A picture containing rectangle, screenshot, text, frame

Description automatically generated

**Giảng viên lý thuyết:** PGS.TS Nguyễn Đình Thuân

**Giảng viên thực hành**: Nguyễn Minh Nhật

**Nhóm sinh viên thực hiện**: Nhóm 20

1. Cao Hoài Sang MSSV: 21522541

2. Thi Thành Công MSSV: 21521897

3. Nguyễn Trần Gia Kiệt MSSV: 21522258

4. Nguyễn Hoàng Đăng Khoa MSSV: 21520999

5. Cù Ngọc Hoàng MSSV: 21552086

**Môn học: PHÂN TÍCH DỮ LIỆU KINH DOANH**

ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA HỆ THỐNG THÔNG TIN

BÁO CÁO LAB04

**BẢNG PHÂN CÔNG CÔNG VIỆC**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cao Hoài Sang | Thi Thành Công | Nguyễn Trần Gia Kiệt | Cù Ngọc Hoàng | Nguyễn Hoàng Đăng Khoa |
| 4.1 |  |  |  |  | x |
| 4.2 |  |  |  | x |  |
| 4.3 | x | x | x |  |  |

**CÁC TÀI NGUYÊN KHÁC**

|  |  |
| --- | --- |
| Tên tài nguyên | Đường dẫn |
| Mã nguồn | [Ấn vào đây](https://uithcm-my.sharepoint.com/:f:/g/personal/21522541_ms_uit_edu_vn/EjH7fEKJ2GlGqxgxPS1EtBsBd2fv39cie7hR4cgWfJkQ9g?e=W3wmSt) |

**Mục Lục**

[1. Câu 1 3](#_Toc165582580)

[a. Yêu cầu: 3](#_Toc165582581)

[b. Bài giải: 3](#_Toc165582582)

[2. Câu 2 24](#_Toc165582583)

[a. Yêu cầu: 24](#_Toc165582584)

[b. Bài giải: 25](#_Toc165582585)

[i. ARIMA 25](#_Toc165582586)

[ii. SES - Simple exponential smoothing 32](#_Toc165582587)

[3. Câu 3 33](#_Toc165582588)

[a. Yêu cầu: 33](#_Toc165582589)

[b. Bài giải 33](#_Toc165582590)

[i. Các phương thức đánh giá mô hình 33](#_Toc165582591)

[ii. Áp dụng cho Lab2 37](#_Toc165582592)

[iii. Áp dụng cho Lab3 38](#_Toc165582593)

[iv. Áp dụng cho 4.1 và 4.2 41](#_Toc165582594)

*Danh sách hình ảnh*

[Figure 1. Dataset 7](#_Toc165717133)

[Figure 2. Nhận diện chuỗi dừng đồ thị 8](#_Toc165717134)

[Figure 3. Đồ thị Gas Use (Excel) 8](#_Toc165717135)

[Figure 4. Đồ thị Gas Use (d =1) 9](#_Toc165717136)

[Figure 5. Đồ thị Gas Use (d=2) 9](#_Toc165717137)

[Figure 6. Độ trễ tưng lag 10](#_Toc165717138)

[Figure 7. Tổng kết giá trị tưng phần 10](#_Toc165717139)

[Figure 8. Kết quá ACF của từng lag 11](#_Toc165717140)

[Figure 9. Đồ thị ACF Gas Use (Excel) 11](#_Toc165717141)

[Figure 10. Ma trận 12](#_Toc165717142)

[Figure 11. Tính PACF từng lag 12](#_Toc165717143)

[Figure 12. Đồ thị Gas Use PACF (Excel) 12](#_Toc165717144)

[Figure 13. So sánh giữa giá trị thực thế và dự báo (Gas Use) 14](#_Toc165717145)

[Figure 14. Đồ thị so sánh giữa thực tế và dự báo (Gas Use) 15](#_Toc165717146)

[Figure 15. Đồ thị Electric Use (Excel) 15](#_Toc165717147)

[Figure 16.Giá trị Electric Use của ACF và PACF 16](#_Toc165717148)

[Figure 17. Đồ thị Electric Use ACF (Excel) 16](#_Toc165717149)

[Figure 18. Đồ thị Electric Use PACF (Excel) 16](#_Toc165717150)

[Figure 19. So sánh giữa giá trị thực tế và dự đoán Electric Use 17](#_Toc165717151)

[Figure 20. Đồ thị so sánh giữa giá trị thực tế và dự đoán Electric Use (Excel) 18](#_Toc165717152)

[Figure 21. So sánh Gas Use các dự báo của các hệ số a 18](#_Toc165717153)

[Figure 22. Đồ thị Gas Use so sánh các hệ số a 19](#_Toc165717154)

[Figure 23. So sánh Electric Use các dự báo của các hệ số a 19](#_Toc165717155)

[Figure 24. Đồ thị Electric Use so sánh các hệ số a 19](#_Toc165717156)

[Figure 25. Tạo Data frame đối với mỗi cột 20](#_Toc165717157)

[Figure 26. Kiểm định ADF cho gas\_data 20](#_Toc165717158)

[Figure 27. Kết quả ADF sau sai phân 21](#_Toc165717159)

[Figure 28. Đồ thị Gas Use ACF (Python) 21](#_Toc165717160)

[Figure 29. Đồ thị Gas Use PACF (Python) 21](#_Toc165717161)

[Figure 30. Chia và huấn luyện mô hình ARIMA 22](#_Toc165717162)

[Figure 31. Đồ thị Gas Use kết quả dự đoán ARIMA 22](#_Toc165717163)

[Figure 32. Huấn luyện mô hình SARIMA 22](#_Toc165717164)

[Figure 33. Đồ thị Gas Use kết quả dự đoán mô hình SARIMA 23](#_Toc165717165)

[Figure 34. Kiểm định ADF cho electricity\_data 23](#_Toc165717166)

[Figure 35. Đồ thị Electric Use ACF (Python) 24](#_Toc165717167)

[Figure 36. Đồ thị Electric Use PACF (Python) 24](#_Toc165717168)

[Figure 37. Chia và huấn luyệ mô hình ARIMA 24](#_Toc165717169)

[Figure 38. Đồ thị Electric Use so sánh giữa giá trị thực tế và dự báo (Python) 25](#_Toc165717170)

[Figure 39. So sánh giữa giá trị dụ đoán và thực tế 25](#_Toc165717171)

[Figure 40. Mô hình SES 25](#_Toc165717172)

[Figure 41. Đồ thị SES của Electric Use và Gas Use 26](#_Toc165717173)

[Figure 42. So sánh SES Gas Use 26](#_Toc165717174)

[Figure 43. So sánh SES Electric Use 26](#_Toc165717175)

[Figure 44 Dữ liệu về các chỉ số không khí tại Hà Nội từ 1/1/2014 - 25/3/2023 27](#_Toc165717176)

[Figure 45 AQI,lagged, diff AQI 28](#_Toc165717177)

[Figure 46 Kết quả hồi qui tuyến tính 29](#_Toc165717178)

[Figure 47 ACF tại các lag 29](#_Toc165717179)

[Figure 48 Biểu đồ ACF 30](#_Toc165717180)

[Figure 49 PACF tại các lag 30](#_Toc165717181)

[Figure 50 Biểu đồ PACF 31](#_Toc165717182)

[Figure 51 Dự đoán trên tập test của mô hình ARIMA(1,0,1) 31](#_Toc165717183)

[Figure 52 Biểu đồ giữa dữ liệu dự đoán và thực tế của ARIMA(1,0,1) 32](#_Toc165717184)

[Figure 53 Dự đoán trên tập test của mô hình ARIMA(2,0,1) 32](#_Toc165717185)

[Figure 54 Biểu đồ giữa dữ liệu dự đoán và thực tế của ARIMA(2,0,1) 33](#_Toc165717186)

[Figure 55 Kiểm định tính dừng bằng python 33](#_Toc165717187)

[Figure 56 Vẽ biểu đồ ACF bằng python 34](#_Toc165717188)

[Figure 57. Vẽ biểu đồ PACF bằng python 34](#_Toc165717189)

[Figure 58 Kết quả dự đoán với ARIMA(1,0,1) 35](#_Toc165717190)

[Figure 59. Kết quả dự đoán với ARIMA(2,0,1) 35](#_Toc165717191)

[Figure 60 Giá trị dự đoán tại các a khác nhau 36](#_Toc165717192)

[Figure 61 Hàm SES bằng python 36](#_Toc165717193)

[Figure 62. Kết quả dự doán bằng SES 36](#_Toc165717194)

[Figure 63 Làm sạch, tách tập train và test 41](#_Toc165717195)

[Figure 64 Sử dụng model LinearRegression của thư viện Sklearn 41](#_Toc165717196)

[Figure 65 Kết quả đánh giá 41](#_Toc165717197)

[Figure 66 Thêm các cột mới 42](#_Toc165717198)

[Figure 67 Thay đổi tên cột 43](#_Toc165717199)

[Figure 68 thêm dữ liệu và huấn luyện mô hình 44](#_Toc165717200)

[Figure 69 Tính RMSE 44](#_Toc165717201)

[Figure 70 Xây dựng và sử dụng hàm tính MAPE 44](#_Toc165717202)

[Figure 71 Chia dataset làm 2 phần 45](#_Toc165717203)

[Figure 72 Chia tập train/test cho Gas 45](#_Toc165717204)

[Figure 73 Chia tập train/test cho Electric 45](#_Toc165717205)

[Figure 74 Huấn luyện mô hình và quá trình test trên Gas 46](#_Toc165717206)

[Figure 75 Huấn luyện mô hình và quá trình test trên Electric 46](#_Toc165717207)

[Figure 76 Đánh giá mô hình ARIMA trên Gas 46](#_Toc165717208)

[Figure 77 Đánh giá mô hình ARIMA trên Electric 47](#_Toc165717209)

[Figure 78 Kiểm định mô hình với các giá trị RMSE, MAPE, MSE 47](#_Toc165717210)

[Figure 79 Kiểm định mô hình với các giá trị RMSE, MAPE, MSE 47](#_Toc165717211)

[Figure 80 Thực hiện kiểm định với phần sử dụng giá trị "Gas Use" 48](#_Toc165717212)

[Figure 81 Thực hiện kiểm định với phần sử dụng giá trị "Electric Use" 48](#_Toc165717213)

[Figure 82 Thực hiện kiểm định SES với python 49](#_Toc165717214)

[Figure 83 Tính toán 51](#_Toc165717215)

[Figure 84 Tính RMSE 51](#_Toc165717216)

[Figure 85 Tính toán 52](#_Toc165717217)

[Figure 86 Tính toán lần lượt các giá trị sum và n 53](#_Toc165717218)

[Figure 87 Giá trị MAPE 53](#_Toc165717219)

[Figure 88 Các giá trị RMSE, MAPE, MSE của mô hình ARIMA (1,0,1) 54](#_Toc165717220)

[Figure 89 Các giá trị RMSE, MAPE, MSE của mô hình ARIMA (2,0,1) 54](#_Toc165717221)

1. Câu 1
   1. Yêu cầu:
      * + Áp dụng mô hình ARIMA và một mô hình máy học tự chọn vào bộ dữ liệu Gas & Electric.
   2. Bài giải:
      1. Cơ sở lý thuyết:

Dự báo chuỗi thời gian là một lớp mô hình quan trọng trong thống kê, kinh tế lượng và machine learning. Ta gọi lớp mô hình này là chuỗi thời gian (time series) là vì mô hình được áp dụng trên các chuỗi đặc thù có yếu tố 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.

* **Mô hình ARIMA**: Dựa trên giả thuyết chuỗi dừng và phương sai sai số không đổi. Mô hình sử dụng đầu vào chính là những tín hiệu quá khứ của chuỗi được dự báo để dự báo nó. Các tín hiệu đó bao gồm: chuỗi tự hồi qui AR (auto regression) và chuỗi trung bình trượt MA (moving average). Hầu hết các chuỗi thời gian sẽ có xu hướng tăng hoặc giảm theo thời gian, do đó yếu tố chuỗi dừng thường không đạt được. Trong trường hợp chuỗi không dừng thì ta sẽ cần biến đổi sang chuỗi dừng bằng sai phân. Khi đó tham số đặc trưng của mô hình sẽ có thêm thành phần bậc của sai phân d và mô hình được đặc tả bởi 3 tham số ARIMA(p, d, q).
* **Mô hình SARIMA**: Về bản chất đây là mô hình ARIMA nhưng được điều chỉnh đặc biệt để áp dụng cho những chuỗi thời gian có yếu tố mùa vụ. Như chúng ta đã biết về bản chất ARIMA chính là mô hình hồi qui tuyến tính nhưng mối quan hệ tuyến tính thường không giải thích tốt chuỗi trong trường hợp chuỗi xuất hiện yếu tố mùa vụ. Chính vì thế, bằng cách tìm ra chu kì của qui luật mùa vụ và loại bỏ nó khỏi chuỗi ta sẽ dễ dàng hồi qui mô hình theo phương pháp ARIMA.
* **Mô hình SES:** Simple exponential smoothing (SES) là một phương pháp dự đoán dữ liệu theo thời gian. Nó sử dụng tổng có trọng lượng của dữ liệu trước trong dãy dữ liệu theo thời gian để đoán các dữ liệu trong tương lai.

