# **Tiền xử lý dữ liệu**

## **Đọc và kiểm tra dữ liệu**

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.A screen shot of a computer

AI-generated content may be incorrect.

* Lọc các biến numeric và các biến categorical

A screen shot of a computer

AI-generated content may be incorrect.

## **Xử lý giá trị thiếu**

* Điền giá trị “median” cho các giá trị thiếu ở biến numeric
* Điền chuỗi “unknow” cho các giá trị thiếu ở biến categorical

A screen shot of a computer program

AI-generated content may be incorrect.

## **Tạo biến giả (dummy) cho ngày trong tuần:**

* get\_dummies chuyển mỗi giá trị của Publication\_Day (ví dụ “Sunday”, “Monday”…) thành một cột nhị phân.
* Loại bỏ cột gốc và ghép các cột dummy vào DataFrame.

A screen shot of a computer program

AI-generated content may be incorrect.

# **Chạy mô hình**

## **Chia tập train/test:**

* Lấy 80% dòng đầu làm train, 20% dòng sau làm test.
* Dùng .copy() để tránh cảnh báo SettingWithCopy khi chỉnh sửa sau này.

A computer screen with text

AI-generated content may be incorrect.

## **Xác định biến mục tiêu (target):**

* y\_train, y\_test là chuỗi thời gian bạn muốn dự đoán (ở đây là thời gian nghe podcast).
* steps là số bước forecast = kích thước tập test.

A computer screen shot of text

AI-generated content may be incorrect.

## **Khởi tạo và huấn luyện 03 mô hình Kalman:**

* **Local Level**: chỉ có thành phần mức độ (μt).
* **Local Linear Trend**: thêm thành phần xu hướng.
* **Level + Weekly Seasonality**: thêm seasonal tuần (S=7).
* fit(disp=False) huấn luyện và tắt log hiển thị.

A screen shot of a computer code

AI-generated content may be incorrect.

## **Dự đoán (Forecast):**

* Với mỗi mô hình đã fit, gọi get\_forecast(steps) để sinh dự đoán cho steps bước kế tiếp.
* Lấy về chuỗi giá trị trung bình dự báo.

A computer screen with text

AI-generated content may be incorrect.

## **Đánh giá hiệu suất:**

* Tính **RMSE** và **MAE** giữa y\_test và y\_pred cho từng mô hình.

A screen shot of a computer code

AI-generated content may be incorrect.

## **Trực quan hóa:**

A computer screen shot of text

AI-generated content may be incorrect.

1. **LocalLevel – Original vs Filtered**

* **Original (xanh nhạt)**: dữ liệu gốc rất dao động, nhiễu lớn, không rõ xu hướng.
* **Filtered (cam)**: lọc xong chỉ còn một đường gần như phẳng, khởi đầu có nhảy rồi ngay lập tức hội tụ quanh ~45.
* **Nhận xét**: Bộ lọc hơi “quá khớp” với giả thiết mức độ quá địa phương nên hầu như bỏ qua mọi biến động thực tế, chỉ giữ lại một hằng số.

A blue and orange line

AI-generated content may be incorrect.

1. **LocalLevel – Forecast vs Actual**

* **Actual (xanh nhạt)**: tập test vẫn dao động mạnh.
* **Forecast (cam)**: đường dự đoán hoàn toàn phẳng, bằng giá trị lọc cuối của train (~45).
* **Nhận xét**: Mô hình đơn giản này dự báo không thay đổi theo thời gian, nên không phù hợp với chuỗi nhiều nhiễu và không-stationary.

A blue and orange line

AI-generated content may be incorrect.

1. **LocalLinearTrend – Original vs Filtered**

* **Filtered**: không còn “cứng” như LocalLevel, đã ghi nhận một phần biến động chậm hơn nhưng vẫn rất dập dờn trong khoảng ~35–60.
* **Nhận xét**: Thêm thành phần trend giúp bộ lọc bám theo dao động chậm, nhưng biên độ bị hạn chế—nó vẫn chưa bắt kịp độ dao động nhanh của dữ liệu gốc.

A blue and orange sound wave

AI-generated content may be incorrect.

1. **LocalLinearTrend – Forecast vs Actual**

* **Forecast**: đường dự báo hơi có xu hướng giảm dần (từ ~40 xuống ~39), nhưng vẫn gần như hằng số.
* **Nhận xét**: Mặc dù có trend component, mô hình chỉ extrapolate một cách rất nhẹ nhàng, không thể theo kịp đỉnh/trough ngắn hạn trong tập test.

A blue and orange line

AI-generated content may be incorrect.

1. **Level+WeeklySeason – Original vs Filtered**

* **Filtered**: gần như trùng với LocalLevel—chỉ một mức độ ổn định ~45, hoàn toàn không thấy pattern tuần (seasonal=7).
* **Nhận xét**: Rõ ràng mô hình không tìm ra thành phần mùa vụ. Có thể do dữ liệu index không phải datetime hay dummy weekday chưa được đưa vào state-space, nên seasonal component mặc định “chết yểu”.

A blue and orange line

AI-generated content may be incorrect.

1. **Level+WeeklySeason – Forecast vs Actual**

* **Forecast**: phẳng, giống hệt LocalLevel.
* **Nhận xét**: Mô hình này không cải thiện được forecast so với LocalLevel vì seasonal không được học.

A blue and orange line

AI-generated content may be incorrect.

**Kết luận**

* **Tất cả các forecast đều gần như phẳng** và không bám theo dao động thực tế → cần điều chỉnh hoặc khởi tạo tham số Q/R hợp lý hơn, hoặc thêm features (chẳng hạn dummy weekday) vào phần state-space để mô hình “thấy” mùa vụ.
* **Filtered states quá mượt** hoặc quá “cứng”, cho thấy mô hình state-space đang ưu tiên quá mức cho process noise so với observation noise (Q quá nhỏ, R quá lớn) hoặc ngược lại—cần tune MLE hoặc đặt giá trị khởi điểm thủ công.
* **X-axis** đang là index dòng, nếu chuyển sang datetime sẽ trực quan hơn khi quan sát seasonal/week pattern.

**Link github:**

<https://github.com/nguyentiendat1401/TimeSeries/tree/main/thucHanh1>