**3.4.2 Mô hình Hồi quy tuyến tính**

**3.4.2.1 Đánh giá dữ liệu trước khi thực hiện thuật toán**

**A graph with orange and blue lines

AI-generated content may be incorrect.**

*Hình : Trực quan hóa xu hướng giá vàng theo thời gian*

Dựa vào nhận xét phần trực quan hoá dữ liệu đã trình bày, dữ liệu dao động mạnh, với các biến động đáng kể trong một số giai đoạn cụ thể. Đồng thời, dữ liệu thể hiện xu thế tăng dài hạn rõ ràng, với tốc độ tăng trưởng gia tăng, đặc biệt trong giai đoạn gần đây. Dựa trên các đặc điểm này, dữ liệu giá được đánh giá là phù hợp để áp dụng mô hình dự báo thời gian dựa trên hồi quy tuyến tính.

**3.4.2.2 Xây dựng mô hình**

**Bước 1: Thêm cột Time**

|  |
| --- |
|  |

*Hình :Hiển thị bảng sau khi thêm cột time*

Sử dụng thư viện numpy để tạo một cột mới gọi là Time trong một DataFrame (df). Cột này được điền bằng một dãy số nguyên từ 0 đến độ dài của chỉ mục DataFrame. Phương thức head() sau đó được sử dụng để hiển thị vài hàng đầu tiên của DataFrame, bao gồm các cột Date, Buy Price, Sell Price, và Time.

**Bước 2: Xác định các đặc trưng và mục tiêu từ dữ liệu huấn luyện**

|  |  |
| --- | --- |
| **Buy\_Price** | **Sell\_Price** |
|  |  |

*Hình : Các đặc trưng và mục tiêu huấn luyện*

Sử dụng dữ liệu thời gian (Time) làm đặc trưng đầu vào để huấn luyện mô hình dự đoán. Mục tiêu đầu ra là hai biến giá gồm Buy Price (giá mua) và Sell Price (giá bán). Cụ thể, khi dự đoán giá mua, biến mục tiêu là Buy Price; tương tự, khi dự đoán giá bán, biến mục tiêu là Sell Price. Việc lựa chọn đặc trưng và mục tiêu như vậy giúp mô hình học được mối quan hệ giữa thời gian và giá, từ đó đưa ra các dự đoán phù hợp cho từng trường hợp cụ thể.

**Bước 3: Huấn luyện mô hình**

|  |  |
| --- | --- |
| **Buy\_Price** | **Sell\_Price** |
|  |  |
| **Kết quả:** | **Kết quả:** |

*Hình : Kết quả xây dựng mô hình hồi quy*

**Nhận xét:**

Mô hình hồi quy tuyến tính theo thời gian (Time-series forecasting with regression) được áp dụng cho cả chuỗi giá mua và giá bán cho thấy xu hướng tăng rõ rệt. Đường hồi quy màu cam trong cả hai biểu đồ thể hiện một xu hướng tuyến tính tăng dần, trong khi dữ liệu thực tế dao động quanh đường này với biên độ ngày càng lớn.

Cụ thể, Ở giai đoạn đầu (từ điểm 0 đến khoảng 1500), cả giá mua và giá bán đều dao động quanh đường hồi quy với biên độ vừa phải, cho thấy mô hình có thể phản ánh tương đối xu hướng trong giai đoạn này. Tuy nhiên, từ khoảng điểm 4000 trở đi, dữ liệu thực tế thể hiện các biến động mạnh, với các đợt tăng giá đột ngột (đỉnh giá vượt 80), trong khi đường hồi quy vẫn giữ nguyên độ dốc tuyến tính, không bắt kịp được sự thay đổi nhanh của chuỗi dữ liệu. Tại điểm cuối cùng trên biểu đồ (~5500), giá mua và giá bán thực tế đều đạt trên 85, trong khi mô hình chỉ dự đoán khoảng 60, chênh lệch khoảng 25 đơn vị – phản ánh mô hình hồi quy tuyến tính đang đánh giá thấp các biến động mạnh mẽ gần đây.

Vì vậy, mô hình này không phản ánh được các dao động phi tuyến và các cú sốc đột biến trong dữ liệu.

**3.4.2.3 Đánh giá mô hình**

|  |  |
| --- | --- |
| **Buy\_Price** | **Sell\_Price** |
|  |  |
| **Kết quả:** | **Kết quả:** |

*Hình : Kết quả đánh giá mô hình*

**Mô hình dự đoán giá mua (Buy\_Price):** Trên tập dữ liệu, mô hình hồi quy tuyến tính cho thấy sai số dự đoán ở mức trung bình, với độ lệch tuyệt đối trung bình (MAD) là **7.32** và tổng bình phương sai số (SSE) là **409301.35**. Điều này có nghĩa là, trung bình mỗi dự đoán của mô hình lệch khoảng 7.32 đơn vị so với giá mua thực tế. Mặc dù xu hướng chung được mô hình mô tả khá tốt qua đường hồi quy, nhưng biểu đồ cho thấy ở giai đoạn sau (sau điểm 4000), mô hình dự báo thấp hơn nhiều so với thực tế khi giá tăng vọt, làm tăng tổng sai số (SSE) đáng kể.

**Mô hình dự đoán giá bán (Sell\_Price):** Tương tự, mô hình hồi quy tuyến tính áp dụng cho giá bán có MAD là **7.50** và SSE là **432874.63**, cao hơn một chút so với mô hình giá mua. Như biểu đồ minh họa, sự khác biệt giữa giá thực tế và đường dự đoán ngày càng tăng về cuối chuỗi, đặc biệt khi giá bán thực tế vượt mốc 85 trong khi mô hình chỉ dự đoán ở mức khoảng 60. Điều này dẫn đến sai số tổng thể lớn, cho thấy mô hình chưa theo kịp các biến động phi tuyến mạnh mẽ trong dữ liệu.

**Tổng kết:** Cả hai mô hình đều thể hiện khả năng nhận diện xu hướng tuyến tính của chuỗi thời gian, nhưng chưa phản ánh tốt các biến động lớn và đột ngột. Giá trị MAD ở mức khoảng 7.3–7.5 và SSE trên 400,000