Bộ dữ liệu: Bộ dữ liệu bao gồm hai ba cột (Month, Gas use, Electric use) cho ta biết lượng sử dụng của ga và điện trong thời gian hai năm theo từng tháng. Để có thể dự báo theo từng tháng, ta sẽ phải tách hai cột gas use and electric use ra thành hai và áp dụng mô hình theo từng cột.

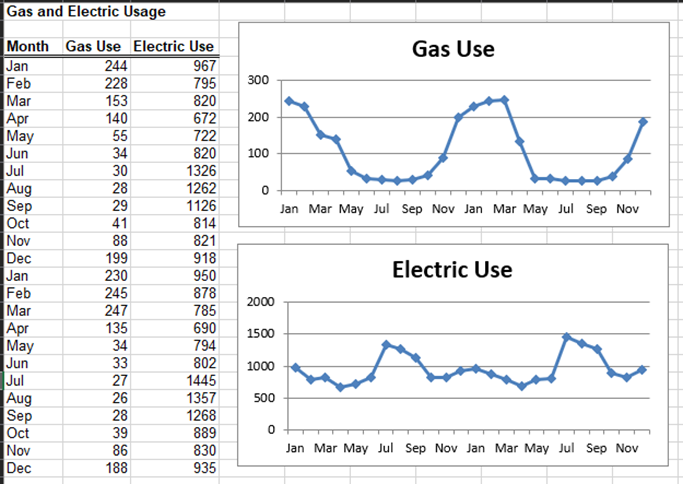


Figure 1. Dataset

* + 1. Excel:
       - ARIMA:

Kiểm tra một chuỗi thời gian có tính dừng là một chuỗi các giá trị mean, variance, autocorrelation không thay đổi theo thời gian và nó không bao hàm yếu tố xu thế.

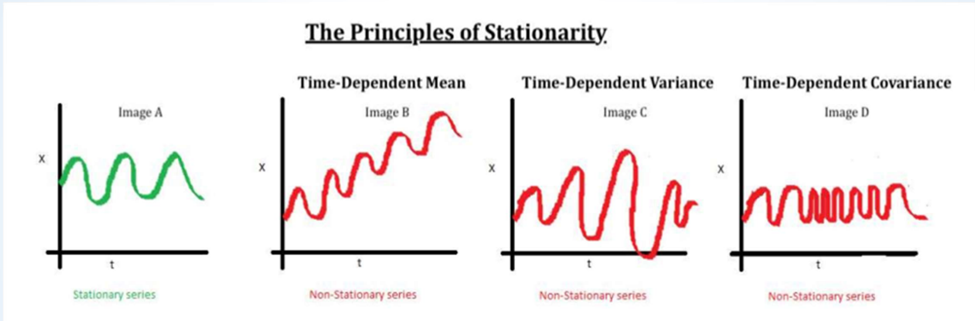


Figure 2. Nhận diện chuỗi dừng đồ thị

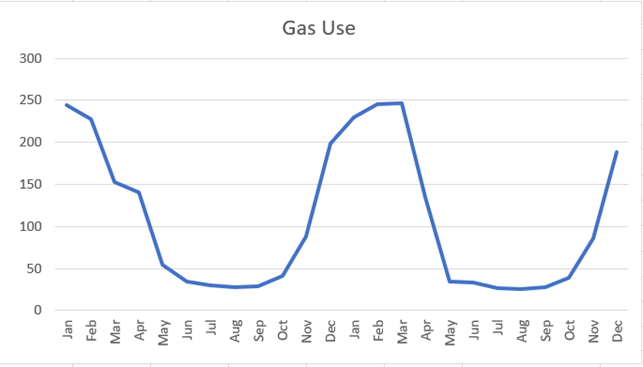


Figure 3. Đồ thị Gas Use (Excel)

Dựa vào đồ thị được phát thảo trong Excel, ta thấy dataset chưa có dấu hiệu của chuỗi dừng, vì vậy ta cần thực hiện phương pháp sai phân, so sánh giữa giá trị hiện tại của nó và giá trị hiện tại. Ngoài ra khi thực hiện việc so sánh sai phân sẽ cho ta giá trị d trong bộ ba giá trị ARIMA. (Số lần sai phân bằng d)

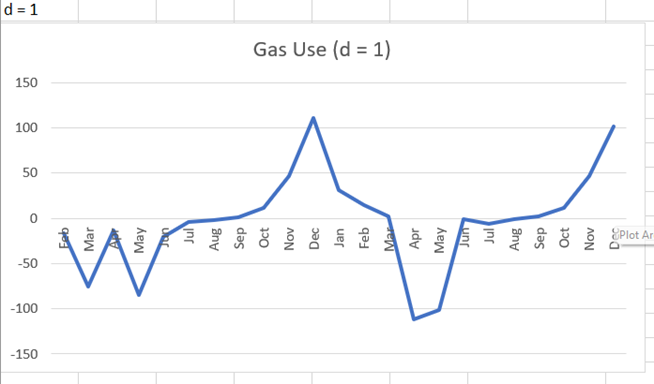


Figure 4. Đồ thị Gas Use (d =1)

Áp dụng sai phân lần 1 (d = 1) có dấu hiệu chuỗi dừng, để chắc chắn ta ta sẽ thực hiện một lần nữa.

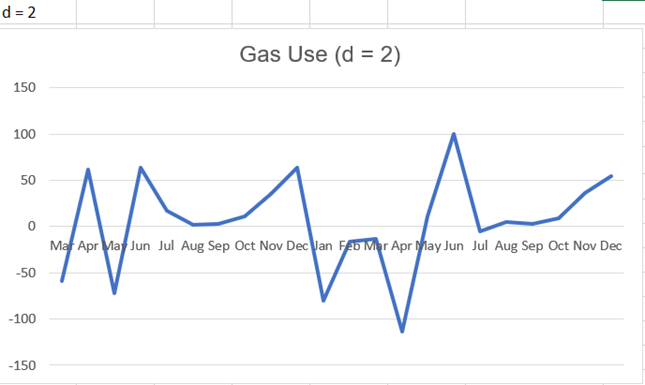


Figure 5. Đồ thị Gas Use (d=2)

Với lần hai (d = 2) ta thấy có dấu hiệu chuỗi dừng, vì vậy ta sẽ chấp nhận giá trị d = 2.

Để có thể xác định được hai giá trị còn lại là p and q, ta sẽ kiếm thông qua hàm tự tương quan ACF (Autocorrelation function) và hàm tự tương quan từng phần PACF (Partial autocorrelation function) với biểu đồ của chúng.

Tính ACF(k) tại độ trễ k =

Với:

* yt : chuỗi thời gian dừng tại thời điểm t.
* yt+k : chuỗi thời gian dừng tại thời điểm t + k.
* µ : giá trị trung bình của chuỗi dừng.

Tùy thuộc vào từng trường hợp, số lượng “lag” (độ trễ) sẽ do người dùng tự quyết định, tại đây ta sẽ áp dụng cho 4 lag để xem biểu đồ. Tính độ trễ của từng lag.



Figure 6. Độ trễ tưng lag

Sau đó tổng lại toàn bộ của từng lag, ta sẽ có như hình sau:

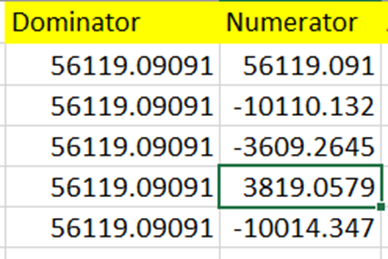


Figure 7. Tổng kết giá trị tưng phần

Với vế bên phải đại diện cho tử số của công thức, còn vế trái đại diện cho mẫu số công thức, ta sử dụng hàm DEVSQ để trên cột giá trị dataset ban đầu.

Sau khi ta chia và lấy giá trị ACF của từng lag:

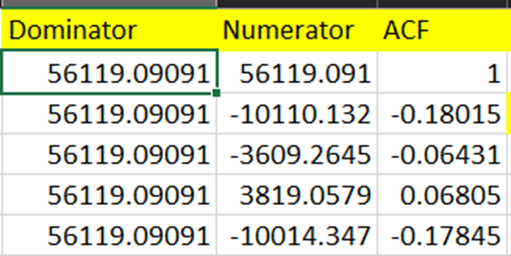


Figure 8. Kết quá ACF của từng lag

Ta nhận thấy, lag đầu tiên cho giá trị là một và các lag sau dần giảm xuống âm, phát họa lên đồ thị, ta có:

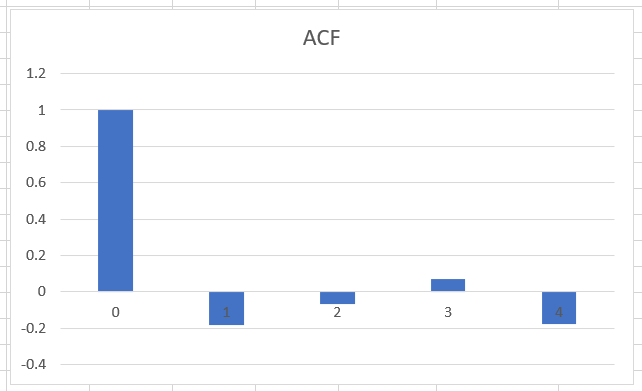


Figure 9. Đồ thị ACF Gas Use (Excel)

Tiếp theo, ta cần tính Hàm tự tương quan từng phần PACF bằng cách giải ma trận sau với mỗi k:

Ma trận giữa hàm tự tương quan ACF và hàm tự tương quan từng phần PACF được biết đến là thuật toán đệ quy Durbin – Levinson, dùng để tính toán chỉ số PACF từ ACF có được.

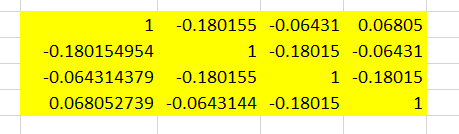


Figure 10. Ma trận

Sử dụng kết hợp hàm MMULT và MINVERSE trong thanh công thức:

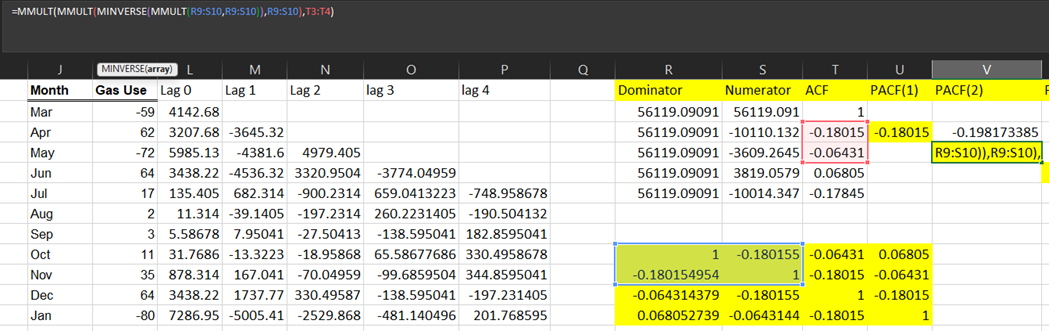


Figure 11. Tính PACF từng lag

Sau khi tính ra được, các giá trị trên đường chéo chính chính là giá trị PACF tương ứng với mỗi lag. Phác thảo đồ thị PACF ta có:



Figure 12. Đồ thị Gas Use PACF (Excel)

Dựa vào đồ thị của ACF có PACF ta có thể thấy:

**Đối với thành phần AR (p):**

* **ACF:**
  + ACF tại lag 1 là -0.180154954, chỉ ra mối tương quan âm mạnh giữa chuỗi tại thời điểm hiện tại và giá trị ngay trước đó.
  + ACF tại lag 2 là -0.064314379, gần với 0 và cho thấy mối tương quan yếu hơn so với lag 1.
  + ACF tại lag 3 và 4 là dương nhưng tương đối nhỏ, chỉ ra mối tương quan yếu hơn.
  + ACF tại lag 5 là âm, nhưng gần với 0, cho thấy mối tương quan nhỏ.
* **PACF:**
  + PACF tại lag 1 là -0.100016296, thể hiện chỉ số lơn đáng kể.
  + PACF tại lag 2 là 0.038929167, chỉ ra nhỏ hơn.
  + PACF tại lag 3 là -0.171630592, lại thể hiện chỉ số đáng kể.
* **Kết luận**:
  + Giá trị ACF âm tại lag 1 và chỉ số đáng kể tại lag 1 trong PACF cho thấy một mối tương quan mạnh và quan hệ trực tiếp đáng kể giữa chuỗi tại thời điểm hiện tại và giá trị tại thời điểm trước đó, có bằng chứng cho ta chọn p = 1.
  + Các giá trị ACF tiếp theo trở nên nhỏ hơn và các giá trị PACF giảm đều, cho thấy sự giảm dần của các mối tương quan và quan hệ trực tiếp cho các lag cao hơn, ủng hộ việc chọn p=1.

