Foreign\_tourist

**1. Giới thiệu dữ liệu:**

Tệp CSV có 1895 dòng. Các cột chính bao gồm:

* **Năm**: Năm thu thập dữ liệu.
* **Tháng**: Tháng của năm.
* **Chỉ tiêu**: Chỉ tiêu khách du lịch (đường không, đường biển, đường bộ, theo quốc gia/vùng lãnh thổ).
* **Ước tính mỗi tháng**: Số lượng khách du lịch ước tính trong tháng.
* **Tổng lượt khách cả năm**: Tổng số lượt khách của năm.
* **So với tháng trước (%)**: Tỷ lệ phần trăm so với tháng trước.
* **So sánh tháng này ở năm trước (%)**: Tỷ lệ phần trăm so với cùng tháng năm trước.
* **So sánh tổng lượt khách với cùng kỳ năm trước (%)**: Tỷ lệ so sánh tổng lượt khách với cùng kỳ năm trước.

**2. Vấn đề tiền xử lý:**

* **Kiểm tra dữ liệu thiếu (missing values)**: Một số cột có giá trị thiếu, đặc biệt là "So sánh tháng này ở năm trước (%)" và "So sánh tổng lượt khách với cùng kỳ năm trước (%)".
* **Dữ liệu kiểu object**: Cần chuyển đổi các cột như "Ước tính mỗi tháng" và "Tổng lượt khách cả năm" sang kiểu số.
* **Kiểm tra outlier**: Đặc biệt với các cột tỷ lệ phần trăm (%), có thể có giá trị bất thường.
* **Biến dummy (nếu cần)**: "Chỉ tiêu" có thể được mã hóa để sử dụng trong mô hình dự báo.
* **Đa cộng tuyến**: Kiểm tra sự tương quan giữa các biến.

### ****Phân tích Kết Quả Mô Hình ARIMA (4,1,5):****

#### **1. Tổng quan mô hình:**

* **Mô hình được chọn:** ARIMA(4,1,5)
* **Số quan sát:** 1874
* **Thời gian huấn luyện:** 45.431 giây
* **Log Likelihood:** -11672.1
* **AIC (Akaike Information Criterion):** 23364.2 (Thấp là tốt)
* **BIC (Bayesian Information Criterion):** 23419.55
* **HQIC:** 23384.59

#### **2. Phân tích các tham số:**

| **Tham số** | **Giá trị (coef)** | **P-value** | **Giải thích** |
| --- | --- | --- | --- |
| **ar.L1** | -1.3507 | 0.000 | Thành phần tự hồi quy bậc 1 có ý nghĩa |
| **ar.L2** | -1.5930 | 0.000 | Thành phần tự hồi quy bậc 2 có ý nghĩa |
| **ar.L3** | -1.2621 | 0.000 | Thành phần tự hồi quy bậc 3 có ý nghĩa |
| **ar.L4** | -0.9065 | 0.000 | Thành phần tự hồi quy bậc 4 có ý nghĩa |
| **ma.L1** | 0.5036 | 0.000 | Thành phần trung bình trượt bậc 1 có ý nghĩa |
| **ma.L2** | 0.5913 | 0.000 | Thành phần trung bình trượt bậc 2 có ý nghĩa |
| **ma.L3** | 0.0749 | 0.004 | Thành phần trung bình trượt bậc 3 có ý nghĩa |
| **ma.L4** | -0.0080 | 0.709 | Không có ý nghĩa (P-value > 0.05) |
| **ma.L5** | -0.6979 | 0.000 | Thành phần trung bình trượt bậc 5 có ý nghĩa |

* **Sigma2:** 1.524e+04 – Phương sai của sai số (càng nhỏ càng tốt).

#### **3. Các kiểm định thống kê:**

| **Kiểm định** | **Giá trị** | **Ý nghĩa** |
| --- | --- | --- |
| **Ljung-Box (L1):** | 7.43 | Kiểm định phần dư, nếu nhỏ cho thấy phần dư là white noise (tốt). |
| **Prob(Q):** | 0.01 | Phần dư có thể chưa hoàn toàn là white noise. |
| **Heteroskedasticity (H):** | 0.91 | Dữ liệu không có phương sai thay đổi. |
| **Prob(H):** | 0.26 | Không có heteroskedasticity (giá trị này > 0.05 là tốt). |
| **Skew (Độ lệch):** | 1.62 | Dữ liệu hơi lệch phải. |
| **Kurtosis:** | 6.81 | Phân phối dữ liệu có đuôi dài (kurtosis > 3). |

### ****4. Đánh giá mô hình:****

* **P-value cho tất cả các thành phần ar (1 đến 4) và ma (1 đến 3, 5)** đều rất nhỏ (p < 0.05), cho thấy các tham số này có ý nghĩa.
* **Thành phần ma.L4 có p-value = 0.709 (không có ý nghĩa thống kê).** Có thể thử loại bỏ hoặc giảm bậc của MA để tinh chỉnh mô hình.
* **AIC/BIC thấp:** Mô hình đang khá tốt nhưng có thể thử nghiệm thêm các giá trị (p, d, q) để giảm AIC/BIC hơn nữa.

