**Evaluate metric**

| ARIMA | Train - test - validate | MAE | MAPE | RMSE | MSE | R-squared |
| --- | --- | --- | --- | --- | --- | --- |
| 70 - 20 -10 | 2025.80 | 19.05 | 2287.82 | 5234108.27 | -3.1977 |
| **75 - 10 - 15** | 952.52 | 8.95 | 1097.31 | 1204092.03 | -0.0659 |
| 65 - 10 - 25 | 3426.99 | 31.02% | 3756.95 | 14114678.32 | -4.9159 |

**Lý thuyết**

* **ARIMA (Autoregressive Intergrated Moving Average)**: biểu diễn phương trình hồi

qui tuyến tính đa biến (multiple linear regression) của các biến đầu vào (còn gọi

là biến phụ thuộc).

* Là một trong những mô hình phổ biến nhất trong phân tích chuỗi thời gian.
* Mô phỏng và dự báo các chuỗi thời gian, đặc biệt là khi dữ liệu không có tính

chất mùa vụ (seasonal).

* ARIMA kết hợp ba thành phần chính:
  + **AR** (AutoRegressive): Tự hồi quy
  + **I** (Integrated): Tích hợp
  + **MA** (Moving Average): Trung bình động
* Mô hình ARIMA có 3 tham số chính thường được gọi là **p, d, và q.**
  + **p: Bậc của phần Tự hồi quy (Autoregressive)**: Số lượng giá trị trong quá khứ

được sử dụng để dự báo giá trị hiện tại.

* **d: Bậc sai phân (differencing) liên quan**: Số lần lấy sai phân của chuỗi thời

gian để đạt được độ ổn định, chuỗi dừng.

* **q: Bậc của phần Trung bình động (Moving Average)**. Số lượng nhiễu trắng

(các giá trị sai lệch ngẫu nhiên) trong quá khứ được sử dụng để dự báo giá trị hiện tại.

* Những tham số này thường được viết theo dạng: **ARIMA(p, d, q).**
* **AR(p)**
  + Trong mô hình AR, giá trị hiện tại của chuỗi thời gian được dự báo dựa trên các giá trị trong quá khứ của chính nó.
  + Mỗi giá trị của chuỗi thời gian tại thời điểm t có thể được giải thích bằng một tổ hợp tuyến tính của các giá trị trước đó (độ trễ p: p các giá trị trước đó).
  + Mô hình **AR(p)** tại thời điểm t:

yₜ = c + φ₁yₜ₋₁ + φ₂yₜ₋₂ + ⋯ + φₚyₜ₋ₚ + εₜ

* Trong đó:
  + yₜ: giá trị tại thời điểm t
  + yₜ₋₁, yₜ₋₂, yₜ₋ₚ: các giá trị tại các thời điểm trước đó
  + φ₁, φ₂, φₚ: các hệ số tự hồi quy tại các lags tương ứng
  + ₚ: số độ trễ được đưa vào mô hình (các lags)
  + εₜ: sai số tự nhiên (residuals) tại thời điểm t
* **I(d)**
  + Lấy sai phân (differencing) chuỗi thời gian để làm cho chuỗi trở thành **dữ liệu dừng** (stationary - các đặc điểm thống kê không thay đổi theo thời gian).
  + Sai phân chỉ ra sự khác nhau giữa giá trị hiện tại và giá trị trước đó. Phân tích sai phân làm cho ổn định giá trị trung bình của chuỗi dữ liệu, biến đổi chuỗi thành một chuỗi dừng.

I(d) = Δᵈ(yₜ)

* **MA(q)**
  + Mô hình hóa mối quan hệ giữa giá trị tại thời điểm t và các sai số (residuals) trong quá khứ. MA có vai trò điều chỉnh dự báo của mô hình dựa trên các sai số từ các dự báo trước đó.
  + Mô hình MA(q) mô trả giá trị chuỗi tại thời điểm t là một tổ hợp tuyến tính của các sai số ngẫu nhiên trong quá khứ:

yₜ = μ + εₜ + θ₁εₜ₋₁ + θ₂εₜ₋₂ + ⋯ + θqεₜ₋q

* Trong đó:
  + yₜ: giá trị tại thời điểm t
  + μ: Trung bình của chuỗi thời gian (nếu có).
  + εₜ: Sai số ngẫu nhiên (residual) tại thời điểm t. Sai số này là sự khác biệt giữa giá trị thực tế và giá trị dự báo. Nếu giá trị dự báo của mô hình là chính xác.
  + εₜ₋₁, εₜ₋₂, εₜ₋q: các sai số ngẫu nhiên tại các thời điểm trước đó (các lags)
  + θ₁, θ₂, θq: Các hệ số của mô hình MA tại các độ trễ 1, 2,...q. Xác định mức độ ảnh hưởng của các sai số quá khứ (từ các độ trễ trước đó) đến giá trị hiện tại.
  + q: Số độ trễ (lags) của sai số đưa vào mô hình MA.

**Tham khảo:**

**“Introduction to ARIMA models”** [**https://people.duke.edu/~rnau/411arim.htm**](https://people.duke.edu/~rnau/411arim.htm)

# **“How To Forecast Time-Series Using Autoregression**” https://github.com/egorhowell/Youtube/blob/main/Time-Series-Crash-Course/17.%20ARIMA.ipynb