BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC THĂNG LONG**

Logo

Description automatically generated

**BÁO CÁO**

**MACHINE LEARNING**

**WALMART RECRUITING - STORE SALES FORECASTING**

**GIẢNG VIÊN: NHÓM SV THỰC HIỆN: A46219\_Lê Hải Sơn**

**Nguyễn Thế Năng A46456\_Đỗ Hoàng Hiệp**

**A43038\_Kim Văn Toản**

**A42264\_Phạm Duy Đạt**

**HÀ NỘI – 2024**

**MỤC LỤC**

[Chương 1. Tóm tắt và giới thiệu bài toán 1](#_Toc185715227)

[1.1 Tóm tắt 1](#_Toc185715228)

[1.2 Giới thiệu bài toán 1](#_Toc185715229)

[Chương 2. Dữ liệu và tiền xử lý 3](#_Toc185715230)

[2.1 Mô tả dữ liệu 3](#_Toc185715231)

[2.1.1 Nguồn gốc dữ liệu: 3](#_Toc185715232)

[2.1.2 Các biến chính: 3](#_Toc185715233)

[Chương 3. Cách tiếp cận 4](#_Toc185715234)

[3.1 Tiền xử lý dữ liệu 4](#_Toc185715235)

[3.2 Phương pháp mô hình hóa 4](#_Toc185715236)

[3.3 Đánh giá hiệu suất 5](#_Toc185715237)

[3.4 Trực quan hóa 5](#_Toc185715238)

[Chương 4. Kết quả và kết luận 6](#_Toc185715239)

[Chương 5. Tài liệu tham khảo 7](#_Toc185715240)

# Tóm tắt và giới thiệu bài toán

## Tóm tắt

Báo cáo này trình bày ứng dụng mô hình ARIMA (AutoRegressive Integrated Moving Average) trong dự báo doanh thu hàng tuần của các cửa hàng trong tập dữ liệu *Walmart Recruiting - Store Sales Forecasting*. Quá trình phân tích bao gồm tiền xử lý dữ liệu, kiểm tra tính dừng, lựa chọn tham số mô hình, và đánh giá hiệu năng dự báo. Kết quả cho thấy ARIMA có thể dự đoán hiệu quả doanh thu ngắn hạn, nhưng có hạn chế khi xử lý các yếu tố phi tuyến hoặc biến động lớn.

## Giới thiệu bài toán

Walmart Recruiting - Store Sales Forecasting là một bài toán khoa học dữ liệu kinh điển tập trung vào việc xây dựng các mô hình dự đoán để dự báo doanh số trong tương lai cho các cửa hàng Walmart. Dự báo doanh thu là một nhiệm vụ quan trọng trong ngành bán lẻ, đặc biệt đối với các chuỗi cửa hàng lớn như Walmart. Dự báo chính xác giúp quản lý hàng tồn kho, tối ưu hóa chuỗi cung ứng, và cải thiện trải nghiệm khách hàng. Dựa trên tập dữ liệu *Walmart Recruiting - Store Sales Forecasting*, báo cáo này áp dụng mô hình ARIMA để dự báo doanh thu hàng tuần của một cửa hàng cụ thể.

* Mục tiêu nghiên cứu:
* Dự đoán doanh số bán hàng tại các cửa hàng Walmart.
* Phân tích các yếu tố ảnh hưởng đến doanh số như ngày lễ, kinh tế vĩ mô và loại cửa hàng.
* Phương pháp chính:
* Sử dụng các kỹ thuật **phân tích dữ liệu** và **mô hình** **học máy** (Machine Learning).
* Áp dụng các mô hình dự báo theo thời gian (Time-series) để dự đoán doanh số.
* Kết quả đạt được:
* Xây dựng mô hình dự báo doanh số chính xác với các chỉ số đánh giá hiệu suất tốt như MASE.
* Phân tích ảnh hưởng của ngày lễ và các yếu tố kinh tế đến doanh số bán hàng.
* Ý nghĩa của nghiên cứu:
* Giúp Walmart tối ưu hóa quản lý hàng tồn kho và nguồn lực.
* Hỗ trợ đưa ra quyết định chiến lược nhằm tăng doanh số bán hàng và nâng cao hiệu quả hoạt động.
* Cung cấp giải pháp thực tế cho các doanh nghiệp bán lẻ trong việc dự đoán và lập kế hoạch kinh doanh.

# Dữ liệu và tiền xử lý

## Mô tả dữ liệu

### Nguồn gốc dữ liệu:

* Dữ liệu được thu thập từ hoạt động kinh doanh bán lẻ thực tế của Walmart.
* Nó bao gồm thông tin về doanh số bán hàng hàng tuần của các phòng ban tại từng cửa hàng Walmart. Dữ liệu cũng bổ sung thông tin kinh tế và các yếu tố ảnh hưởng khác như thời tiết, giá nhiên liệu, và tuần có ngày lễ.