**Đối với thành phần MA (q):**

* **ACF:**
  + ACF tại lag 1 là âm, nhưng tương đối nhỏ.
  + Các giá trị ACF cho tất cả các lag đều tương đối nhỏ và trở nên không đáng kể sau lag 1.
* **PACF:**
  + Các giá trị PACF cho thấy các đỉnh đáng kể tại lag 1 và sau đó dần giảm, trở nên không đáng kể cho các lag cao hơn.
* **Kết luận:**
  + Cả ACF và PACF gợi ý rằng chuỗi không có tương quan tự tự đáng kể ở các lag cao hơn lag 1. Các giá trị ACF tương đối nhỏ và trở nên không đáng kể sau lag 1, và các giá trị PACF giảm dần và trở nên không đáng kể cho các lag cao hơn. Có thể thấy có thể không có thành phần Moving Average đáng kể ở các lag cao hơn lag 1 (q=0).

Sau khi hoàn thành ước lượng bộ ba biến p, d và q ta có ARIMA(1,2,0) và xây dựng mô hình:

trong đó:

* y(t): quan sát dừng tại thời điểm t.
* e(t): sai số dựa báo quá khứ.
* a0, a1 : các hệ số phân tích hồi quy.

Chia tập dữ liệu ra thành train và test theo tỉ lệ 80/20, ta có cần dự đoán giá trị của năm thàng cuối trong 24 tháng của bộ dữ liệu và so sánh với giá trị ban đầu.

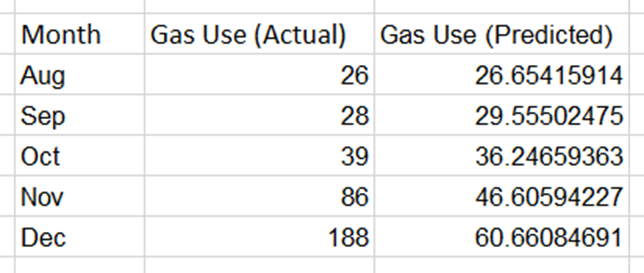


Figure 13. So sánh giữa giá trị thực thế và dự báo (Gas Use)

Phác thảo lên đồ thị

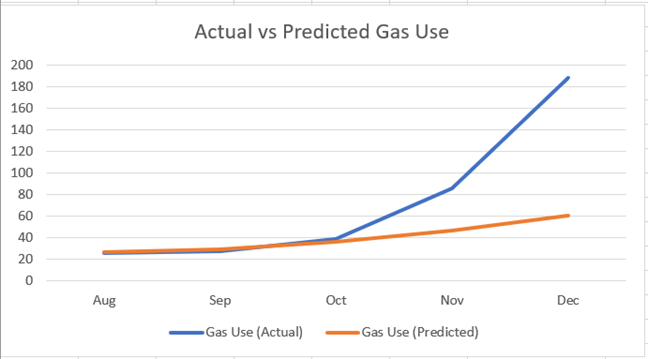


Figure 14. Đồ thị so sánh giữa thực tế và dự báo (Gas Use)

Nhận xét: model ARIMA(1,2,0) cho kết quả dự đoán sát với giá trị thực tế trong ba tháng đầu nhưng có dấu hiệu tăng chậm so với hai tháng kế tiếp.

Đối với các cột Electric Use, áp dụng tương tự các bước trên ta có:

Kiểm tra tính dừng:

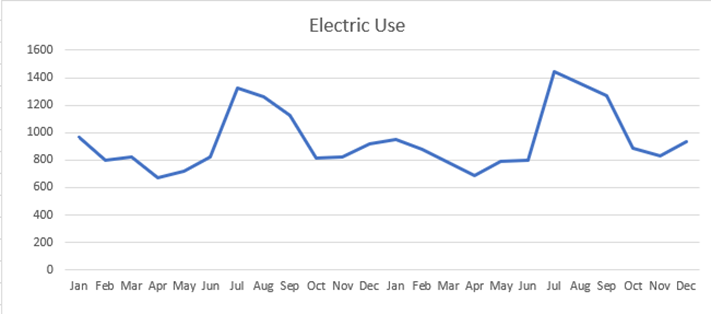


Figure 15. Đồ thị Electric Use (Excel)

Khác với bên Gas Use, cột Electric Use đã có dấu hiệu dừng nên không cần khai triển sai phân để biến chuỗi thành chuỗi dừng (d = 0).

Tính trị giá ACF và PACF với các độ trễ lag, ở đây ta cũng sử dụng 4 lag:

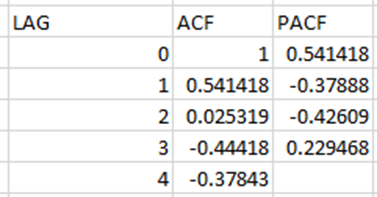


Figure 16.Giá trị Electric Use của ACF và PACF

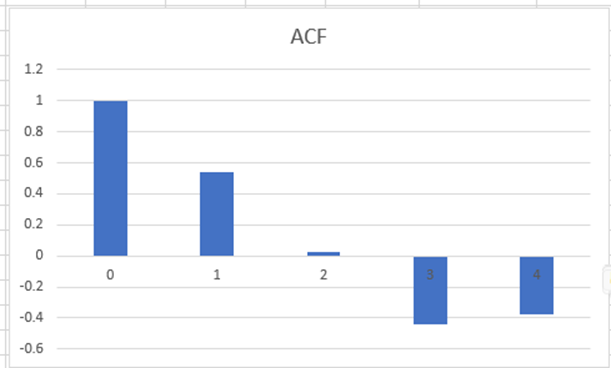


Figure 17. Đồ thị Electric Use ACF (Excel)

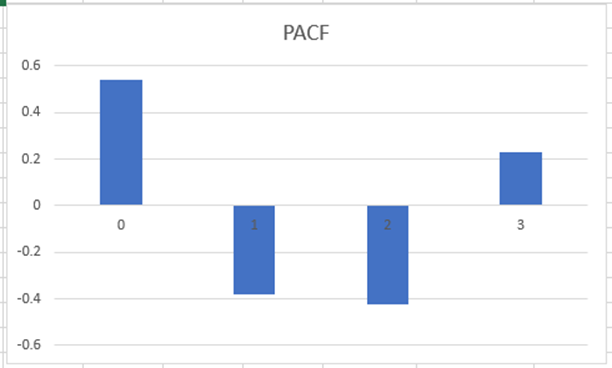


Figure 18. Đồ thị Electric Use PACF (Excel)

Dựa vào đồ thị của ACF có PACF ta có thể thấy:

* **ACF:**
  + ACF tại lag 1 khoảng 0.541418, và giảm dần khi tăng lag.
  + Các giá trị ACF tại lag 2 và 3 tương đối nhỏ (0.025319 và -0.44418), và không có tương quan tự đáng kể vượt quá lag 3.
* **PACF:**
  + PACF tại lag 1 là -0.37888, thể hiện chỉ số âm không đáng kể.
  + PACF tại lag 2 là -0.42609, có giá trị nhỏ hơn cả lag 1.
  + PACF tại lag 3 là 0.244968, lại thể hiện chỉ số dương, tăng đột ngột so với ta hai lag trước đó.

Theo đó ta có thể rút ra:

* **Thành phần AR (*p*)**: Chúng ta có thể xem xét việc bao gồm lag 1 và có thể lag 3 trong thành phần AR, vì chúng có các giá trị tương quan tự khá cao và tương quan tự bộ phận đáng kể. Do đó, *p* có thể là 1 hoặc 3.
* **Thành phần MA (*q***): Các giá trị ACF tại lag 2 và 3 tương đối nhỏ, chỉ ra khả năng tương quan với các lag đó. Tuy nhiên, vì các giá trị PACF tại lag 2 và 3 không quá đáng kể, có thể không cần bao gồm chúng trong thành phần MA. Do đó, *q* có thể là 0 hoặc 1.

Từ các kết luận trên ta sẽ chọn thử bộ ba số p, d và q là 1, 0 và 1 cho mô hình ARIMA để dự đoán kết quả.

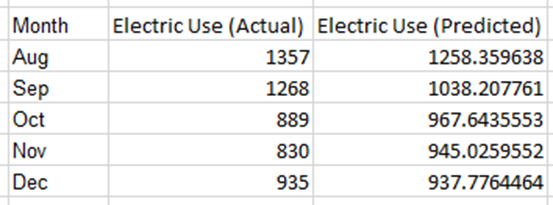


Figure 19. So sánh giữa giá trị thực tế và dự đoán Electric Use

Phác thảo lên đồ thị:

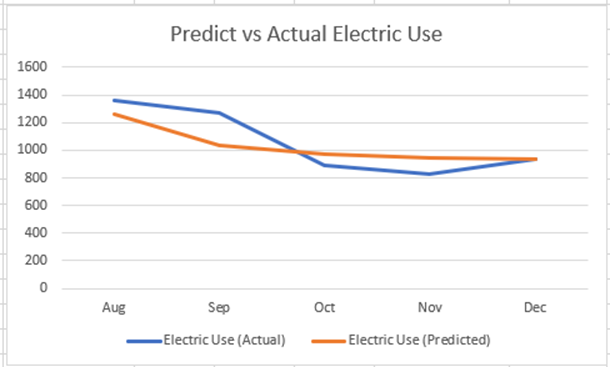


Figure 20. Đồ thị so sánh giữa giá trị thực tế và dự đoán Electric Use (Excel)

**Nhận xét**: Mô hình dự đoán lượng điện sử dụng hơn giá trị thực tron hai tháng đầu và dưới mức giá trị thực trong ba tháng sau, cho thấy bộ số (1,0,1) chưa cho kết quả phù hợp và cần chọn lại giá trị của p và q.

* + - * SES

Trong phương pháp smoothing đơn giản (SES), hệ số san bằng mũ, thường được ký hiệu là α (alpha), là một tham số quyết định độ quan trọng của quan sát gần nhất khi dự báo các giá trị trong tương lai. Đối với mô hình SES ta chỉ cần chọn hệ số α để đưa vào mô hình và dự đoán.

Trong đó:

* Ft+1: Giá trị dự đoán tại thời điểm t + 1.
* α: Hệ số san bằng hàm số mũ,0 ≤ α ≤ 1.
* Yt: Giá trị thực tại thời điểm t.
* Ft: Giá trị dự đoán tại thời điểm t.

Để có thể xem hệ số san bằng mũ nào có cho giá trị gần với giá trị thực nhất, ta sẽ thử với các giá α bằng 0.1, 0.2, 0.3 và 0.5.

Đối với cột Gas Use:

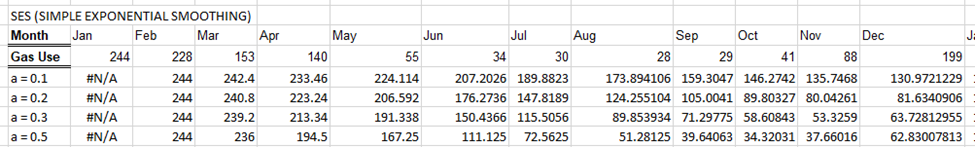


Figure 21. So sánh Gas Use các dự báo của các hệ số a

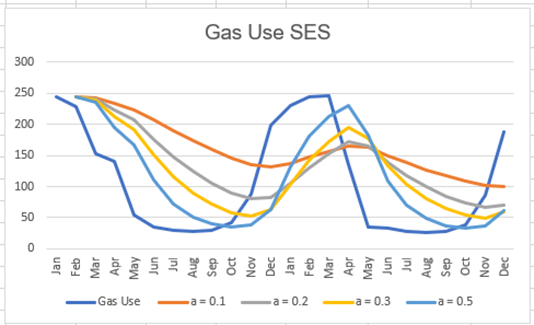


Figure 22. Đồ thị Gas Use so sánh các hệ số a

Đối với cột Electric Use:

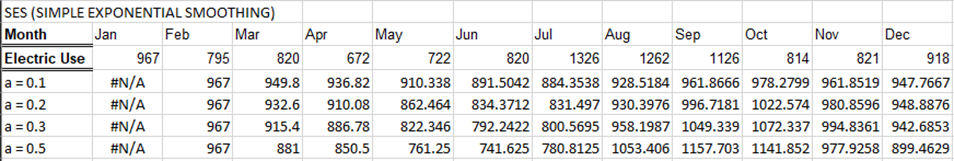


Figure 23. So sánh Electric Use các dự báo của các hệ số a

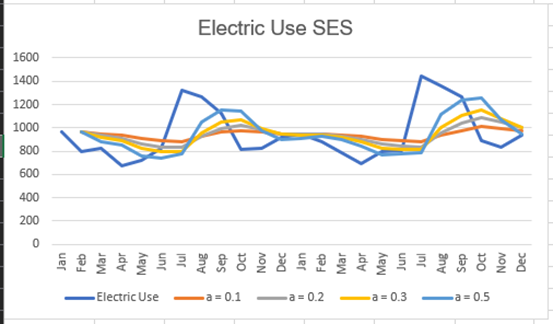


Figure 24. Đồ thị Electric Use so sánh các hệ số a

**Nhận xét**: Ở cả hai cột dự báo, α = 0.5 cho giá trị dự báo gần với giá trị thật nhất. Ta rút ra được rằng trong kết quả của dự báo đã tính tới tác động của tất cả các giai đoạn trước đó và trọng số của giai đoạn càng gần càng lớn theo cấp số nhân. Hệ số α trong mô hình cần lớn thì mô hình càng nhạy bén với sự biến động của số liệu và ngược lại.

