Bộ Công Thương

Trường Đại Học Kinh Tế Kỹ Thuật Công Nghiệp

Khoa Khoa Học Ứng Dụng

BÁO CÁO TỔNG KẾT
HỌC PHẦN: ĐỒ ÁN 2
Dự báo doanh số bán hàng Walmart

Thành viên	**Mã sinh viên**
Trần Thanh Hoa	22174600115
Lê Thị Phương Linh	22174600057
Nguyễn Thị Phương Anh	22174600085
Nguyễn Quang Huy	22174600113

(GV hướng dẫn : ThS. Lê Hằng Anh)

Hà Nội, 2025

Bộ Công Thương

Trường Đại Học Kinh Tế Kỹ Thuật Công Nghiệp

Khoa Khoa Học Ứng Dụng

BÁO CÁO TỔNG KẾT
HỌC PHẦN: ĐỒ ÁN 2
Dự báo doanh số bán hàng Walmart

Thành viên	**Mã sinh viên**
Trần Thanh Hoa	22174600115
Lê Thị Phương Linh	22174600057
Nguyễn Thị Phương Anh	22174600085
Nguyễn Quang Huy	22174600113

(GV hướng dẫn : ThS. Lê Hằng Anh)

Hà Nội, 2025

PHIẾU ĐĂNG KÝ ĐỀ TÀI

1. Tên đề tài: Dự báo doanh số bán hàng Walmart

2. Thông tin nhóm sinh viên:

Sinh viên 1 (Nhóm trưởng):

- **Họ và tên:** Trần Thanh Hoa
- **Mã sinh viên:** 22174600115
- **Điện thoại:** 0886679585
- **Email:** Tthoa.dhkl16a1hn@sv.uneti.edu.vn

Sinh viên 2:

- **Họ và tên:** Nguyễn Thị Phương Anh
- **Mã sinh viên:** 22174600085
- **Điện thoại:** 0352 797 453
- **Email:** Ntpanh.dhkl16a1hn@sv.uneti.edu.vn

Sinh viên 3:

- **Họ và tên:** Nguyễn Quang Huy
- **Mã sinh viên:** 22174600113
- **Điện thoại:** 086 560 1815
- **Email:** Nqhuy.dhkl16a1hn@sv.uneti.edu.v

Sinh viên 4:

- **Họ và tên:** Lê Thị Phương Linh
- **Mã sinh viên:** 22174600047
- **Điện thoại:** 0866724363
- **Email:** ltplinh.dhkl16a1hn@sv.uneti.edu.vn

3. Tóm tắt nội dung đề tài: Dự báo doanh số bán hàng là một bài toán quan trọng trong lĩnh vực thương mại và quản lý chuỗi cung ứng, giúp các doanh nghiệp tối ưu hóa kế hoạch sản xuất, phân phối và tồn kho. Đề tài này tập trung vào việc xây dựng mô hình dự đoán doanh số bán hàng cho các cửa hàng của Walmart – một trong những

chuỗi bán lẻ lớn nhất thế giới – dựa trên dữ liệu lịch sử, các sự kiện khuyến mãi và yếu tố mùa vụ. Dữ liệu được sử dụng là bộ "Walmart Sales Forecasting", bao gồm thông tin doanh số bán hàng theo tuần của 45 cửa hàng khác nhau, với chi tiết theo các phòng ban (departments), thời gian khuyến mãi (holiday events), và các yếu tố đặc trưng khác như chỉ số kinh tế và thời tiết. Mục tiêu chính là dự đoán chính xác doanh số tương lai, từ đó hỗ trợ Walmart trong việc ra quyết định chiến lược kinh doanh và quản lý hàng tồn kho.

Việc xử lý và phân tích dữ liệu sẽ bao gồm các bước tiền xử lý, khám phá dữ liệu, xây dựng mô hình học máy (như hồi quy, cây quyết định, hoặc LSTM cho chuỗi thời gian), và đánh giá hiệu quả mô hình dựa trên các chỉ số như RMSE hoặc MAPE. Kết quả từ đề tài không chỉ mang lại giá trị cho Walmart mà còn có thể áp dụng rộng rãi cho các doanh nghiệp bán lẻ khác trong công tác dự báo doanh thu.

Ngày 9 tháng 4 năm 2025

Nhóm trưởng

Hoa

Trần Thanh Hoa

ĐỀ CƯƠNG CHI TIẾT ĐỀ TÀI

1. Tên đề tài: Dự báo doanh số bán hàng Walmart

2. Mục tiêu đề tài: Xây dựng một hệ thống dự báo doanh số bán hàng cho các cửa hàng Walmart dựa trên dữ liệu lịch sử, các sự kiện khuyến mãi và yếu tố mùa vụ. Thông qua việc phân tích và xử lý dữ liệu bán hàng theo tuần từ 45 cửa hàng trên toàn nước Mỹ, đề tài hướng đến việc khám phá các xu hướng, chu kỳ và tác động của các dịp lễ, chương trình khuyến mãi đến doanh thu. Từ đó, áp dụng các mô hình học máy và mô hình chuỗi thời gian để dự đoán doanh số trong tương lai một cách chính xác. Kết quả dự báo sẽ hỗ trợ doanh nghiệp trong việc lập kế hoạch hàng tồn kho, tối ưu hóa nguồn lực và xây dựng chiến lược kinh doanh hiệu quả hơn. Đồng thời, đề tài cũng so sánh hiệu suất giữa các mô hình dự báo để lựa chọn phương pháp tối ưu, góp phần nâng cao khả năng ứng dụng của các thuật toán phân tích dữ liệu trong lĩnh vực bán lẻ.

3. Tổng quan tình hình nghiên cứu thuộc lĩnh vực đề tài: Dự báo doanh số bán hàng là một lĩnh vực nghiên cứu quan trọng trong khoa học dữ liệu và phân tích kinh doanh, thu hút sự quan tâm mạnh mẽ từ giới học thuật cũng như doanh nghiệp. Trong những năm gần đây, với sự phát triển của trí tuệ nhân tạo và học máy, các phương pháp dự báo truyền thống như hồi quy tuyến tính đã dần được bổ sung hoặc thay thế bởi các mô hình hiện đại như Random Forest và mạng nơ-ron sâu (Deep Learning).

Một số nghiên cứu nổi bật có thể kể đến như công trình của Hyndman & Athanasopoulos (2018) với mô hình dự báo chuỗi thời gian phi tuyến, hay nghiên cứu của Bojer và Meldgaard (2020) sử dụng XGBoost và các đặc trưng mùa vụ để dự báo doanh số bán lẻ với độ chính xác cao. Ngoài ra, mạng nơ-ron hồi tiếp (RNN) và đặc biệt là LSTM đã được áp dụng rộng rãi trong bài toán dự báo doanh số theo chuỗi thời gian dài, với khả năng ghi nhớ thông tin lâu dài và xử lý dữ liệu có tính tuần hoàn theo thời gian.