### Các biến chính:

|  |  |  |  |
| --- | --- | --- | --- |
| **Tên Biến** | **Mô Tả** | **Loại Biến** | **Đơn Vị** |
| Store | Mã số của cửa hàng. | Categorical |  |
| Dept | Mã số của phòng ban. | Categorical |  |
| Date | Ngày bắt đầu tuần. | Date/Time | YYYY-MM-DD |
| Weekly\_Sales  (Biến mục tiêu) | Doanh số bán hàng hàng tuần của phòng ban. | Continuous | USD |
| IsHoliday | Tuần có ngày lễ hay không (0: Không, 1: Có). | Binary | 0/1 |
| Temperature | Nhiệt độ trung bình tại khu vực. | Continuous | Độ F |
| Fuel\_Price | Giá nhiên liệu tại khu vực. | Continuous | USD |
| CPI | Chỉ số giá tiêu dùng (Consumer Price Index) tại khu vực. | Continuous |  |
| Unemployment | Tỷ lệ thất nghiệp tại khu vực. | Continuous | % (Phần trăm) |
| Type | Loại cửa hàng (A: lớn, B: trung bình, C: nhỏ). | Categorical | A/B/C |
| Size | Diện tích cửa hàng. | Continuous | Square feet |
| MarkDown1-MarkDown5 | Mỗi biến đại diện cho một loại giảm giá hoặc khuyến mãi khác nhau được áp dụng tại cửa hàng. | Continuous |  |

# Cách tiếp cận

## Tiền xử lý dữ liệu

1. **Lọc dữ liệu**: Tập trung vào Store 1, Dept 1.
2. **Chỉ mục thời gian**: Sử dụng Date làm chỉ mục và sắp xếp theo thời gian tăng dần.
3. **Xử lý giá trị thiếu**: Điền NaN trong các cột bị trễ (Lag\_1) bằng 0.
4. **Tạo đặc trưng**:

* Time: Số tuần tính từ điểm bắt đầu.
* Lag\_1: Doanh số tuần trước.
* Thành phần Fourier: Mô hình hóa chu kỳ.

1. **Làm mượt dữ liệu**: Sử dụng trung bình trượt để quan sát xu hướng dài hạn.
2. **Trực quan hóa**: Kiểm tra xu hướng và mùa vụ để hiểu rõ hơn về dữ liệu.

## Phương pháp mô hình hóa

**a. Linear Regression**

* Sử dụng DeterministicProcess:
* Tạo các đặc trưng xu hướng (trend) tuyến tính hoặc phi tuyến tính.
* Xây dựng mô hình hồi quy tuyến tính đơn giản để mô hình hóa xu hướng dài hạn.
* Dự đoán và dự báo:
* Dự đoán xu hướng trong giai đoạn dữ liệu lịch sử và mở rộng dự báo cho tương lai.

**b. Moving Average**

Sử dụng trung bình trượt (moving average) để quan sát và phân tích xu hướng mượt của dữ liệu.

**c. XGBoost**

* Residuals + Boosting:
* Trích xuất các giá trị dư (residuals) từ mô hình tuyến tính.
* Huấn luyện mô hình XGBoost trên phần dư để cải thiện độ chính xác bằng cách nắm bắt các biến động ngắn hạn.

**d. ARIMA và SARIMA**

* ARIMA:
* Áp dụng mô hình ARIMA để dự đoán doanh số với các tham số thủ công (order).
* SARIMA:
* Sử dụng mô hình SARIMA để kết hợp yếu tố mùa vụ trong dự đoán doanh số bán hàng.

**e. Auto ARIMA**

Tự động hóa việc lựa chọn tham số mô hình ARIMA bằng cách sử dụng thư viện pmdarima.

**f. Phân tích mùa vụ (Seasonality Analysis)**

Sử dụng Fourier Transform (CalendarFourier) và phân tích phân rã chuỗi thời gian (STL) để nắm bắt các yếu tố mùa vụ trong dữ liệu.

## Đánh giá hiệu suất

* Sử dụng các thước đo để đánh giá hiệu suất dự đoán, bao gồm:
* **Mean Absolute Scaled Error (MASE)**: Đo lường mức độ chính xác của dự báo.
* Đánh giá đồ thị: So sánh trực quan giữa dữ liệu thực tế và giá trị dự đoán.

## Trực quan hóa

* Vẽ biểu đồ để trực quan hóa xu hướng, biến động, và dự báo:
* Biểu đồ so sánh doanh số thực tế với xu hướng.
* Biểu đồ lag plot để phân tích tương quan tự động (autocorrelation).
* Biểu đồ dự báo ngắn hạn và dài hạn.

# Kết quả và kết luận

Kết quả dự báo doanh thu cho cửa hàng được trình bày như sau:

* **MASE:** 0.7059008368435488

Dự báo cho thấy mô hình ARIMA phù hợp để nắm bắt xu hướng ngắn hạn, nhưng hiệu quả giảm đối với các tuần có kỳ nghỉ lễ (biến động lớn).

**Đồ thị kết quả:**

* **Dự báo so với giá trị thực:** Đường dự báo khớp tốt với giá trị thực trong giai đoạn ngắn.
* **Sai số dự báo:** Sai số tăng vào các tuần có kỳ nghỉ lễ do biến động doanh thu.

**Kết luận:**

Mô hình ARIMA cho thấy hiệu quả trong việc dự báo doanh thu hàng tuần cho cửa hàng Walmart. Kết quả nhấn mạnh tầm quan trọng của việc chọn tham số mô hình và xử lý dữ liệu. Tuy nhiên, ARIMA gặp hạn chế khi đối mặt với các yếu tố phi tuyến hoặc các biến động lớn trong dữ liệu (như kỳ nghỉ lễ).

# Tài liệu tham khảo

1. Tập dữ liệu *Walmart Recruiting - Store Sales Forecasting* từ Kaggle.[Ấn vào đây](https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data)
2. Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. [Ấn vào đây](https://otexts.com/fpp3/)
3. Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*. [Ấn vào đây](https://students.aiu.edu/submissions/profiles/resources/onlineBook/G6j8h8_Time_Series_Analysis_Forecasting5.pdf?utm_source=chatgpt.com)
4. Thư viện Python: pandas, numpy, statsmodels, matplotlib.
5. Các bài viết và tài liệu về mô hình ARIMA và dự báo chuỗi thời gian.