* + 1. Python:
       - ARIMA:

Trong Python, chúng ta sẽ sử dụng thư viện statsmodels để thực hiện phần lớn công việc dự đoán của chúng ta.

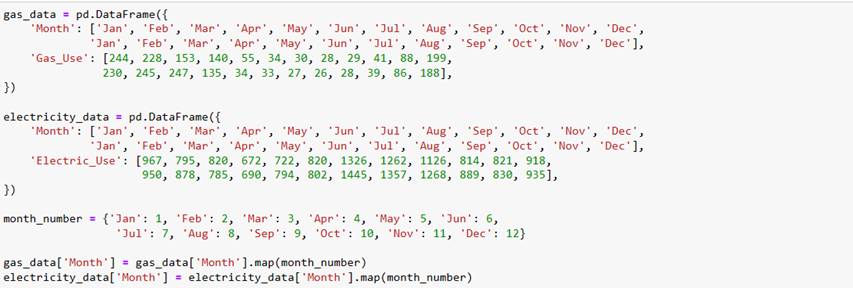


Figure 25. Tạo Data frame đối với mỗi cột

Tương tự, ta cần kiểm tra tính dừng của chuỗi gas\_data.

Ngoài phác họa đồ thị ra, ta có thể sử dụng phương pháp kiểm định Augmented Dickey-Fuller (ADF) để có thể biết được rằng chuỗi có tính dừng không. Trong kiểm định ADF, giả thuyết H0 cho rằng chuỗi có Unit Root và là chuỗi dừng, đối thuyết H1 cho rằng ngược lại. Kết quả sẽ tùy vào giá trị p-value có dưới mức chấp nhận (thường là 0.05) hay không để chấp nhận hay bác bỏ giả thuyết.

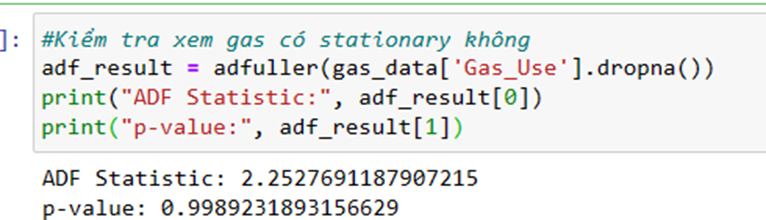


Figure 26. Kiểm định ADF cho gas\_data

Như ta thấy giá trị p-value cao hơn mức chấp nhận 0.05 giống với đồ thị bên Excel, chính vì vậy ta sẽ áp dụng sai phân hai lần và kiểm định ADF mỗi lần để kiểm tra giá trị.

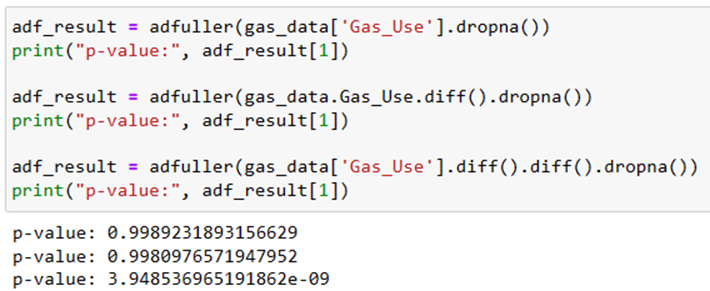


Figure 27. Kết quả ADF sau sai phân

Sau hai lần sai phân, ta có thể thấy p-value đã giảm xuống dưới mức chấp nhận, đúng với đồ thị bên Excel nên ta sẽ chấp nhận hệ số d = 2.

Kế tiếp, cần tính toán giá trị p và q thông qua đồ thị ACF và PACF.

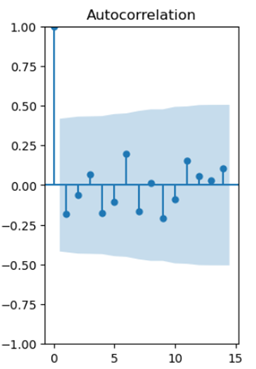


Figure 28. Đồ thị Gas Use ACF (Python)



Figure 29. Đồ thị Gas Use PACF (Python)

Từ các đồ thị trên ta có được bộ ba (1,2,0) ARIMA.

Chia bộ dữ liệu thành tập train và test theo tỉ lệ 80/20, huấn luyện mô hình với chỉ số ta có:

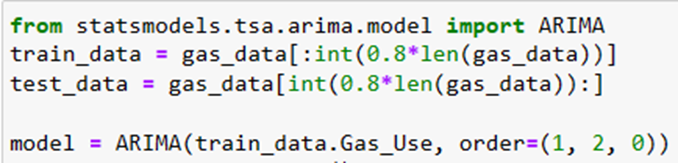


Figure 30. Chia và huấn luyện mô hình ARIMA

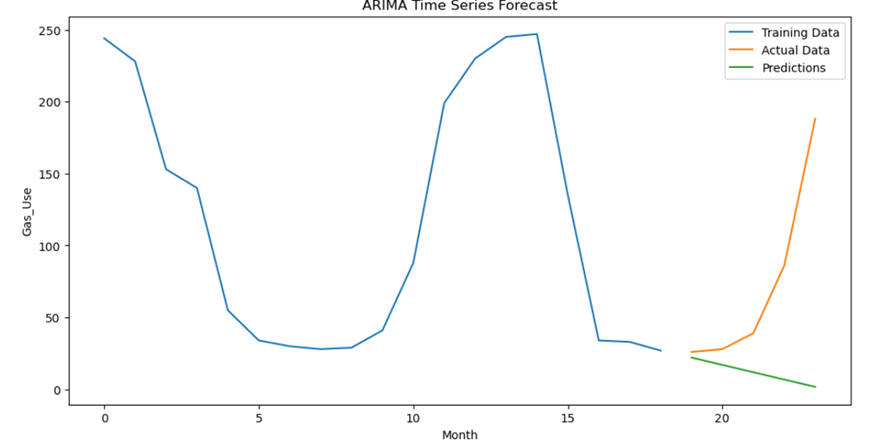


Figure 31. Đồ thị Gas Use kết quả dự đoán ARIMA

**Nhận xét**: có sai lệch nghiêm trọng trong giá trị dự báo của mô hình so với Excel, do có thể bộ dữ liệu có tính đặt trưng tiềm ẩn và sử dụng sai model.

Theo như quan sát đò thị ta có thể dấu hiệu đặc trưng của chuỗi theo mùa (lượng ga sử dụng nhiều vào các tháng đầu năm, giảm vào giữa năm và lại tăng cao vào cuối năm). Chính vì vậy ta sẽ sử dụng mô hình SARIMA (Seasonal ARIMA), là mô hình ARIMA nhưng được điều chỉnh đặc biệt để áp dụng cho những chuỗi thời gian có yếu tố mùa vụ.



Figure 32. Huấn luyện mô hình SARIMA

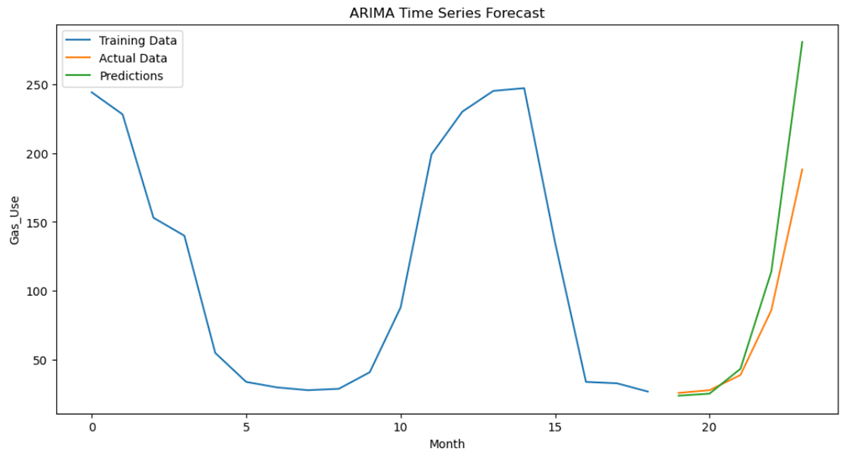


Figure 33. Đồ thị Gas Use kết quả dự đoán mô hình SARIMA

Sử dụng mô hình SARIMA cho ta kết quả dự đoán gần hơn với giá trị thực so với mô hình ARIMA bình thường. Nhưng theo quan sát kết quả vẩn còn lệch khá với tháng sau, cần có điều chỉnh về hệ số p và q cho mô hình khớp hơn.

Áp dụng tương tự vào cột Electric Use ta có:

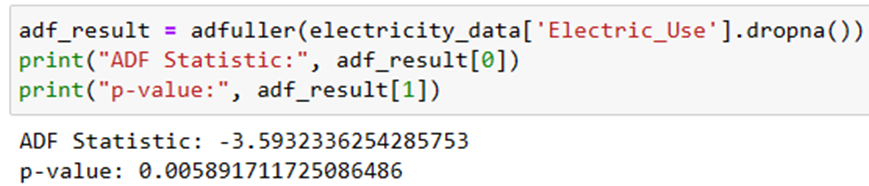


Figure 34. Kiểm định ADF cho electricity\_data

Giá trị p-value của electricity\_data không nằm dưới mức chấp nhận nhưng ta vẫn có thể chấp nhận và không thực hiện sai phân để tránh trường hợp sai phân quá mức (over-differencing).

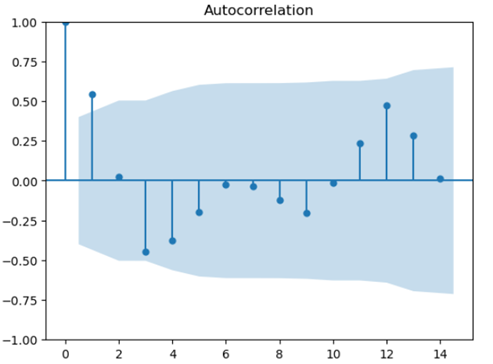


Figure 35. Đồ thị Electric Use ACF (Python)

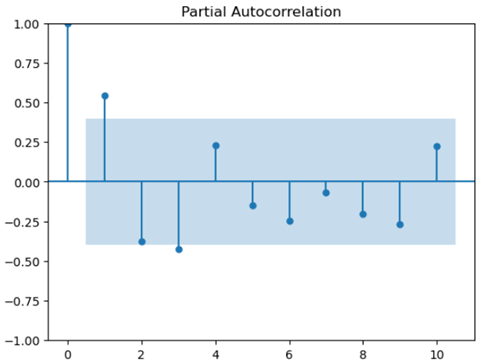


Figure 36. Đồ thị Electric Use PACF (Python)

Dựa vào đồ thị ACF và PACF, ta có được giá trị p và q là 1 và 1, từ đó có được bộ số mô hình ARIMA(1,0,1).

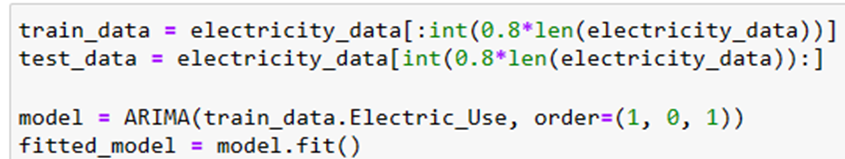


Figure 37. Chia và huấn luyệ mô hình ARIMA

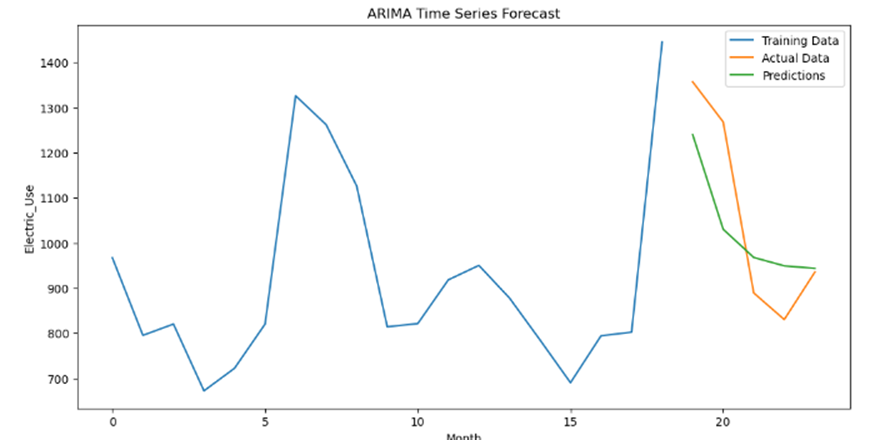


Figure 38. Đồ thị Electric Use so sánh giữa giá trị thực tế và dự báo (Python)