Trong thực tiễn, các công ty lớn như Walmart, Amazon và Target cũng đã đầu tư mạnh mẽ vào hệ thống dự báo doanh thu tự động, tích hợp dữ liệu lớn (Big Data), các chỉ số kinh tế vĩ mô, thời tiết và hành vi tiêu dùng. Ví dụ, Walmart đã từng hợp tác với Kaggle tổ chức cuộc thi "Walmart Recruiting - Store Sales Forecasting", nơi hàng ngàn nhà khoa học dữ liệu xây dựng mô hình dự báo dựa trên dữ liệu thực tế của 45 cửa hàng.

Các hướng nghiên cứu hiện tại không chỉ tập trung vào việc cải thiện độ chính xác của mô hình mà còn chú trọng đến tính giải thích (interpretability), khả năng áp

dụng linh hoạt theo địa phương và khả năng phản ứng nhanh với các yếu tố biến động đột ngột như đại dịch, thiên tai hay sự thay đổi trong xu hướng tiêu dùng. Từ tổng quan này, có thể thấy bài toán dự báo doanh số bán hàng vẫn đang là một lĩnh vực năng động, với tiềm năng phát triển mạnh mẽ và đóng vai trò quan trọng trong việc hỗ trợ ra quyết định kinh doanh chiến lược.

4. Nội dung đề tài: Đề tài "Dự báo doanh số bán hàng" tập trung vào việc nghiên cứu, xây dựng và đánh giá các mô hình dự đoán doanh số bán hàng của hệ thống siêu thị Walmart dựa trên dữ liệu thực tế. Mục tiêu chính là tìm ra mô hình dự báo có độ chính xác cao nhằm hỗ trợ việc lập kế hoạch hàng hóa, quản lý kho bãi và tối ưu hóa hoạt động kinh doanh. Đề tài bao gồm các nội dung chính sau:

4.1. Tìm hiểu và xử lý dữ liệu:

Bộ dữ liệu được sử dụng trong đề tài là "Walmart Sales Forecasting" gồm thông tin doanh số bán hàng theo tuần của 45 cửa hàng và nhiều department khác nhau. Dữ liệu bao gồm các yếu tố liên quan đến thời gian (ngày, tuần, tháng), các dịp lễ (Holiday Events như Lễ Tạ Ơn, Giáng Sinh…), các chỉ số kinh tế (giá xăng, tỷ lệ thất nghiệp, CPI…), và các yếu tố nội tại như khuyến mãi.

Đầu tiên, dữ liệu từ các tệp test.csv,train.csv, stores.csv, features.csv được hợp nhất và xử lý. Các bước tiền xử lý bao gồm chuyển đổi kiểu dữ liệu ngày, mã hóa biến phân loại, xử lý giá trị thiếu, tạo các đặc trưng mới từ ngày như tháng, tuần, thứ, năm,... Đồng thời, dữ liệu được chuẩn hóa nhằm đảm bảo tính nhất quán và phù hợp cho việc huấn luyện mô hình.

4.2. Khám phá dữ liệu và phân tích thống kê:

Trước khi xây dựng mô hình, cần phân tích xu hướng doanh số theo thời gian, mối quan hệ giữa doanh số và các yếu tố ảnh hưởng như thời tiết, dịp lễ hay chương trình khuyến mãi. Việc trực quan hóa dữ liệu bằng các biểu đồ đường, biểu đồ cột, heatmap… giúp hiểu rõ đặc điểm của dữ liệu và định hướng lựa chọn mô hình phù hợp.

4.3. Xây dựng mô hình dự báo:

Nội dung quan trọng nhất của đề tài là thử nghiệm và so sánh nhiều mô hình dự báo khác nhau:

Mô hình truyền thống: Hồi quy tuyến tính (Linear Regression)- phù hợp với chuỗi thời gian có tính quy luật.

Mô hình học máy: Random Forest, Decision Tree, Support Vector Regression (svr), K-Nearest Neighbors (KNN)- có khả năng xử lý dữ liệu phi tuyến tính và nhiều đặc trưng phức tạp.

Mô hình học sâu: LSTM (Long Short-Term Memory) – phù hợp với chuỗi thời gian dài, có khả năng ghi nhớ thông tin lịch sử tốt.

Việc đánh giá mô hình dựa trên các chỉ số như RMSE (Root Mean Square Error), MAE (Mean Absolute Error) hoặc MAPE (Mean Absolute Percentage Error).

4.4. So sánh kết quả và lựa chọn mô hình tối ưu:

Kết quả của các mô hình sẽ được tổng hợp và phân tích để lựa chọn mô hình có độ chính xác cao nhất và phù hợp với yêu cầu thực tiễn. Ngoài ra, yếu tố như thời gian huấn luyện, khả năng mở rộng và tính dễ hiểu của mô hình cũng được cân nhắc.

4.5. Đề xuất và ứng dụng:

Cuối cùng, đề tài sẽ đề xuất một hệ thống dự báo doanh số ứng dụng mô hình tối ưu, có thể tích hợp vào quy trình ra quyết định kinh doanh thực tế tại các hệ thống bán lẻ. Mô hình cũng có thể mở rộng áp dụng cho các ngành hàng khác hoặc các chuỗi cửa hàng tương tự.

Thông qua việc triển khai đề tài, người thực hiện không chỉ rèn luyện kỹ năng xử lý dữ liệu và xây dựng mô hình dự báo, mà còn hiểu rõ hơn về vai trò của khoa học dữ liệu trong hỗ trợ hoạt động kinh doanh hiện đại.

5. Phương pháp thực hiện: Để thực hiện bài toán dự báo doanh số bán hàng, đề tài sẽ tiến hành theo các bước cụ thể như sau:

5.1. Thu thập và xử lý dữ liệu

Sử dụng bộ dữ liệu "Walmart Sales Forecasting" với thông tin doanh số bán hàng theo tuần từ 45 cửa hàng và nhiều department khác nhau. Dữ liệu bao gồm các yếu tố như ngày, sự kiện lễ, khuyến mãi, chỉ số kinh tế, v.v. Tiến hành tiền xử lý dữ liệu bằng cách xử lý giá trị thiếu, loại bỏ ngoại lệ, mã hóa biến phân loại (Label Encoding, One-Hot Encoding), chuẩn hóa dữ liệu và phân chia thành tập huấn luyện và kiểm tra.

5.2. Phân tích và trực quan hóa dữ liệu

Sử dụng các công cụ trực quan hóa như biểu đồ đường, biểu đồ nhiệt (heatmap), biểu đồ phân tán để khám phá xu hướng doanh số, mối quan hệ giữa doanh số và các yếu tố ảnh hưởng như mùa vụ, lễ tết và chương trình khuyến mãi. Vẽ biểu đồ doanh số trung bình theo cửa hàng, department, tháng,...

