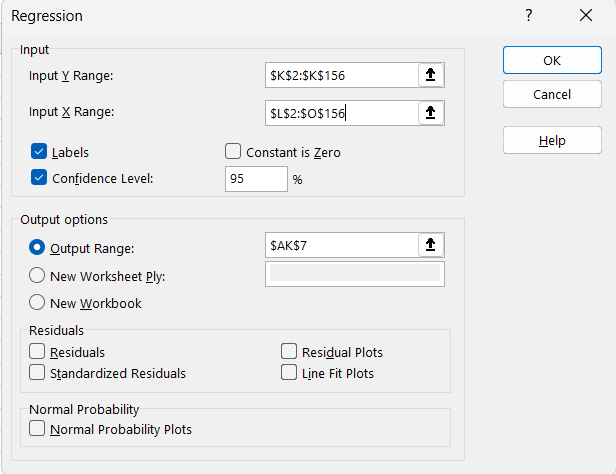
## Cài đặt thuật toán ARIMA

### MS Excel

Ta chọn 4 lag để tìm ra thông số p phù hợp với mô hình:



Sử dụng Data Anlysis



Ta được kết quả:

A screenshot of a data

Description automatically generated with low confidence

Nhận xét: với lag1, lag2, lag3 đều có p-value >0.05. Ta thực hiện lại với chỉ lag1.

A picture containing text, font, number, screenshot

Description automatically generated

Sử dụng Data Analysis:

A screenshot of a computer

Description automatically generated with medium confidence

Kết quả nhân được:

A screenshot of a graph

Description automatically generated with low confidence

Kết luận: từ kết quả trên ta có thể chọn p phù hợp với mô hình là 1.

* Sử dụng Real Statistics với tham số p,d,q là (1,2,2):

A screenshot of a spreadsheet

Description automatically generated

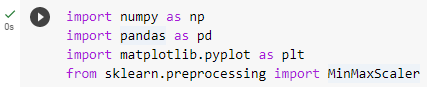
A screenshot of a computer

Description automatically generated with medium confidence

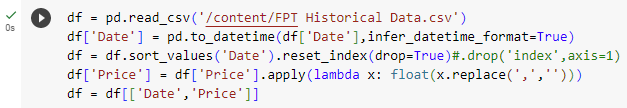
Vậy phương trình có dạng:

### Python

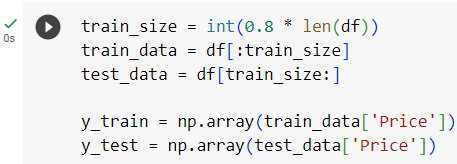
* Import các thư viện cần thiết



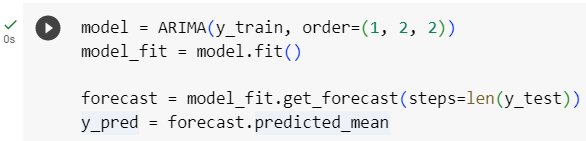
* Đọc dữ liệu từ file csv sử dụng thư viện pandas



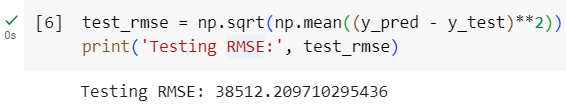
* Chia dữ liệu thành tập train và tập test với tỉ lệ 8:2



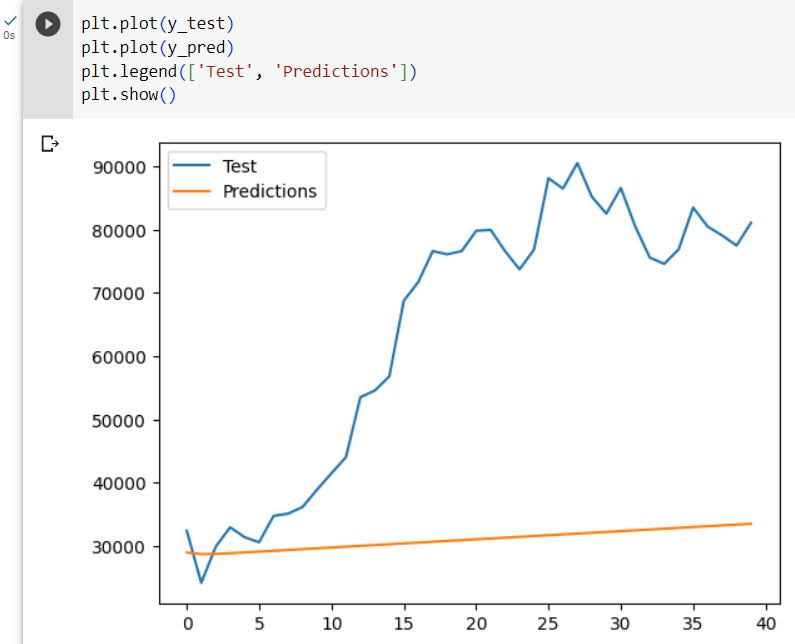
* Khởi tạo và huấn luyện mô hình ARIMA



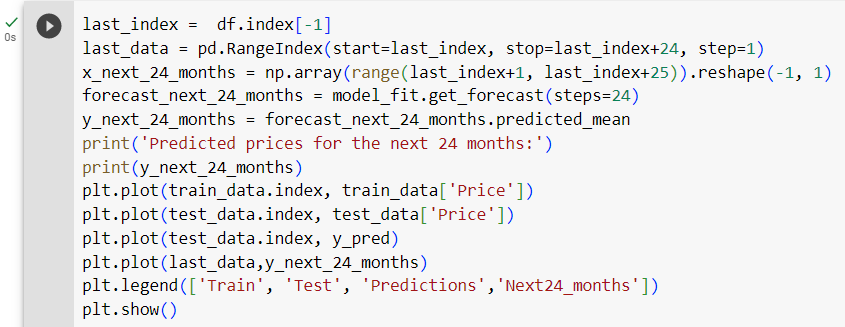
* Dự báo kết quả trên tập test và đánh giá trên độ đo RMSE