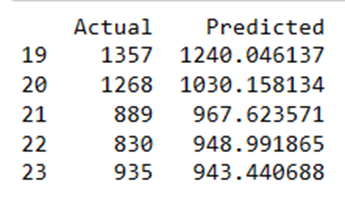


Figure 39. So sánh giữa giá trị dụ đoán và thực tế

**Nhận xét**: khác với cột Gas\_Use, mô hình cho kết quả dự đoán gần giống với kết quả dụ đoán của Excel.

* + - * SES

Để có thể sử dụng mô hình SES trong Python, ta cần phải khởi tạo mô hình riêng cho nó:



Figure 40. Mô hình SES

Do trước đó ta thấy α = 5 cho giá trị tốt nhất, ta cũng sẽ lấy α = 5 làm hệ số để tính:

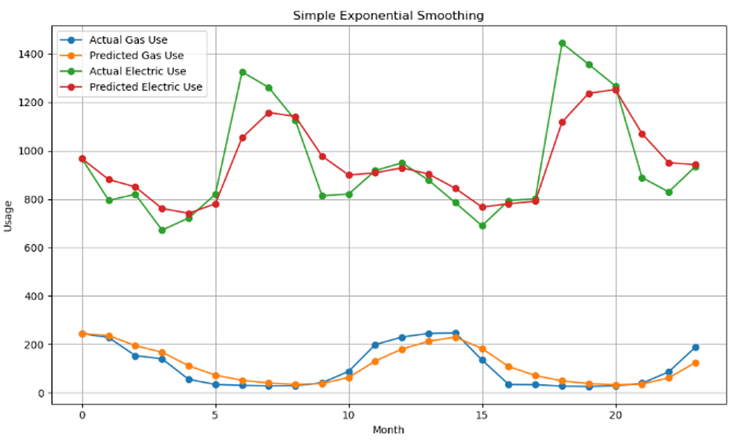


Figure 41. Đồ thị SES của Electric Use và Gas Use

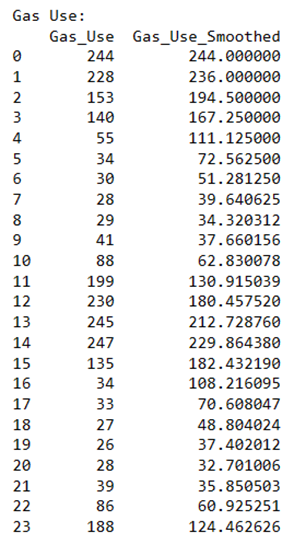


Figure 42. So sánh SES Gas Use

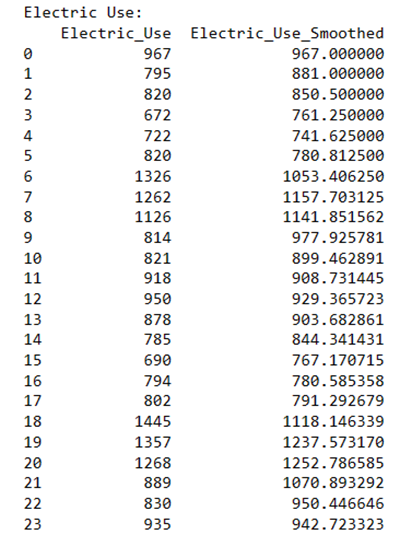


Figure 43. So sánh SES Electric Use

**Nhận xét**: Giá trị kết quả dự đoán gần đúng với giá trị thực và khớp với Excel.

1. Câu 2
   1. Yêu cầu:
      * + Áp dụng mô hình ARIMA và một mô hình máy học tự chọn vào bộ dữ liệu tùy chọn của Việt Nam.
   2. Bài giải:
      1. ARIMA

* Mô hình ARIMA (Autoregressive Integrated Moving Average) là một phương pháp thống kê phổ biến được sử dụng để dự báo chuỗi thời gian
* Mô hình ARIMA được ký hiệu là ARIMA(p, d, q), với p,d,q không âm
* Trong đó:
* p: Số hạng tự hồi quy – AR(p): Biểu thị số lượng giá trị quá khứ của chuỗi thời gian ảnh hưởng đến giá trị hiện tại.
* d: Số lần sai phân -I(d): Chỉ ra số lần chuỗi thời gian cần được sai phân để đạt được tính dừng (stationary).
* q: Số hạng trung bình trượt – MA(q): Biểu thị số lượng lỗi dự báo quá khứ được sử dụng để điều chỉnh dự báo hiện tại.
* Bộ dữ liệu: các chỉ số chất lượng không khí tại Hà Nội từ 1/1/2014 – 25/3/2024

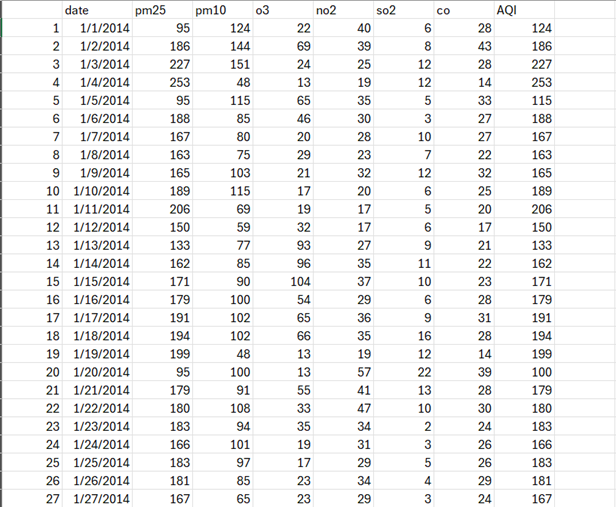


Figure 44 Dữ liệu về các chỉ số không khí tại Hà Nội từ 1/1/2014 - 25/3/2023

* Dùng ARIMA để dự đoán cột AQI
* **Excel:**
* Kiểm định tính dừng của chuỗi: Một chuỗi thời gian có tính dừng là một chuỗi các giá trị mean, variance, autocorrelation không thay đổi theo thời gian và không bao hàm yếu tố xu thế.
* Dùng Dickey-Fuller test

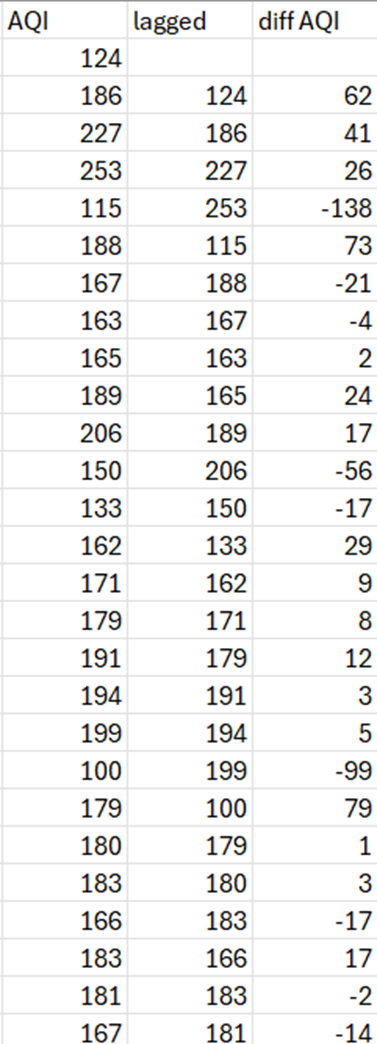


Figure 45 AQI,lagged, diff AQI

* Thực hiện hồi qui tuyến tính với AQI là biến phụ thuộc, lagged và diff AQI là biến độc lập.

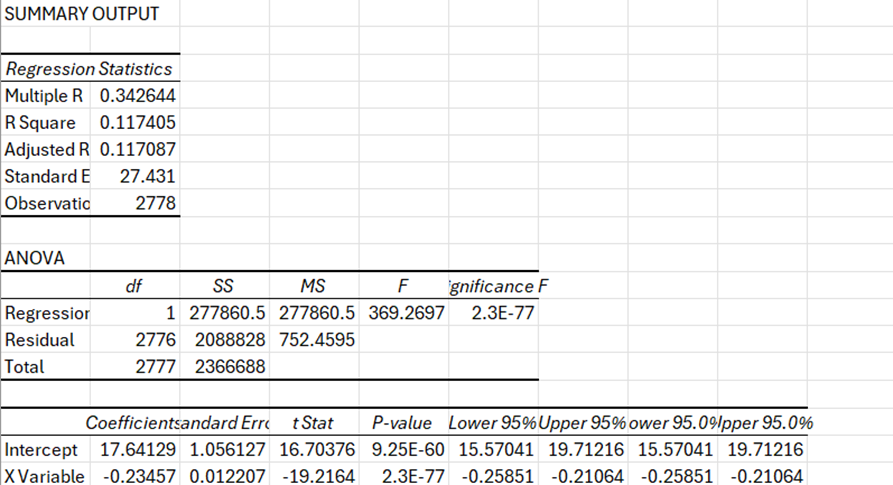


Figure 46 Kết quả hồi qui tuyến tính

* Ta được: (tra bảng Dickey-Fuller với loại biểu đồ liên tục và không có xu hướng)
* Suy ra chuỗi đã cho là chuỗi dừng. Từ đó ta được d=0 (Integrated).
* Xác định q trong MA(q) thông qua ACF.
* Hàm tự tương quan (ACF) là một kỹ thuật thống kê dùng để xác định mức độ tương quan của các giá trị trong một chuỗi thời gian với nhau. ACF thể hiện sự tương quan giữa chuỗi và độ trễ của chính nó.Hệ số tương quan có thể dao động từ -1 (mối quan hệ tiêu cực hoàn hảo) đến +1 (mối quan hệ tích cực hoàn hảo). Hệ số bằng 0 có nghĩa là không có mối quan hệ giữa các biến.
* Tính ACF(k) tại độ trễ k
* Với:
* : giá trị tại thời điểm t.
* : giá trị tại thời điểm t - k.
* µ : giá trị trung bình của chuỗi dừng.
* Ta thực hiện tính ACF cho 10 lags

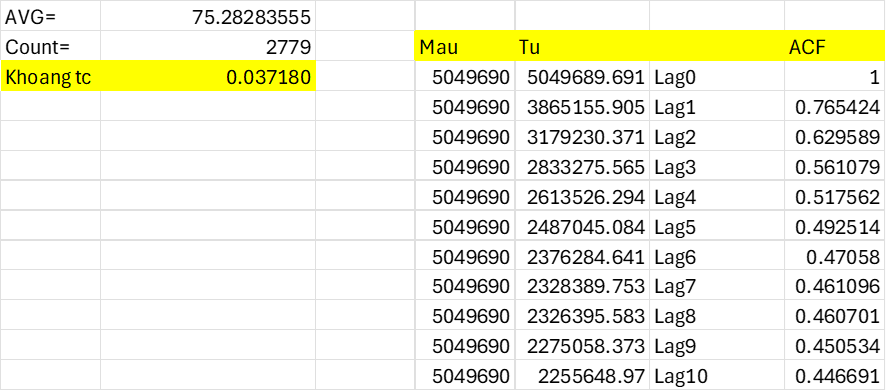


Figure 47 ACF tại các lag

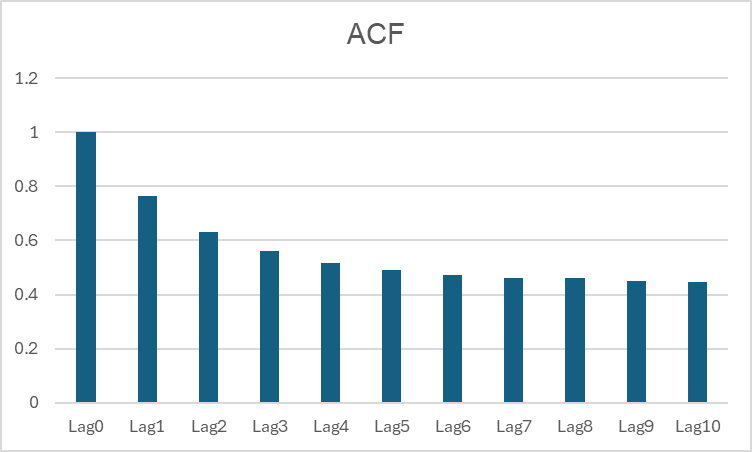


Figure 48 Biểu đồ ACF

* Nhận xét: giá trị ACF tại các lag đều lớn khoảng tin cậy
* chọn: q = 1 (lag có ACF lớn nhất)
* Xác định p trong AR(p) thông qua PACF.
* Hàm tự tương quan một phần (PACF) đo lường mối tương quan từng phần giữa chuỗi thời gian dừng và các giá trị trong quá khứ của chính nó, xem xét và tính toán các giá trị ở tất cả các độ trễ ngắn hơn. Điều này khác với hàm tự tương quan (ACF), hàm này không chịu ảnh hưởng của các độ trễ khác.
* Tính hàm tự tương quan một phần PACF bằng cách giải ma trận sau với mỗi k:
* Ma trận giữa hàm tự tương quan ACF và hàm tự tương quan từng phần PACF được biết đến là thuật toán đệ quy Durbin – Levinson, dùng để tính toán chỉ số PACF từ ACF có được.