5.3. Xây dựng mô hình

Áp dụng và so sánh nhiều mô hình dự báo bao gồm:

Mô hình truyền thống: Hồi quy tuyến tính (Linear Regression)

Mô hình học máy: Random Forest, Decision Tree, Support Vector Regression (SVR), K-Nearest Neighbors (KNN)

Mô hình học sâu: LSTM (Long Short-Term Memory)

5.4. Đánh giá mô hình

Sử dụng các chỉ số như MAE (Mean Absolute Error), RMSE (Root Mean Square Error) và MAPE (Mean Absolute Percentage Error), R^2 Score để đo lường độ chính xác của mô hình.

5.5. Kết luận và đề xuất

So sánh hiệu quả các mô hình để chọn ra mô hình tối ưu nhất, đồng thời đưa ra các đề xuất ứng dụng mô hình vào thực tiễn quản lý doanh số tại các hệ thống bán lẻ. Áp dụng mô hình vào tập dữ liệu kiểm định để thử nghiệm khả năng dự báo thực tế, đồng thời đề xuất hướng ứng dụng mô hình vào hệ thống bán lẻ nhằm hỗ trợ ra quyết định trong kinh doanh.

6. Phân công công việc (dự kiến):

STT	**Họ và tên**	**Mã sinh viên**	**Nội dung công việc được phân công**
1	Lê Thị Phương Linh	22174600057	EDA-Phân tích dữ liệu Viết word 4.1: EDA-Phân tích dữ liệu
2	Trần Thanh Hoa	22174600115	Code giải thích mô hình và đưa ra insight. Code triển khai mô hình và tạo dự báo. Làm word: - Phiếu đăng ký đề tài - Đề cương chi tiết đề tài - Phần Mở đầu - Chương 1: Giới thiệu - Chương 2: Cơ sở lý thuyết - Chương 4: Mục 4.2 và 4.3 - Chương 5: Kết luận
3	Nguyễn Quang Huy	22174600113	Code xây dựng và đánh giá mô hình. Viết word Chương 4: Mục 4.2
4	Nguyễn Thị Phương Anh	22174600085	Thu thập và khám phá dữ liệu. Tiền xử lý dữ liệu Viết word chương 3: Thực nghiệm và xử lý dữ liệu Làm PowerPoint

7. Dự kiến kết quả đạt được: Sau quá trình nghiên cứu và thực hiện, đề tài "Dự báo doanh số bán hàng" kỳ vọng đạt được một số kết quả cụ thể như sau:

- Xây dựng được quy trình dự báo doanh số hoàn chỉnh:

Thông qua việc xử lý dữ liệu, phân tích đặc trưng và áp dụng nhiều mô hình khác nhau, đề tài sẽ thiết lập một quy trình chuẩn cho bài toán dự báo doanh số bán hàng. Quy trình này có thể tái sử dụng và mở rộng cho các bộ dữ liệu khác nhau trong các lĩnh vực tương tự như bán lẻ, tiêu dùng nhanh, hoặc thương mại điện tử.

- So sánh và lựa chọn mô hình tối ưu:

Bằng việc áp dụng và đánh giá hiệu suất của nhiều mô hình học máy (Decision Tree, Random Forest, SVR, KNN, Linear Regression) và mô hình học sâu (LSTM), đề tài sẽ chỉ ra mô hình nào phù hợp nhất với đặc điểm dữ liệu chuỗi thời gian trong kinh doanh bán lẻ. Mô hình tối ưu được chọn sẽ có sai số dự báo thấp nhất (thông qua các chỉ số RMSE, MAE, MAPE), đồng thời có khả năng ứng dụng cao trong thực tế.

- Hiểu rõ các yếu tố ảnh hưởng đến doanh số:

Thông qua phân tích dữ liệu, đề tài sẽ xác định được những yếu tố chính ảnh hưởng đến doanh số bán hàng như các dịp lễ, chương trình khuyến mãi, xu hướng mùa vụ, và các biến kinh tế vĩ mô. Từ đó, hỗ trợ các nhà quản lý đưa ra các chiến lược kinh doanh hiệu quả và tối ưu hóa kế hoạch nhập hàng, tồn kho.

- Tạo nền tảng ứng dụng mô hình trong thực tiễn:

Kết quả của đề tài có thể được sử dụng như một module dự báo doanh số trong hệ thống quản lý bán hàng, giúp doanh nghiệp bán lẻ như Walmart nâng cao khả năng lập kế hoạch kinh doanh, phân phối hàng hóa hợp lý, tiết kiệm chi phí và nâng cao doanh thu.

Nhìn chung, đề tài không chỉ giúp đánh giá hiệu quả các mô hình dự báo mà còn mở ra hướng ứng dụng thực tiễn trong lĩnh vực thương mại, góp phần vào việc ra quyết định chiến lược dựa trên dữ liệu.

Ngày 9 tháng 4 năm 2025

Nhóm trưởng

Hoa

Trần Thanh Hoa

MỞ ĐẦU

Trong những năm gần đây, cùng với sự phát triển mạnh mẽ của công nghệ và dữ liệu lớn (Big Data), các doanh nghiệp trên toàn thế giới ngày càng quan tâm đến việc ứng dụng các mô hình dự báo trong quản lý và điều hành hoạt động kinh doanh. Đặc biệt trong lĩnh vực bán lẻ, việc dự đoán chính xác doanh số bán hàng không chỉ giúp tối ưu hóa chuỗi cung ứng, quản lý hàng tồn kho hiệu quả mà còn đóng vai trò quan trọng trong việc ra quyết định chiến lược, gia tăng doanh thu và giảm thiểu rủi ro tài chính. Walmart – tập đoàn bán lẻ hàng đầu thế giới – với mạng lưới rộng lớn và lượng dữ liệu giao dịch khổng lồ, là minh chứng rõ ràng cho nhu cầu dự báo doanh số chính xác. Tuy nhiên, doanh số bị chi phối bởi nhiều yếu tố như thời tiết, lễ tết, khuyến mãi, vị trí cửa hàng, xu hướng tiêu dùng,… Vì vậy, việc xây dựng một hệ thống dự báo thông minh là vô cùng cần thiết.

Trước thực tế đó, em đã thực hiện đề tài "DỰ BÁO DOANH SỐ BÁN HÀNG CỦA WALMART" nhằm tìm hiểu, phân tích và áp dụng các mô hình học máy hiện đại để dự đoán doanh số, từ đó đưa ra các đề xuất mang tính ứng dụng cao cho các doanh nghiệp bán lẻ nói chung và Walmart nói riêng. Đây là một đề tài có tính thực tiễn cao, không chỉ phục vụ cho công việc học tập mà còn có thể ứng dụng hiệu quả trong đời sống và hoạt động kinh doanh thực tế.