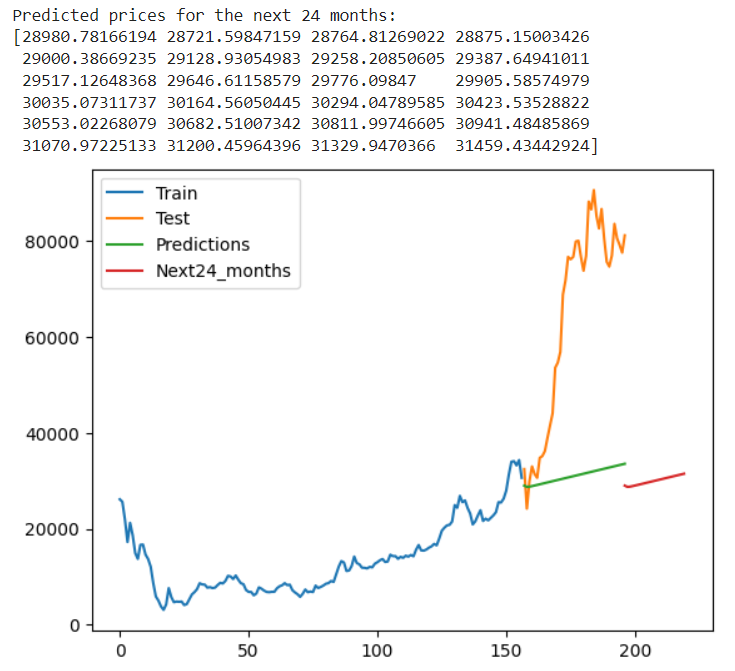


Trực quan hóa:



* Dự báo 24 tháng tiếp theo và trực quan hóa giá trị dự báo





## Cài đặt thuật toán Học máy – Support Vector Regressor

### Python

* Đọc dữ liệu từ file csv sử dụng thư viện pandas

A screen shot of a computer

Description automatically generated with medium confidence

* Chia dữ liệu thành tập train và tập test với tỉ lệ 8:2

A screen shot of a computer program

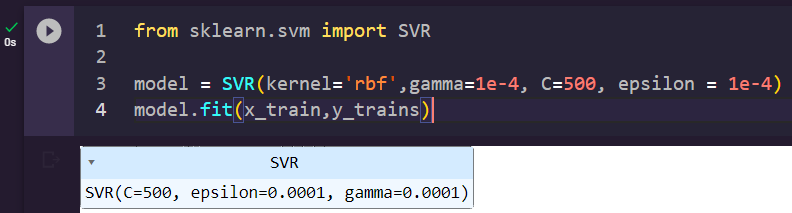
Description automatically generated with low confidence

* Chuẩn hóa dữ liệu sử dụng MinmaxScaler

A screenshot of a computer code

Description automatically generated with low confidence

* Khởi tạo và huấn luyện mô hình Support Vector Regressor:
  + kernel = “rbf”
  + gamma = 1e-4
  + epsilon = 1e-4
  + C = 500



* Dự báo trên tập test và đánh giá kết quả trên độ đo RMSE

A screen shot of a computer code

Description automatically generated with low confidence

* Dự báo 24 tháng tiếp theo tính từ tháng 06/2023

A screenshot of a computer program

Description automatically generated with low confidence

* Trực quan hóa giá trị dự báo trên tập test và 24 tháng tiếp theo

A screen shot of a computer code

Description automatically generated with low confidence

A picture containing text, screenshot, diagram, plot

Description automatically generated

BẢNG PHÂN CÔNG CÔNG VIỆC

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Công việc** | | **Minh Triết** | **Kim Liên** | **Ngọc Hiền** | **Ngọc Linh** | **Đức Tài** |
| Tổng quan bộ dữ liệu Gas & Electric | | **X** |  |  |  |  |
| Cài đặt thuật toán ARIMA trên bộ dữ liệu Gas & Electric | MS Excel |  | **X** |  |  |  |
| Python |  |  | **X** |  |  |
| Cài đặt thuật toán Học máy | | **X** |  |  |  |  |
| Tổng quan bộ dữ liệu giá cổ phiếu FPT | | **X** |  |  |  |  |
| Cài đặt thuật toán ARIMA trên bộ dữ liệu giá cổ phiếu FPT | MS Excel |  |  |  |  | **X** |
| Python |  |  |  | **X** |  |
| Cài đặt thuật toán Học máy | | **X** |  |  |  |  |
| **Hoàn thành** | | **100%** | **100%** | **100%** | **100%** | **100%** |

TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | Slide “LAB04. Phân tích chuỗi thời gian” - Nguyễn Minh Nhựt |

[2] Quick way to find p, d and q values for ARIMA.

https://analyticsindiamag.com/quick-way-to-find-p-d-and-q-values-for-arima/

[3] Bài 19 - Mô hình ARIMA trong time series.

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

[4] Auto ARIMA – API reference.

https://alkaline-ml.com/pmdarima/modules/generated/pmdarima.arima.auto\_arima