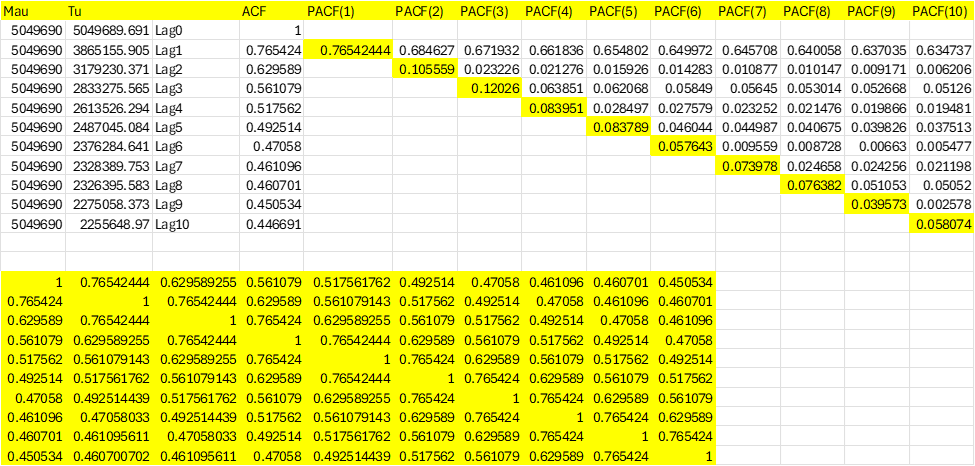


Figure 49 PACF tại các lag

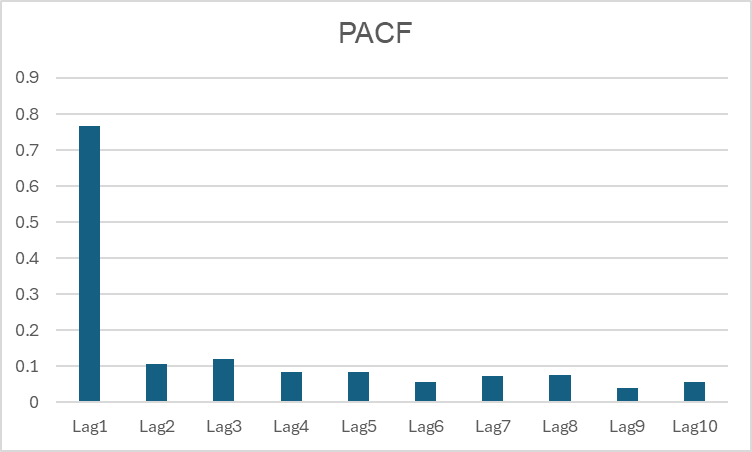


Figure 50 Biểu đồ PACF

* Nhận xét: PACF ở các lag đều lớn hơn độ tin cậy
* Chọn p = 1 hoặc p = 2
* Ta được 2 mô hình: ARIMA(1,0,1) và ARIMA(2,0,1)

Chia tập dữ liệu thành 2 tập train test với tỉ lệ 8:2, sử dụng công cụ real statistics trong excel để dự đoán

A screenshot of a spreadsheet

Description automatically generated

Figure 51 Dự đoán trên tập test của mô hình ARIMA(1,0,1)

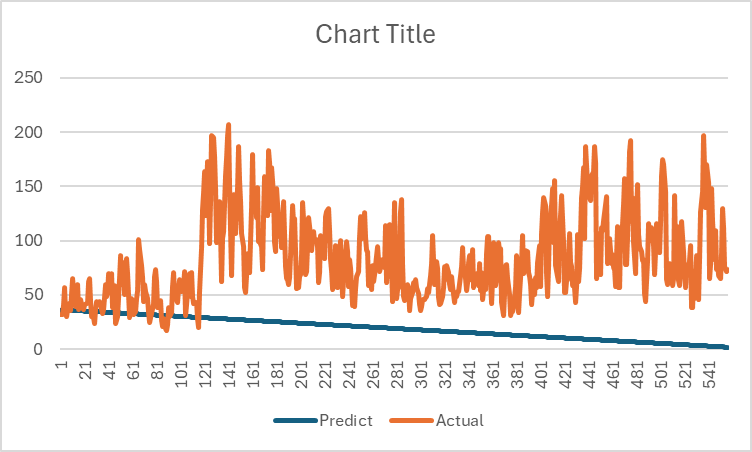


Figure 52 Biểu đồ giữa dữ liệu dự đoán và thực tế của ARIMA(1,0,1)

* ARIMA(2,0,1)

A screenshot of a spreadsheet

Description automatically generated

Figure 53 Dự đoán trên tập test của mô hình ARIMA(2,0,1)

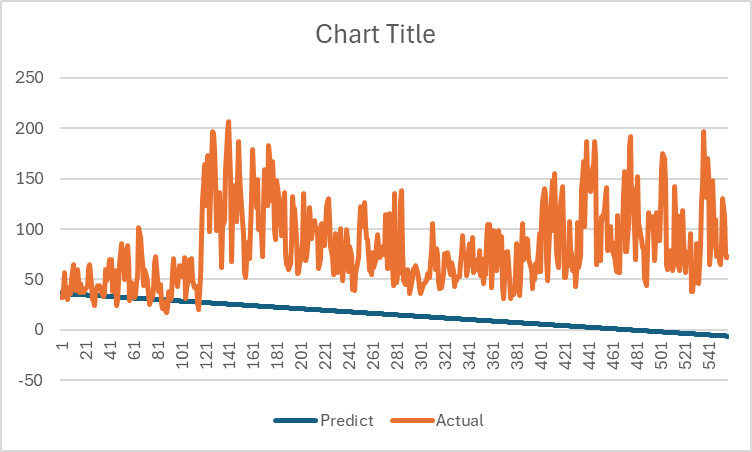


Figure 54 Biểu đồ giữa dữ liệu dự đoán và thực tế của ARIMA(2,0,1)

* Nhận xét:
  + Mô hình ARIMA(1,0,1) có RMSE thấp hơn nên phù hợp hơn.
  + Dữ liệu dự đoán có xu hướng đi xuống, mô hình ARIMA không thích hợp để dự đoán chuỗi thời gian dài.
* **Python:**
* Thực hiện kiểm định tính dừng:

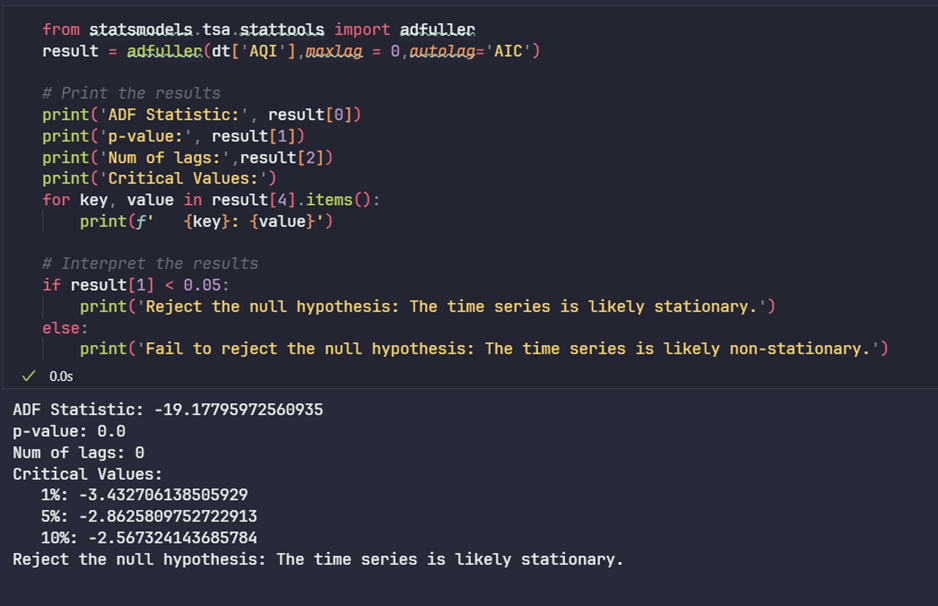


Figure 55 Kiểm định tính dừng bằng python

* Vẽ biểu đồ ACF



Figure 56 Vẽ biểu đồ ACF bằng python

* Nhận xét: tương tự với excel, các ACF đều lớn hơn khoảng tin cậy, chọn q =1 (ACF lớn nhất).
* Vẽ biểu đồ PACF

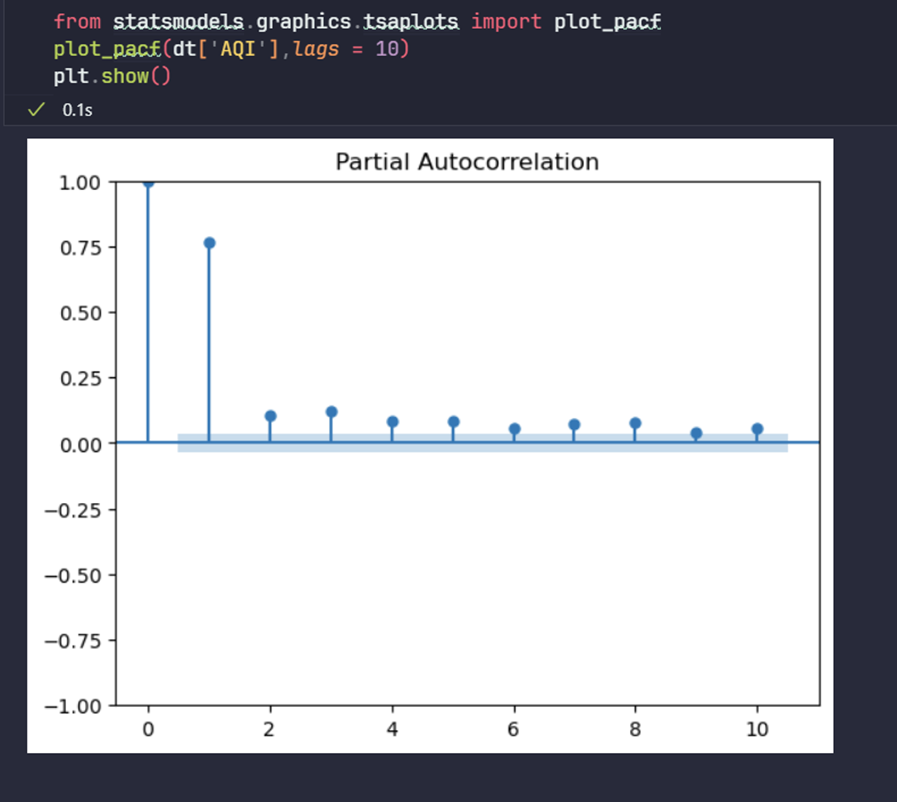


Figure 57. Vẽ biểu đồ PACF bằng python

* Nhận xét: PACF tăng đột biến ở lag 1 sau đó giảm xuống ở lag 2, từ lag 2 trở đi không có sự thay đổi nhièu. Chọn p =1 hoặc p = 2
* Chia tập train và test với tỉ lệ 8:2
* ARIMA(1,0,1)

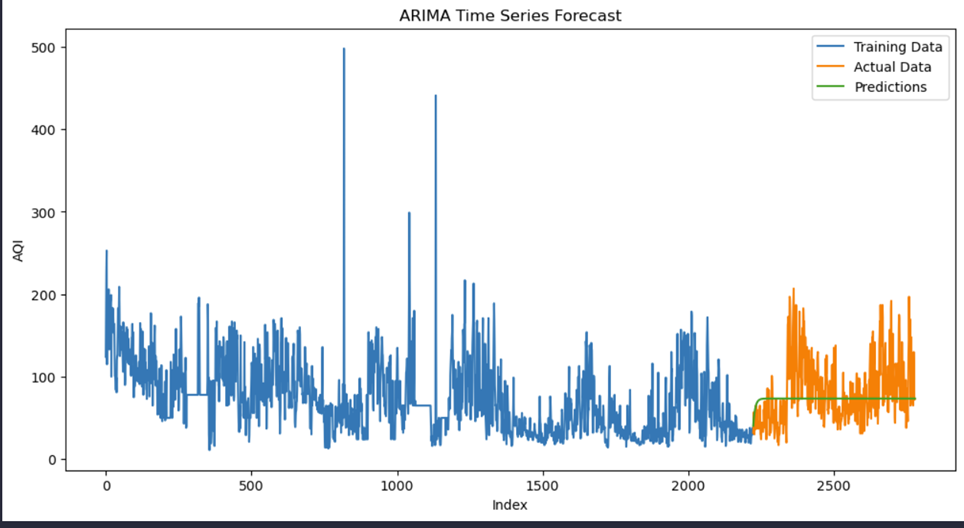


Figure 58 Kết quả dự đoán với ARIMA(1,0,1)

* ARIMA(2,0,1)