Trong quá trình thực hiện đề tài, em đã nhận được rất nhiều sự hướng dẫn tận tình, chỉ bảo quý báu và những góp ý chân thành từ giảng viên Cô Lê Hằng Anh. Em xin gửi lời cảm ơn sâu sắc đến cô vì đã giúp đỡ em trong suốt thời gian nghiên cứu và hoàn thiện đề tài. Mặc dù đã nỗ lực hết mình, nhưng do hạn chế về mặt kiến thức, kinh nghiệm và kỹ năng thực hành, bài làm của em chắc chắn không tránh khỏi những thiếu sót. Em rất mong nhận được những ý kiến đóng góp quý báu để bài nghiên cứu này được hoàn thiện hơn.

Đồ án bao gồm các phần được trình bày theo chương như sau:
Chương 1: Giới thiệu
Chương 2: Cơ sở lý thuyết và mô hình dự báo
Chương 3: Thực nghiệm và xử lý dữ liệu
Chương 4: Kết quả đạt được và đánh giá mô hình
Chương 5: Kết luận

MỤC LỤC

PHIẾU ĐĂNG KÝ ĐỀ TÀI ..... **ii**
ĐỀ CƯƠNG CHI TIẾT ĐỀ TÀI ..... **iv**
MỞ ĐẦU ..... **ix**
CHƯƠNG 1: GIỚI THIỆU **1**
1.1 Bối cảnh và lý do chọn đề tài 1
1.2 Mục tiêu của đề tài 1
1.3 Phạm vi nghiên cứu 2
1.4 Phương pháp nghiên cứu 2
1.5 Ý nghĩa thực tiễn 2
CHƯƠNG 2: CƠ SỞ LÝ THUYẾT **3**
2.1 Tổng quan về dự báo doanh số bán hàng 3
2.2. Tổng quan về Walmart và bài toán đặt ra 3
2.3. Một số khái niệm và kỹ thuật liên quan 4
2.3.1. Học máy (Machine Learning) 4
2.3.2. Các mô hình học máy phổ biến 4
2.3.3. Các chỉ số đánh giá mô hình 7
2.4. Quy trình xây dựng mô hình dự báo 8
CHƯƠNG 3: THỰC NGHIỆM VÀ XỬ LÝ DỮ LIỆU **9**
3.1. Thu thập dữ liệu 9
3.1.1. Đọc dữ liệu 9
3.1.2. Gộp hai bảng features và stores theo cột "Store" 11
3.1.3. Kiểm tra sự xuất hiện lặp lại của các bản ghi theo ngày trong tập dữ liệu 12
3.2. Tiền xử lý dữ liệu 14
3.2.1. Kiểm tra các giá trị bị thiếu 14
3.2.2 Xử lý ngoại lệ, ngoại lai 17
CHƯƠNG 4: KẾT QUẢ ĐẠT ĐƯỢC **19**
4.1. EDA - Phân tích dữ liệu 19
4.1.1. Phân phối doanh số 19
4.1.2. Phân bố doanh số theo thời gian 22
4.1.3. Ảnh hưởng ngày nghỉ lễ IsHoliday ảnh hưởng doanh số thế nào? Markdown có hiệu quả không? Giảm giá có làm doanh số tăng? 29
4.1.4. Tương quan giữa các đặc trưng với doanh số 33
4.2. Xây dựng mô hình 37
4.2.1. Mô hình Máy học Truyền thống 37
4.2.2. Mô hình Deep Learning - Mạng LSTM 39
4.2.3. Time Series models – ARIMA, SARIMA 40
4.2.3.1. Kiểm tra tính dừng 40

4.2.3.2. Huấn luyện và Dự báo 41
4.2.4. Best model 42
4.3. Giải thích mô hình và đưa ra insight 43
4.3.1. Vẽ Partial Dependence.....44
4.4. Triển khai mô hình và tạo dự báo.....46
4.4.1. Dự báo giữ chân khách hàng.....46
4.4.1. Dự báo rủi ro theo cửa hàng.....50
CHƯƠNG 5: KẾT LUẬN 55
TÀI LIỆU THAM KHẢO 57

MỤC LỤC HÌNH VẼ

Hình 4.2 : Hiệu suất bán hàng của từng loại cửa hàng vào dịp lễ 20
Hình 4. 3 Tổng doanh thu theo năm 23
Hình 4.4 Tổng doanh thu hàng tháng 24
Hình 4. 5: Doanh số hàng tháng theo từng năm 26
Hình 4. 6 : Doanh số trung bình hàng tuần theo các năm 28
Hình 4. 7 : Doanh số trung bình : Tuần lễ với Tuần thường 30
Hình 4. 9 : Markdown 4 và Weekly Sales 32
Hình 4.10 : Ma trận tương quan giữa các đặc trưng với doanh số 33
Hình 4.11: Ảnh hưởng của giá xăng và nhiệt độ với doanh số 35
Hình 4.12 : So sánh RMSE và MAE của các mô hình truyền thống 38
Hình 4.13: Lịch sử huấn luyện mô hình LSTM 39
Hình 4.14: Kiểm tra tính dừng của chuỗi thời gian Weekly_Sales tổng hợp hàng tuần 40
Hình 4.15: So sánh dự báo ARIMA, SARIMA và dữ liệu thực tế Grid Search 41
Hình 4.16: So sánh hiệu suất tổng thể của các mô hình (RMSE, R^2 và MAE) 42
Hình 4.17: Biểu đồ top 15 đặc trưng quan trọng nhất 43
Hình 4.18: Biểu đồ PDP - Ảnh hưởng của Dept, Size và Store đến dự đoán 44

CHƯƠNG 1: GIỚI THIỆU

1.1 Bối cảnh và lý do chọn đề tài

Trong thời đại công nghệ 4.0, dữ liệu được xem là nguồn tài nguyên quý giá đối với mọi doanh nghiệp. Việc thu thập, phân tích và khai thác dữ liệu một cách hiệu quả không chỉ giúp doanh nghiệp hiểu rõ hơn về khách hàng mà còn hỗ trợ ra quyết định chính xác, kịp thời và mang tính chiến lược. Một trong những ứng dụng nổi bật của dữ liệu trong hoạt động kinh doanh là dự báo doanh số bán hàng.

Đối với các doanh nghiệp bán lẻ, đặc biệt là các tập đoàn lớn như Walmart, việc dự báo doanh số là yếu tố then chốt giúp quản lý tốt chuỗi cung ứng, giảm thiểu hàng tồn kho, tối ưu hóa nguồn lực và tăng lợi nhuận. Với hệ thống hàng nghìn cửa hàng trải dài khắp nước Mỹ và hàng triệu giao dịch mỗi ngày, Walmart sở hữu một kho dữ liệu khổng lồ, mở ra cơ hội lớn để áp dụng các phương pháp phân tích và học máy nhằm dự báo xu hướng tiêu dùng và lên kế hoạch kinh doanh hiệu quả hơn.