### ****5. Hướng cải thiện:****

1. **Kiểm tra lại phần dư (Residuals):**  
   Nếu phần dư không phải white noise, có thể thực hiện tuning lại tham số hoặc chuyển sang **SARIMA** nếu dữ liệu có tính mùa vụ.
2. **Loại bỏ tham số không ý nghĩa:**
   * Thử huấn luyện lại mô hình với ARIMA(4,1,4) (loại bỏ ma.L4).
3. model = ARIMA(data, order=(4,1,4))
4. result = model.fit()
5. print(result.summary())
6. **Tuning thêm:**
   * Dùng auto\_arima với phạm vi lớn hơn cho (p, d, q).
7. auto\_model = auto\_arima(data, seasonal=False, start\_p=0, start\_q=0,
8. max\_p=6, max\_q=6, trace=True)

### ****Kết luận:****

* Mô hình ARIMA(4,1,5) hoạt động khá tốt với dữ liệu của bạn.
* Có thể tối ưu thêm bằng cách loại bỏ các thành phần không ý nghĩa (ma.L4).
* Nếu AIC/BIC vẫn cao, có thể thử SARIMA để xử lý dữ liệu có tính mùa vụ.

**Phân tích biểu đồ dự báo ARIMA(4,1,5) và ARIMA(4,1,4):**

**1. Tổng quan về biểu đồ:**

* **Đường xanh dương:** Dữ liệu thực tế (quá khứ).
* **Đường cam:** Dự báo từ mô hình ARIMA(4,1,5).
* **Đường xanh lá:** Dự báo từ mô hình ARIMA(4,1,4).

**2. Những điều biểu đồ thể hiện:**

1. **Dữ liệu thực tế (Đường xanh dương):**
   * Có sự sụt giảm mạnh vào giai đoạn giữa (khoảng năm 2020-2021), có thể là do tác động của COVID-19 hoặc một yếu tố đặc biệt.
   * Sau năm 2022, dữ liệu tăng mạnh trở lại, cho thấy sự phục hồi.
   * Dữ liệu có tính chu kỳ rõ ràng với các đỉnh và đáy (tính mùa vụ hoặc dao động theo tháng).
2. **Dự báo ARIMA (Đường cam và xanh lá):**
   * **Cả hai mô hình ARIMA(4,1,5) và ARIMA(4,1,4)** cho thấy dự báo khá tương đồng, dao động nhẹ và ổn định trong khoảng thời gian dự báo (2024-2025).
   * **ARIMA(4,1,5)** có dự báo dao động nhiều hơn một chút so với ARIMA(4,1,4).

**3. Ý nghĩa của dự báo:**

* **Sự ổn định:**
  + Cả hai mô hình đều cho thấy sự ổn định trong tương lai gần (12 tháng tiếp theo).
  + Không có sự tăng trưởng hoặc sụt giảm mạnh nào được dự báo.
  + Dự báo dao động nhẹ, cho thấy sự phục hồi chậm nhưng ổn định.
* **Khả năng mô hình:**
  + Mô hình ARIMA đã bắt kịp xu hướng dữ liệu sau giai đoạn sụt giảm.
  + **Không có dấu hiệu dự báo sai lệch lớn**, chứng tỏ mô hình hoạt động tốt với dữ liệu.

**4. Kết luận:**

* **Dự báo ổn định, không có biến động lớn.**
* **ARIMA(4,1,5) và ARIMA(4,1,4)** cho kết quả tương đương, có thể chọn mô hình đơn giản hơn (ARIMA(4,1,4)) để tránh overfitting.
* Nếu mong muốn mô hình dự báo có xu hướng tăng/giảm rõ ràng hơn, có thể:
  + Kiểm tra và xử lý thêm tính mùa vụ (SARIMA).
  + Thử nghiệm các mô hình khác như **Prophet hoặc LSTM (deep learning)**.

Nếu bạn muốn đi sâu vào phân tích hoặc tinh chỉnh mô hình, mình sẵn sàng hỗ trợ thêm!