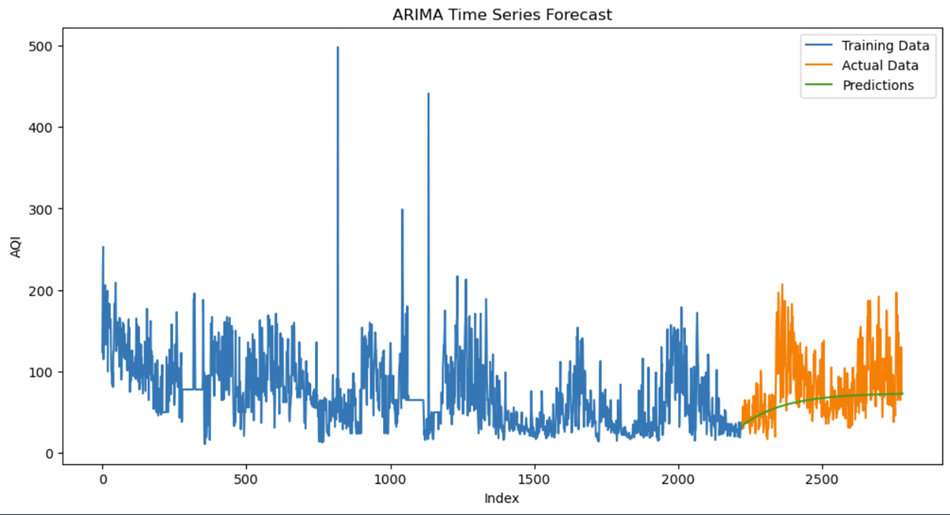


Figure 59. Kết quả dự đoán với ARIMA(2,0,1)

* Từ hai hình trên ta nhận xét : ARIMA(2,0,1) phù hợp hơn
  + 1. SES - Simple exponential smoothing
* SES: Làm mịn theo cấp số nhân là một kỹ thuật được sử dụng để dự báo dữ liệu chuỗi thời gian bằng cách làm mịn các biến động trong dữ liệu. Làm mịn theo cấp số nhân dựa trên giả định rằng các giá trị tương lai của chuỗi thời gian là một hàm của các giá trị trong quá khứ của nó. Phương pháp này hoạt động tốt khi chuỗi thời gian có xu hướng và/hoặc thành phần theo mùa.
* Mô hình:
* Với
* Trong đó:

|  |  |
| --- | --- |
|  | Giá trị dự đoán thứ t+1,t |
|  | Hệ số làm trơn, có già trị từ 0 đến 1 |

* **Excel**
* Thực hiện dự đoán trên bộ dữ liệu với từ 0.1 – 0.9

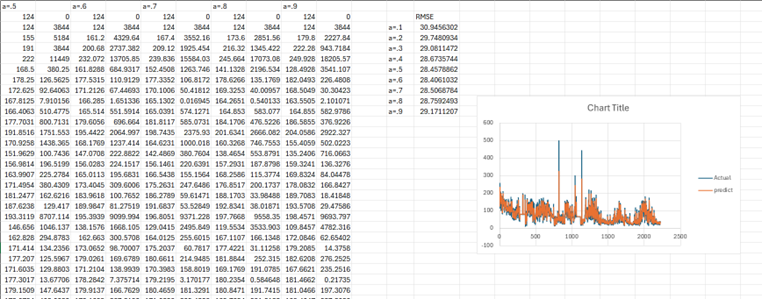


Figure 60 Giá trị dự đoán tại các a khác nhau

* Kết quả cho thấy = .6 có độ phù hợp cao nhất
* **Python**
* Tạo hàm SES

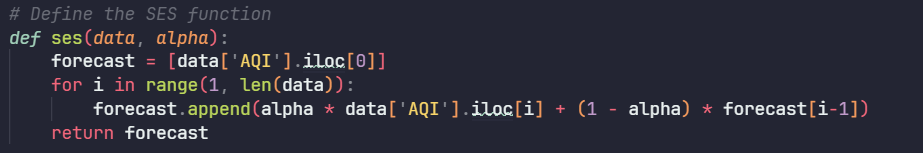


Figure 61 Hàm SES bằng python

A graph showing a number of different colored lines

Description automatically generated with medium confidence

Figure 62. Kết quả dự doán bằng SES

* Nhận xét: kết quả dự đoán gần trùng khớp với thực tế

1. Câu 3
   1. Yêu cầu:
      * + Phương pháp để đánh giá sự chính xác của một thuật toán hay mô hình dự đoán? Áp dụng các phương pháp ấy cho các bài lab 2, 3, 4.1, 4.2.
   2. Bài giải
      1. Các phương thức đánh giá mô hình

Mean Absolute Error (MAE)

* MAE biểu thị mức trung bình của sai số tuyệt đối giữa giá trị thực tế và giá trị dự đoán trong tập dữ liệu, nó đo mức trung bình trong phần dư [1].

Với:

+ là giá trị dự đoán

+ là giá trị dự thực

+ là tổng số điểm dữ liệu

* Độ đo sai được sử dụng để đánh giá giữa mô hình dự đoán và tập dữ liệu dùng để kiểm tra trong bài toán hồi quy. Chỉ số càng nhỏ thì mô hình càng chính xác. Ngoài ra, MSE được biết đến là mạnh mẽ hơn đối với các *yếu tố ngoại lai* so với MAE do MSE được tính bằng bình phương sai số.

Mean Squared Error (MSE)

* MSE trong một phép ước lượng là trung bình bình phương sai số. Cho thấy sai số bình phương trung bình giữa các giá trị dự đoán và thực tế [1].

Với:

+ là giá trị dự đoán

+ là giá trị dự thực

+ là tổng số điểm dữ liệu

* MSE càng thấp thì dự báo càng tốt.

Hệ số xác định hay (R-Squared)

* biểu thị tỷ lệ phần trăm phương sai trong biến phụ thuộc được giải thích bởi *mô hình hồi quy tuyến tính* [2].

Với:

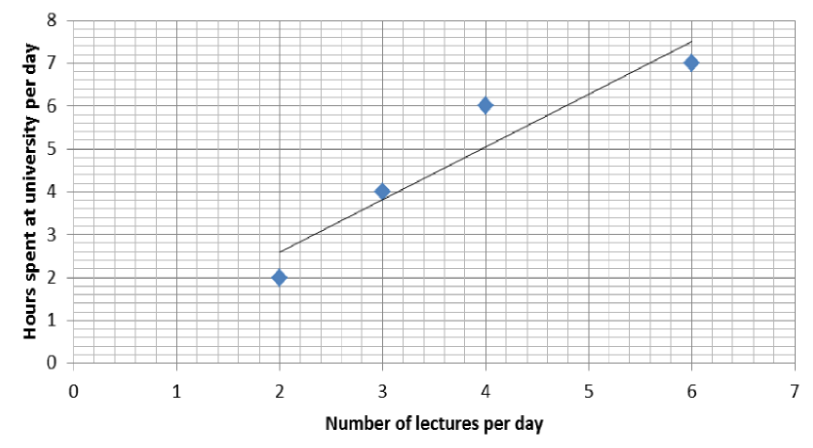
+ là giá trị dự đoán

+ là giá trị dự thực

+ là tổng số điểm dữ liệu

+ là giá trị trung bình

* Ví dụ chúng ta có đồ thị cho biết ảnh hưởng của số tiết học (bài giảng) mỗi ngày và số giờ dành cho việc ở trường đại học. Phương trình hồi quy được biểu thì bằng đường thẳng như trong ảnh. .



Áp dụng công thức tìm được . Điều này cho chúng ta biết 89.5% số giờ ở trường đại học được giải thích bởi số tiết học (bài giảng) trong ngày.

Root mean square error (RMSE):

* là một độ đo thống kê để đo lường mức độ lỗi của một mô hình dự đoán so với dữ liệu thực tế.
* RMSE được tính bằng cách lấy căn bậc hai của trung bình của bình phương của sự sai lệch giữa dự đoán và giá trị thực tế:
* Với:
  + : giá trị thực tế
  + : giá trị dự đoán
  + : số lượng mẫu
* RMSE càng nhỏ, mô hình dự đoán càng chính xác. Nó thường được sử dụng làm độ đo để so sánh sự hiệu quả của các mô hình dự đoán khác nhau.

Mean Absolute Percentage Error (MAPE)

* là một phép đo đánh giá mức độ sai lệch tương đối giữa các giá trị dự đoán và giá trị thực tế trong dự đoán.
* MAPE được tính bằng công thức như sau:
* Với:
  + : giá trị thực tế
  + : giá trị dự đoán
  + : số lượng mẫu
* MAPE cung cấp một phần trăm đo lường về mức độ chênh lệch của dự đoán so với thực tế, giá trị càng thấp thì mô hình dự đoán càng chính xác. MAPE thường được sử dụng để đánh giá hiệu suất của các mô hình dự đoán trong các lĩnh vực như dự báo kinh doanh, dự báo thời tiết, hoặc dự đoán chứng khoán.
  + 1. Áp dụng cho Lab2
* Với bộ dataset VIETNAM NATIONAL HIGHSCHOOL EXAM SCORE 2018 (Available: <https://www.kaggle.com/code/kerneler/starter-vietnam-national-highschool-a18f500b-5>). Thực hiện đánh giá cho mô hình hồi quy tuyến tính cho biến phụ thuộc điểm toán, dựa vào 2 biến độc lập điểm hóa và điểm lý.
* Thực hiện loại bỏ dữ liệu trống và chia tâp train/val (test)

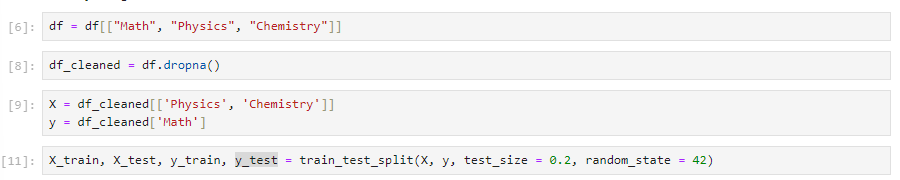


Figure 63 Làm sạch, tách tập train và test

* Sử dụng mô hình LinearRegression trong thư viện *sklearn*

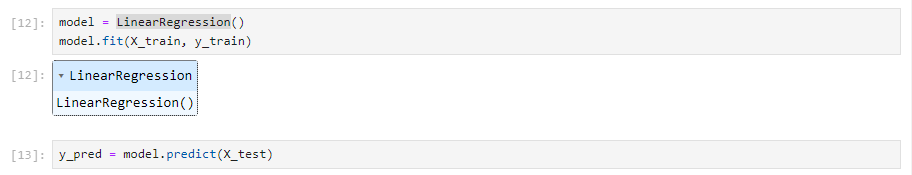


Figure 64 Sử dụng model LinearRegression của thư viện Sklearn

* Tìm các chỉ số MAE, MSE, R-Square

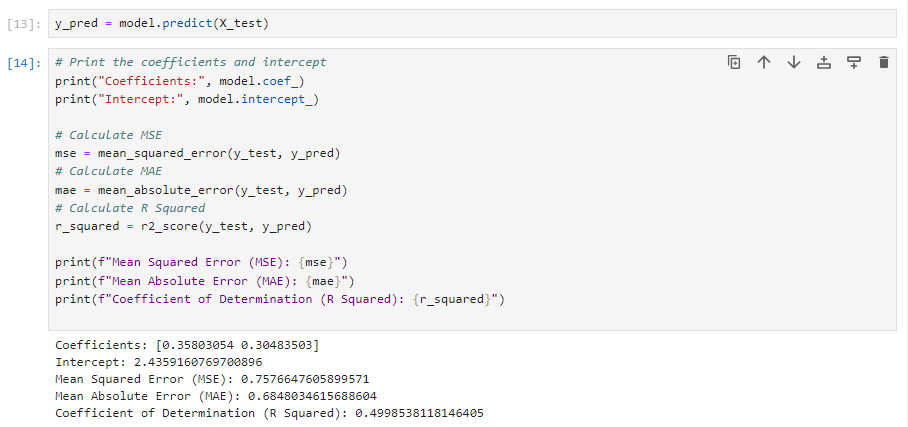
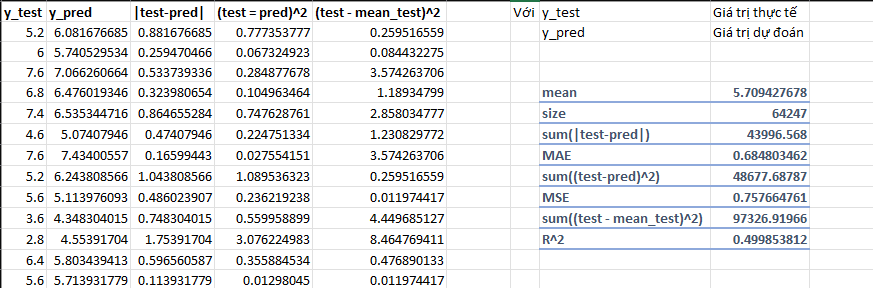


Figure 65 Kết quả đánh giá

* Thực hiện tính theo công thức được đề cập ở phần 3.b trong excels



Ta nhận thấy với MAE = 0.685, MSE = 0.76, R-Squared = 0.4999. Ta nhận thấy mô hình hồi quy tuyến tính cho bài toàn đưa ra kết quả với độ chính xác thấp.