Tuy nhiên, doanh số bán hàng chịu ảnh hưởng từ nhiều yếu tố như thời tiết, mùa vụ, ngày lễ, chương trình khuyến mãi, địa lý từng khu vực, hành vi người tiêu dùng,... Do đó, xây dựng một mô hình dự báo chính xác là một bài toán phức tạp, đòi hỏi sự kết hợp giữa hiểu biết thống kê, kỹ thuật xử lý dữ liệu và khả năng lựa chọn mô hình học máy phù hợp.

Xuất phát từ thực tế trên, em đã chọn đề tài "Dự báo doanh số bán hàng của Walmart" nhằm tìm hiểu quy trình dự báo, lựa chọn và đánh giá các mô hình học máy phù hợp, từ đó góp phần hỗ trợ công tác ra quyết định kinh doanh trong lĩnh vực bán lẻ.

1.2 Mục tiêu của đề tài

- Tìm hiểu và phân tích dữ liệu bán hàng của Walmart.
- Tiền xử lý dữ liệu và trích xuất các đặc trưng liên quan.
- Áp dụng các mô hình học máy để dự báo doanh số bán hàng.
- So sánh hiệu quả giữa các mô hình và đánh giá độ chính xác.
- Đề xuất mô hình dự báo tối ưu phục vụ công tác quản lý doanh số.

1.3 Phạm vi nghiên cứu

- Đề tài tập trung vào tập dữ liệu doanh số bán hàng của Walmart được công bố công khai, bao gồm các thông tin về doanh số theo từng cửa hàng, từng tuần, từng bộ phận (department), cùng với các yếu tố ảnh hưởng như thời tiết, ngày lễ, v.v.
- Các mô hình dự báo được áp dụng bao gồm: Hồi quy tuyến tính, Random Forest, và các mô hình học máy phổ biến khác.
- Phạm vi thời gian của dữ liệu được giới hạn theo tập dữ liệu gốc (thường từ năm 2010 đến năm 2012).

1.4 Phương pháp nghiên cứu

- Phân tích mô tả dữ liệu để hiểu đặc trưng dữ liệu và các xu hướng chính.
- Tiền xử lý dữ liệu: xử lý dữ liệu thiếu, chuẩn hóa, mã hóa biến phân loại.
- Chia tập dữ liệu thành tập huấn luyện và kiểm thử.
- Áp dụng các mô hình học máy và đánh giá độ chính xác bằng các chỉ số như MAE (Mean Absolute Error), RMSE (Root Mean Square Error) và MAPE (Mean Absolute Percentage Error), R^2 so sánh kết quả giữa các mô hình để chọn ra mô hình tốt nhất.

1.5 Ý nghĩa thực tiễn

Việc xây dựng mô hình dự báo doanh số bán hàng có ý nghĩa rất lớn trong việc:

- Giúp các nhà quản lý tại Walmart nói riêng và các doanh nghiệp bán lẻ nói chung lập kế hoạch kinh doanh hiệu quả hơn.
- Tối ưu hóa nguồn hàng, giảm thiểu chi phí vận hành.
- Góp phần nâng cao chất lượng phục vụ khách hàng thông qua việc đảm bảo đủ nguồn cung đúng thời điểm.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1 Tổng quan về dự báo doanh số bán hàng

Dự báo doanh số bán hàng là quá trình sử dụng dữ liệu lịch sử và các yếu tố liên quan để ước lượng lượng hàng hóa hoặc dịch vụ mà một doanh nghiệp sẽ bán ra trong tương lai. Đây là công cụ quan trọng giúp doanh nghiệp:

- Lập kế hoạch tồn kho, chuỗi cung ứng.
- Điều chỉnh chiến lược tiếp thị, khuyến mãi.
- Ra quyết định kinh doanh kịp thời và chính xác.

Các kỹ thuật dự báo truyền thống thường dựa trên thống kê (như trung bình động, hồi quy tuyến tính), trong khi các phương pháp hiện đại tận dụng sức mạnh của học máy để khai thác dữ liệu lớn và mối quan hệ phi tuyến giữa các biến[1].

2.2. Tổng quan về Walmart và bài toán đặt ra

Walmart là một trong những chuỗi bán lẻ lớn nhất thế giới, sở hữu hàng nghìn siêu thị tại Hoa Kỳ và các quốc gia khác. Với dữ liệu bán hàng theo từng tuần, từng cửa hàng và từng bộ phận sản phẩm, bài toán đặt ra là:

Xây dựng một mô hình dự báo doanh số bán hàng tương lai của từng department tại mỗi cửa hàng.

Yêu cầu của bài toán gồm:

- Dự đoán chính xác doanh số bán hàng.
- Cân nhắc đến yếu tố ngày lễ, thời tiết và xu hướng theo mùa.
- So sánh hiệu suất các mô hình khác nhau.

[1] *Trần Minh Triết (2021),* Khai phá dữ liệu (Data Mining)*, NXB Đại học Quốc gia TP.HCM, TP. Hồ Chí Minh.*

2.3. Một số khái niệm và kỹ thuật liên quan

2.3.1. Học máy (Machine Learning)

Học máy là một nhánh của trí tuệ nhân tạo, cho phép hệ thống học từ dữ liệu và cải thiện hiệu suất theo thời gian mà không cần được lập trình rõ ràng. Trong bài toán dự báo doanh số bán hàng của Walmart, học máy được sử dụng để:

- Xây dựng các mô hình dự đoán doanh số dựa vào dữ liệu lịch sử.
- Tự động nhận biết xu hướng, mùa vụ, dịp lễ, tác động của thời tiết và các đặc điểm từng cửa hàng.
- Giảm thiểu sai số trong dự đoán, hỗ trợ quản lý hàng tồn và kế hoạch bán hàng.

2.3.2. Các mô hình học máy phổ biến

❖ Linear Regression (Hồi quy tuyến tính):

Mục đích:

- Dự đoán giá trị doanh số (biến liên tục) dựa trên các yếu tố đầu vào như thời gian, ngày lễ, thời tiết, cửa hàng, bộ phận (department),...

Công thức: [2]

$$Y = \beta_0 + \beta_{1x1} + \beta_{2x2} + \ldots + \beta_{nxn} + \varepsilon$$

Trong đó:

- Y: Doanh số dự báo (sales)
- x1,x2,...,xn: Các đặc trưng (biến đầu vào như tuần, nhiệt độ, sự kiện lễ hội,…)
- $\beta_0 + \beta_1 + \beta_n$: Hệ số hồi quy
- ε: Sai số ngẫu nhiên

Ứng dụng:

- Xây dựng mô hình cơ bản làm chuẩn so sánh.

[2] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R*(2nd ed.). Springer.

- Hiệu quả khi dữ liệu có mối quan hệ tuyến tính giữa các yếu tố.

❖ Random Forest:

Mục đích:

- Dự đoán doanh số với độ chính xác cao hơn bằng cách kết hợp nhiều cây quyết định (Decision Trees).

Nguyên lý:

- Tạo nhiều cây quyết định trên các tập dữ liệu con (subsets).
- Trung bình kết quả dự đoán của các cây để giảm sai số và tránh overfitting.

Dự báo:[3]

$$Y = \frac{1}{T} \sum_{i=1}^{T} h_i(x)$$

Trong đó:

- T: số lượng cây
- $h_i(x)$: dự báo từ cây thứ i**i**

Ứng dụng:

- Tốt với dữ liệu có tính phi tuyến, có nhiều đặc trưng.
- Tự động xử lý tương tác giữa các biến.

❖ Support Vector Regression (SVR):

Mục đích:

- Tìm đường hồi quy sao cho sai số nhỏ hơn một ngưỡng εε, đồng thời mô hình càng đơn giản càng tốt.

[3] *Breiman, L. (2001). Random Forests.* Machine Learning*, 45(1), 5–32.*

Công thức:

- Tối thiểu hóa hàm mục tiêu:[4]

$$min\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi i+\xi_i^*$$

Với ràng buộc:

$$\begin{cases} y_i - w_{xi}^T - b \le \varepsilon + \xi i \\ w_{xi}^T + b - y_i \le \varepsilon + \xi_i^* \\ \xi i, \xi_i^* \ge 0 \end{cases}$$

Trong đó:

- w: vector trọng số
- ξ , ξ_i^*: biến slack cho sai số vượt ngưỡng
- C: hệ số điều chỉnh giữa độ chính xác và độ phức tạp

Ứng dụng:

- Phù hợp cho dữ liệu phi tuyến nhẹ hoặc nhiễu ít.
- Có thể kết hợp với kernel để mở rộng phi tuyến.

❖ KNN – K-Nearest Neighbors Regression

Mục đích:

- KNN là một thuật toán dựa trên khoảng cách, dùng để dự đoán giá trị đầu ra của một điểm dữ liệu mới bằng cách tham chiếu đến K điểm dữ liệu gần nhất trong tập huấn luyện.

Trong dự báo doanh số, KNN sẽ tìm các tuần, cửa hàng, department trong quá khứ có đặc điểm gần giống tuần cần dự đoán và lấy trung bình doanh số của các điểm này để dự báo.

[4] *Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression.* Statistics and Computing*, 14(3), 199–222.*

Cách hoạt động:

1. Tính khoảng cách giữa điểm cần dự đoán và tất cả các điểm trong tập huấn luyện (thường dùng khoảng cách Euclid).
2. Chọn K điểm gần nhất.
3. Dự đoán giá trị đầu ra bằng cách lấy trung bình doanh số của K điểm đó.

Công thức toán học:

Khoảng cách Euclid:[5]

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n} (x_j - x_{ij})^2}$$

Dự báo:

$$\hat{y} = \frac{1}{K} \sum_{i=1}^{K} y_i$$

Trong đó:

- x: điểm dữ liệu cần dự đoán
- x_i: các điểm lân cận (K điểm gần nhất)
- y_i: giá trị đầu ra (doanh số thực tế) của các điểm lân cận

2.3.3. Các chỉ số đánh giá mô hình

❖ MAE (Mean Absolute Error): Sai số tuyệt đối trung bình.[6]

Công thức:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y_i}|$$

❖ RMSE (Root Mean Squared Error): Căn bậc hai của sai số bình phương trung bình.

[5] *James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021).* An Introduction to Statistical Learning: with Applications in R*(2nd ed.). Springer.*

[6] *Géron, A. (2019).* Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow *(2nd ed.). O'Reilly Media.*

Công thức:

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$$

❖ R^2 (R-squared): Hệ số xác định, đánh giá mức độ phù hợp của mô hình với dữ liệu.

Công thức:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}$$

2.4. Quy trình xây dựng mô hình dự báo

1. Khám phá và phân tích dữ liệu (EDA): Phân tích xu hướng, mùa vụ, yếu tố ảnh hưởng.
2. Tiền xử lý dữ liệu: Chuẩn hóa dữ liệu, xử lý giá trị thiếu, mã hóa biến phân loại.
3. Chia dữ liệu huấn luyện và kiểm thử.
4. Huấn luyện mô hình học máy.
5. Đánh giá mô hình: Sử dụng MAE, RMSE, R^2.
6. Triển khai mô hình tốt nhất để dự báo doanh số tương lai.

CHƯƠNG 3: THỰC NGHIỆM VÀ XỬ LÝ DỮ LIỆU

3.1. Thu thập dữ liệu

3.1.1. Đọc dữ liệu

Stores:

	Store	Type	Size
0	1	A	151315
1	2	A	202307
2	3	B	37392
3	4	A	205863

Train:

	Store	Dept	Date	Weekly_Sales	IsHoliday
0	1	1	2/5/2010	24924.5	FALSE
1	1	1	2/12/2010	46039.49	TRUE
2	1	1	2/19/2010	41595.55	FALSE
3	1	1	2/26/2010	19403.54	FALSE
4	1	1	3/5/2010	21827.9	FALSE

Features:

	Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2
0	1	2/5/2010	42.31	2.572	NaN	NaN
1	1	2/12/2010	38.51	2.548	NaN	NaN
2	1	2/19/2010	39.93	2.514	NaN	NaN
3	1	2/26/2010	46.63	2.561	NaN	NaN
4	1	3/5/2010	46.5	2.625	NaN	NaN

	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHolida y
0	NaN	NaN	NaN	211.096358	8.106	FALSE
1	NaN	NaN	NaN	211.24217	8.106	TRUE
2	NaN	NaN	NaN	211.289143	8.106	FALSE
3	NaN	NaN	NaN	211.319643	8.106	FALSE
4	NaN	NaN	NaN	211.350143	8.106	FALSE

Test:

	Store	Dept	Date	IsHoliday
0	1	1	11/2/2012	FALSE
1	1	1	11/9/2012	FALSE
2	1	1	11/16/2012	FALSE
3	1	1	11/23/2012	TRUE
4	1	1	11/30/2012	FALSE

Số lượng mã cửa hàng duy nhất trong stores_df: 45

Số lượng mã cửa hàng duy nhất trong features_df: 45

- Tổng quan bộ dữ liệu:

train.csv – Dữ liệu huấn luyện

Chứa thông tin lịch sử bán hàng:

Store: Mã số của cửa hàng

Dept: Mã số của phòng ban (department)

Date: Ngày cụ thể

Weekly_Sales: Doanh số bán hàng trong tuần (giá trị mục tiêu)

IsHoliday: Tuần có kỳ nghỉ lễ hay không (True/False)

features.csv – Đặc trưng bổ sung theo ngày và cửa hàng

Store: Mã cửa hàng

Date: Ngày

Temperature: Nhiệt độ trung bình (theo vùng của cửa hàng)

Fuel_Price: Giá xăng tại khu vực đó

MarkDown1-5: Giá trị khuyến mãi thuộc 5 chiến dịch khác nhau

CPI: Chỉ số giá tiêu dùng

Unemployment: Tỷ lệ thất nghiệp

IsHoliday: Tuần lễ có nghỉ lễ không (trùng tên với file train.csv)

stores.csv – Thông tin tĩnh về từng cửa hàng

Store: Mã cửa hàng

Type: Loại cửa hàng

Size: Diện tích cửa hàng (số feet vuông)

test.csv – Dữ liệu kiểm tra (không có nhãn)

Có cấu trúc tương tự train.csv nhưng không chứa Weekly_Sales.