* + 1. Áp dụng cho Lab3
* Với bộ dataset “vũng-tàu\_ngã tư giếng nước - tp.vũng tàu-air-quality”. Thực hiện đánh giá cho mô hình hồi quy phi tuyến tính cho biến phụ thuộc là pm10 dựa theo 4 biến độc lập là: pm25, o3, so2 và co
* Trên tập dữ liệu ta thực hiện thêm các cột mới để hình thành hàm phi tuyến:

A screenshot of a table

Description automatically generated

Figure 66 Thêm các cột mới

* Sau đó đổi tên các cột lần lượt thành X1, X2, X3, X4:

A table of numbers and digits

Description automatically generated

Figure 67 Thay đổi tên cột

* Thực hiện thêm các thư viện, dữ liệu và huấn luyện mô hình trên python:

A white rectangular object with text

Description automatically generated with medium confidence

Figure 68 thêm dữ liệu và huấn luyện mô hình

* Thực hiện tính RMSE bằng cách sử dụng thư viện sklearn:

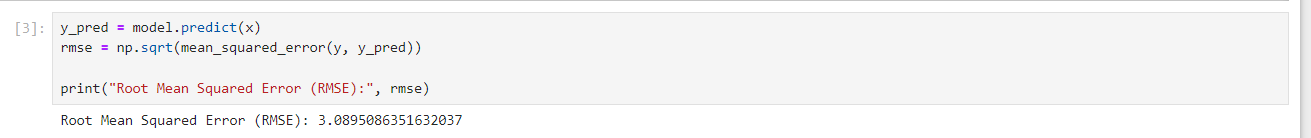


Figure 69 Tính RMSE

* Xây dựng hàm tính MAPE và sử dụng hàm:

A white rectangular object with a white border

Description automatically generated

Figure 70 Xây dựng và sử dụng hàm tính MAPE

* Ta thấy giá trị của RMSE và MAPE khá thấp (3.089 và 13.22%) cho thấy mô hình này có khả năng dự đoán tốt(độ chính xác cao) và ổn định dựa trên các dữ liệu đã kiểm tra.
  + 1. Áp dụng cho 4.1 và 4.2
* Bài 4.1
* Với bộ dataset “Gas & Electric”. Thực hiện đánh giá cho mô hình ARIMA dựa vào 2 biến “Gas Use”, “Electric Use”.
* Chia bộ dataset làm 2 phần, 1 phần ta sẽ sử dụng cột “Gas Use”, phần còn lại sẽ sử dụng cột “Electric Use” và áp dụng mô hình ARIMA vào cả 2 phần này.

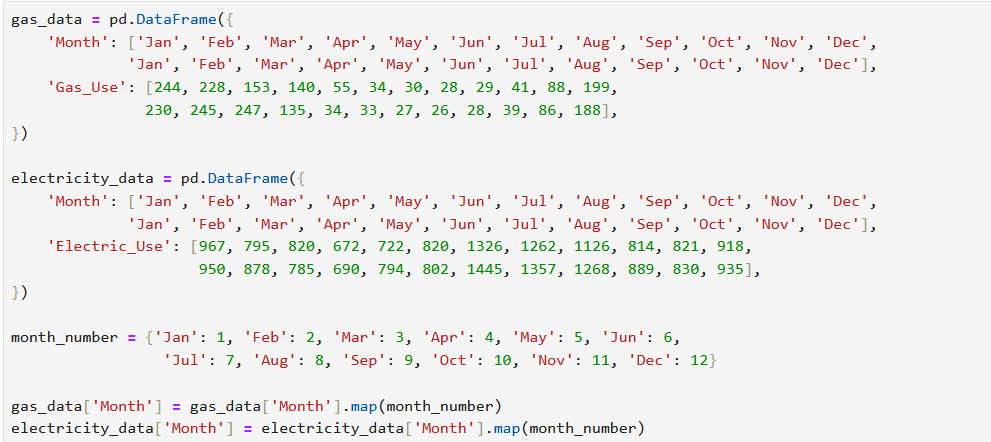


Figure 71 Chia dataset làm 2 phần

* Thực hiện loại bỏ dữ liệu trống và chia tâp train/val (test)

A screenshot of a computer program

Description automatically generated

Figure 72 Chia tập train/test cho Gas

A screenshot of a computer program

Description automatically generated

Figure 73 Chia tập train/test cho Electric

* Thực hiện quá trình huấn luyện mô hình và quá trình Test trên cả 2 phần

A screenshot of a computer program

Description automatically generated

Figure 74 Huấn luyện mô hình và quá trình test trên Gas

A screenshot of a computer program

Description automatically generated

Figure 75 Huấn luyện mô hình và quá trình test trên Electric

* Thực hiện kiểm định đánh giá mô hình ARIMA bằng cách tính toán RMSE, MAPE

A screenshot of a computer code

Description automatically generated

Figure 76 Đánh giá mô hình ARIMA trên Gas

* + Với phần sử dụng giá trị “Gas Use”, khi sử dụng mô hình ARIMA và thực hiện kiểm định, ta thấy giá trị của RMSE là 43.29, có độ sai lệch nhất định so với giá trị thực tế, nhưng sai lệch không quá lớn, ở mức trung bình. Còn với MAPE thì giá trị sai lệnh hơn 20%, ở mức trung bình. Dựa vào kết quả thì có thể thấy mô hình ARIMA có vẻ dự báo hiệu quả với tập dữ liệu được sử dụng, nhưng vẫn cần thêm các đánh giá để cải thiện độ chính xác.

A screenshot of a computer code

Description automatically generated

Figure 77 Đánh giá mô hình ARIMA trên Electric

* + Với phần sử dụng giá trị “Electric Use”, khi sử dụng mô hình ARIMA và thực hiện kiểm định, ta thấy giá trị của RMSE khá cao 134.65 có nghĩa là mô hình chêch lệch đáng kể so với giá trị thực tế. MAPE thấp 10.29% cho thấy dự báo của mô hình tương đối chính xác khi so sánh với tỷ lệ phần trăm. Mô hình dự báo sai lệch đáng kể so với giá trị thực tế, không chính xác về mặt tuyệt đối, mô hình cần được cải thiện.
* Thực hiện kiểm định ARIMA trên Excel.

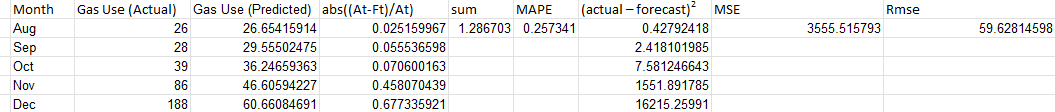


Figure 78 Kiểm định mô hình với các giá trị RMSE, MAPE, MSE

* + Với phần sử dụng “Gas Use”, ta nhận thấy các giá trị RMSE, MAPE, MSE lần lượt là 59.68; 25.73; 3555.51, đều có giá trị ở mức trung bình. Điều đó cho thấy mô hình dự báo sai lệch kha khá so với giá trị thực tế, nhưng vẫn có vài giá trị dự đoán sát với thực tế, nhưng vẫn cần thêm các đánh giá để cải thiện độ chính xác.

A screenshot of a data

Description automatically generated

Figure 79 Kiểm định mô hình với các giá trị RMSE, MAPE, MSE

* + Với phần sử dụng “Electric Use”, ta nhận thấy các giá trị RMSE, MAPE, MSE lần lượt là 128.03; 9.67%; 16391.58, đều có giá trị khá cao. Điều đó cho thấy mô hình dự báo sai lệch đáng kể so với giá trị thực tế, giá trị dự báo sai lệch khá nhiều so với giá trị thực tế. Mô hình không phù hợp để áp dụng, cần cải thiện mô hình để có thể cải thiện độ chính xác của các dự đoán.
* Tương tự, ta tiếp tục đánh giá sự chính xác của phương pháp SES được áp dụng vào dataset “Gas & Electric” nhưng sẽ thực hiện kiểm định trên Excel. Ở trên Excel ta sẽ sử dụng giá trị thực tế và giá trị dự đoán với

A screenshot of a data

Description automatically generated

Figure 80 Thực hiện kiểm định với phần sử dụng giá trị "Gas Use"

* + Với phần sử dụng giá trị “Gas Use”, khi sử dụng phương pháp SES và thực hiện kiểm định, ta thấy giá trị MAPE rất cao 218.68% điều đó cho thấy phương pháp dự báo không chính xác, không thể sử dụng để đưa ra quyết định trên dự báo về trên bộ dữ liệu sử dụng giá trị “Gas”. Có thể dẫn đến sai sót nghiêm trọng trong các ứng dụng thực tế.

A screenshot of a computer

Description automatically generated

Figure 81 Thực hiện kiểm định với phần sử dụng giá trị "Electric Use"

* + Với phần sử dụng giá trị “Electric Use”, khi sử dụng phương pháp SES và thực hiện kiểm định, ta thấp MAPE thấp 19.27% cho thấy dự báo của phương tương đối chính xác khi so sánh với tỷ lệ phần trăm, và theo độ chính xác khi dự đoán của phương pháp được đánh giá là tốt.
* Nhưng khi kiểm định phương pháp SES được áp dụng trên dataset “Gas & Electric” với python. Thì các giá trị kiểm định lại đánh giá phương pháp SES không phù hợp để thực hiện việc dự đoán với tập dataset được sử dụng.

A screenshot of a computer program

Description automatically generated

Figure 82 Thực hiện kiểm định SES với python

* + Các giá trị RMSE, MAPE, MSE, R-square của 2 phần sử dụng “Gas Use” và “Electric Use” đều rất cao lần lượt là (66.51; 150.65; 4424.75; -0.18) và (234.96; 15.93; 55207.41; -0.20), từ đó có thể thấy được trên python thì phương pháp SES không phù hợp để áp dụng vào dataset và dự đoán giá trị trong tương lai. Ta cần xem xét cản thiện hoặc thử nghiệm phương pháp khác để có thể đạt được kết quả tốt hơn.
* Bài 4.2
* Với bộ dữ liệu: các chỉ số chất lượng không khí tại Hà Nội từ 1/1/2014 – 25/3/2024. Thực hiện đánh giá phương pháp SES được áp dụng cho bộ dữ liệu này bằng cách sử dụng phương pháp RMSE và MAPE.
* Công thức RMSE:
* Công thức MAPE:
* Ta sẽ thực hiện đánh giá trên các giá trị dự đoán của α = 0.6. Đầu tiên ta sẽ tính toán RMSE.
* Ta sẽ tính toán giá trị

A table of numbers with black text

Description automatically generated

Figure 83 Tính toán

* Tiếp sau đó ta tính RMSE, kết quả ta nhận được là RMSE = 28.4061

A screenshot of a computer

Description automatically generated

Figure 84 Tính RMSE

* Ta tiếp tục tính toán giá trị MAPE, đầu tiên tính

A table of numbers and equations

Description automatically generated

Figure 85 Tính toán

* Sau khi có các giá trị ta lần lượt tính tổng của các giá trị đó, số lượng phần tử để phục vụ cho việc tính toán MAPE

A screenshot of a calculator

Description automatically generated

Figure 86 Tính toán lần lượt các giá trị sum và n

* Tính toán MAPE bằng cách lấy ta được giá trị MAPE = 0.274181 = 27.4181%.

A screenshot of a spreadsheet

Description automatically generated

Figure 87 Giá trị MAPE

* Nhận xét: ta có hai giá trị RMSE và MAPE lần lượt là 28.4061 và 27.4181 % là các giá trị trung bình, không quá cao cũng không quá thấp cho thấy phương pháp SES này có khả năng dự đoán tốt(độ chính xác tốt) và ổn định dựa trên bộ dữ liệu dùng để kiểm định.
* Với mô hình ARIMA:
  + ARIMA(1,0,1)

A screenshot of a spreadsheet

Description automatically generated

Figure 88 Các giá trị RMSE, MAPE, MSE của mô hình ARIMA (1,0,1)

* + - Các giá trị RMSE: 76.02556, MAPE: 70.0098%, MSE: 5779.886 đều khá cao, cho thấy mô hình dự đoán hoạt động không tốt trên tập dataset được chọn. Có thể cần điều chỉnh hoặc thay đổi mô hình, cải thiện quá trình huấn luyện hoặc sử dụng một phương pháp dự đoán khác để đạt được kết quả tốt hơn.
  + ARIMA(2,0,1)

A screenshot of a data

Description automatically generated

Figure 89 Các giá trị RMSE, MAPE, MSE của mô hình ARIMA (2,0,1)

* + - Các giá trị RMSE:80.23257, MAPE: 75.483%, MSE: 6437.3 đều khá cao, cho thấy mô hình dự đoán hoạt động không tốt trên tập dataset được chọn, và có thể đưa ra các giá trị dự đoán sai lệch nhiều so với giá trị thực tế. Có thể cần điều chỉnh hoặc thay đổi mô hình, cải thiện quá trình huấn luyện hoặc sử dụng một phương pháp dự đoán khác để đạt được kết quả tốt hơn.

**DANH MỤC TÀI LIỆU THAM KHẢO**

|  |  |
| --- | --- |
| [1] | A. Chugh, "MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?," Medium, 8 12 2020. [Online]. Available: https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e. |
| [2] | N. University, "Coefficient of Determination, R-squared," [Online]. Available: https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html#:~:text=%C2%AFy)2.-,R%202%20%3D%201%20%E2%88%92%20sum%20squared%20regression%20(SSR)%20total,from%20the. |

**­­­­**