Dùng để tạo ra dự đoán cuối cùng.

3.1.2. Gộp hai bảng features và stores theo cột "Store"

	Store	Dept	Date	Weekly_Sales	IsHoliday_x	Temperature	Fuel_Price
0	1	1	2/5/2010	24924.5	FALSE	42.31	2.572
1	1	1	2/12/2010	46039.49	TRUE	38.51	2.548
2	1	1	2/19/2010	41595.55	FALSE	39.93	2.514
3	1	1	2/26/2010	19403.54	FALSE	46.63	2.561
4	1	1	3/5/2010	21827.9	FALSE	46.5	2.625

	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5
0	NaN	NaN	NaN	NaN	NaN
1	NaN	NaN	NaN	NaN	NaN
2	NaN	NaN	NaN	NaN	NaN
3	NaN	NaN	NaN	NaN	NaN
4	NaN	NaN	NaN	NaN	NaN

	CPI	IsHoliday_y	Unemployment	Type	Size
0	211.096358	8.106	FALSE	A	151315
1	211.24217	8.106	TRUE	A	151315
2	211.289143	8.106	FALSE	A	151315
3	211.319643	8.106	FALSE	A	151315
4	211.350143	8.106	FALSE	A	151315

Kích thước của bảng features_df: (8190, 12)

Kích thước của bảng train: (421570, 5)

Số lượng cửa hàng duy nhất trong bảng training_df: 45

Khoảng thời gian trong bảng train_df: từ 1/13/2012 đến 9/9/2011

3.1.3. Kiểm tra sự xuất hiện lặp lại của các bản ghi theo ngày trong tập dữ liệu

Để xác định xem có nhiều dòng dữ liệu được ghi nhận cho cùng một ngày hay không, tiến hành kiểm tra số lượng bản ghi có cùng giá trị trong cột Date. Dưới đây là kết quả dữ liệu thể hiện các dòng có cùng ngày 1/13/2012, nhưng khác nhau về Store và Dept, cho thấy rằng hệ thống ghi nhận nhiều giao dịch khác nhau tại các cửa hàng và bộ phận khác nhau trong cùng một ngày.

	Store	Dept	Date	Weekly_Sales	IsHoliday
101	1	1	1/13/2012	16894.4	FALSE
244	1	2	1/13/2012	43353.09	FALSE
387	1	3	1/13/2012	13822.49	FALSE
530	1	4	1/13/2012	36582.36	FALSE
673	1	5	1/13/2012	19281.61	FALSE
...	...	...	...	...	...
420971	45	93	1/13/2012	1642.06	FALSE
421105	45	94	1/13/2012	3743.16	FALSE
421248	45	95	1/13/2012	50721.08	FALSE
421393	45	97	1/13/2012	6433.64	FALSE
421528	45	98	1/13/2012	677.71	FALSE

Dựa vào kết quả kiểm tra, thấy rằng cùng một ngày xuất hiện trong nhiều dòng khác nhau. Điều này là hợp lý vì mỗi dòng trong tập dữ liệu đại diện cho doanh thu của một bộ phận (Dept) tại một cửa hàng (Store) cụ thể trong một tuần nhất định (Date). Như vậy, với một ngày, hệ thống sẽ ghi nhận nhiều dòng dữ liệu tương ứng với từng tổ

	Store	Dept	Date	Weekly_Sales	IsHoliday_x	Temperature	Fuel_Price	Unemployment
0	1	1	2/5/2010	24924.5	FALSE	42.31	2.572	FALSE
1	1	1	2/12/2010	46039.49	TRUE	38.51	2.548	TRUE
2	1	1	2/19/2010	41595.55	FALSE	39.93	2.514	FALSE
3	1	1	2/26/2010	19403.54	FALSE	46.63	2.561	FALSE
4	1	1	3/5/2010	21827.9	FALSE	46.5	2.625	FALSE

Hợp Store – Dept

	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	IsHoliday_y	Type	Size
0	NaN	NaN	NaN	NaN	NaN	211.096358	8.106	A	151315
1	NaN	NaN	NaN	NaN	NaN	211.24217	8.106	A	151315
2	NaN	NaN	NaN	NaN	NaN	211.289143	8.106	A	151315
3	NaN	NaN	NaN	NaN	NaN	211.319643	8.106	A	151315
4	NaN	NaN	NaN	NaN	NaN	211.350143	8.106	A	151315

	Store	Dept	Date	Weekly_Sales	IsHoliday_x	Temperature	Fuel_Price
0	1	1	2/5/2010	24924.5	FALSE	42.31	2.572
1	1	1	2/12/2010	46039.49	TRUE	38.51	2.548
2	1	1	2/19/2010	41595.55	FALSE	39.93	2.514
3	1	1	2/26/2010	19403.54	FALSE	46.63	2.561
4	1	1	3/5/2010	21827.9	FALSE	46.5	2.625

3.2. Tiền xử lý dữ liệu

3.2.1. Kiểm tra các giá trị bị thiếu

Store	**0**
Dept	0
Date	0
Weekly_Sales	0
IsHoliday_x	0
Temperature	0
Fuel_Price	0
MarkDown1	270889
MarkDown2	310322
MarkDown3	284479
MarkDown4	286603
MarkDown5	270138
CPI	0
Unemployment	0
IsHoliday_y	0
Type	0
Size	0

Các biến chủ chốt (Store, Dept, Date, Weekly_Sales, IsHoliday, Temperature, Fuel_Price, CPI, Unemployment, Type, Size) đều không có giá trị thiếu, đảm bảo tính đầy đủ cho phân tích doanh thu và điều kiện kinh tế – xã hội.

Các cột MarkDown1 – MarkDown5 có số lượng missing rất lớn (từ ~270 000 đến ~ 310 000 bản ghi).

Sau khi xử lý, xóa một cột "IsHoliday" không cần thiết ta được kết quả:

- Thống kê mô tả

	Store	Dept	Date	Weekly_Sales	index	Temperature	Fuel_Price
count	421570	421570	421570	421570	421570	421570	421570
mean	22.200546	44.260317	30:32.0	15981.25812	3929.692801	60.090059	3.361027
min	1	1	2/5/2010 0:00	-4988.94	0	-2.06	2.472
25%	11	18	10/8/2010 0:00	2079.65	1927	46.68	2.933
50%	22	37	6/17/2011 0:00	7612.03	3875	62.09	3.452
75%	33	74	2/24/2012 0:00	20205.8525	5873	74.28	3.738
max	45	99	10/26/2012 0:00	693099.36	8150	100.14	4.468
std	12.785297	30.492054	NaN	22711.18352	2327.429021	18.447931	0.458515

	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	Size
count	421570	421570	421570	421570	421570	421570	421570	421570

mean	2590.074819	879.974298	468.087665	1083.132268	1662.772385	171.201947	7.960289	136727.9157
min	0	-265.76	-29.1	0	0	126.064	3.879	34875
25%	0	0	0	0	0	132.022667	6.891	93638
50%	0	0	0	0	0	182.31878	7.866	140167
75%	2809.05	2.2	4.54	425.29	2168.04	212.416993	8.572	202505
max	88646.76	104519.54	141630.61	67474.85	108519.28	227.232807	14.313	219622
std	6052.385934	5084.538801	5528.873453	3894.529945	4207.629321	39.159276	1.863296	60980.58333

Nhận xét:

Sau khi điền giá trị NaN cho các cột Markdown bằng 0 và hợp nhất dữ liệu, chúng ta có 421.570 bản ghi đầy đủ cho các cột sau. Dưới đây là các nhận xét chính về phân phối và đặc điểm của các biến:

Store & Dept

Cả hai cột là chỉ số ID, không phải biến liên tục.

Giá trị trải đều từ 1 đến 45 cho Store và 1 đến 99 cho Dept, cho thấy đủ bao phủ các cửa hàng và bộ phận.

Weekly_Sales

Mean ≈ 15.981, Std ≈ 22.711 → phân phối rất phân tán.

Min = −4.988 (xuất hiện âm có thể do điều chỉnh trả hàng hoặc lỗi ghi nhận).

Max = 693.099 → tồn tại nhiều outlier doanh thu cao bất thường.

Median = 7.612, 75% = 20.206 → phân phối nghiêng phải, đa số giao dịch ở mức thấp hơn trung vị 7.612.

Temperature (°F)

Mean ≈ 60.09, Std ≈ 18.45 → dao động từ −2.06 đến 100.14.

25–75% nằm trong [46.68; 74.28] thể hiện sự khác biệt rõ theo mùa.

Fuel_Price (USD/gal)

Mean ≈ 3.361, Std ≈ 0.459 → biến động tương đối nhỏ.

Giá thấp nhất 2.472, cao nhất 4.468, phù hợp xu hướng giá xăng tại Mỹ giai đoạn 2010–2012.

MarkDown1–MarkDown5

Median và 25% đều = 0 → hơn một nửa các tuần không có chiến dịch giảm giá.

Mean ở MarkDown1 ≈ 2.590, MarkDown5 ≈ 1.663, các MarkDown khác thấp hơn.

Max rất lớn (MarkDown2 up to 104.520) và có cả giá trị âm (ví dụ MarkDown2 min = −265.76), cho thấy khả năng lỗi dữ liệu hoặc ghi nhận lại điều chỉnh giá.

Cần kiểm tra và loại bỏ/điều chỉnh các giá trị âm nếu không hợp lý.

CPI & Unemployment

CPI: Mean ≈ 171.20, Std ≈ 39.16, dao động từ 126.06 đến 227.23.

Unemployment: Mean ≈ 7.96%, Std ≈ 1.86%, dao động từ 3.88% đến 14.31%.

Các biến này ổn định hơn, phù hợp với bối cảnh kinh tế 2010–2012.

Size

Mean ≈ 136.728, Std ≈ 60.981, Min = 34.875, Max = 219.622 (đơn vị là kích thước cửa hàng).

Phân phối trải rộng, cần cân nhắc chuẩn hóa hoặc log-transform khi đưa vào mô hình.

3.2.2 *Xử lý ngoại lệ, ngoại lai*
Số lượng ngoại lệ: 35381

Tỷ lệ ngoại lệ: 8.42%

Từ kết quả trên, đưa ra nhận xét sau: Tỷ lệ ngoại lệ gần 8.5% (35 381 trên 421 570 bản ghi) cho thấy rằng không chỉ có một vài điểm bất thường lẻ tẻ, mà chúng chiếm một tỉ trọng đáng kể trong tập dữ liệu.

Thay thế các ngoại lệ bằng giá trị trung vị của 'Weekly_Sales', kết quả thu được:

Số lượng ngoại lệ sau khi thay thế: 0

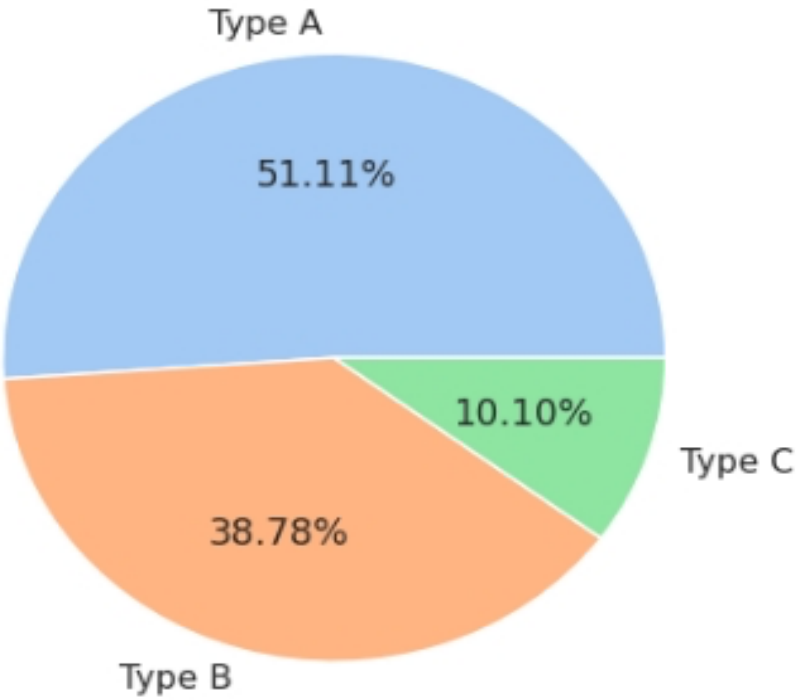
CHƯƠNG 4: KẾT QUẢ ĐẠT ĐƯỢC

4.1. EDA - Phân tích dữ liệu

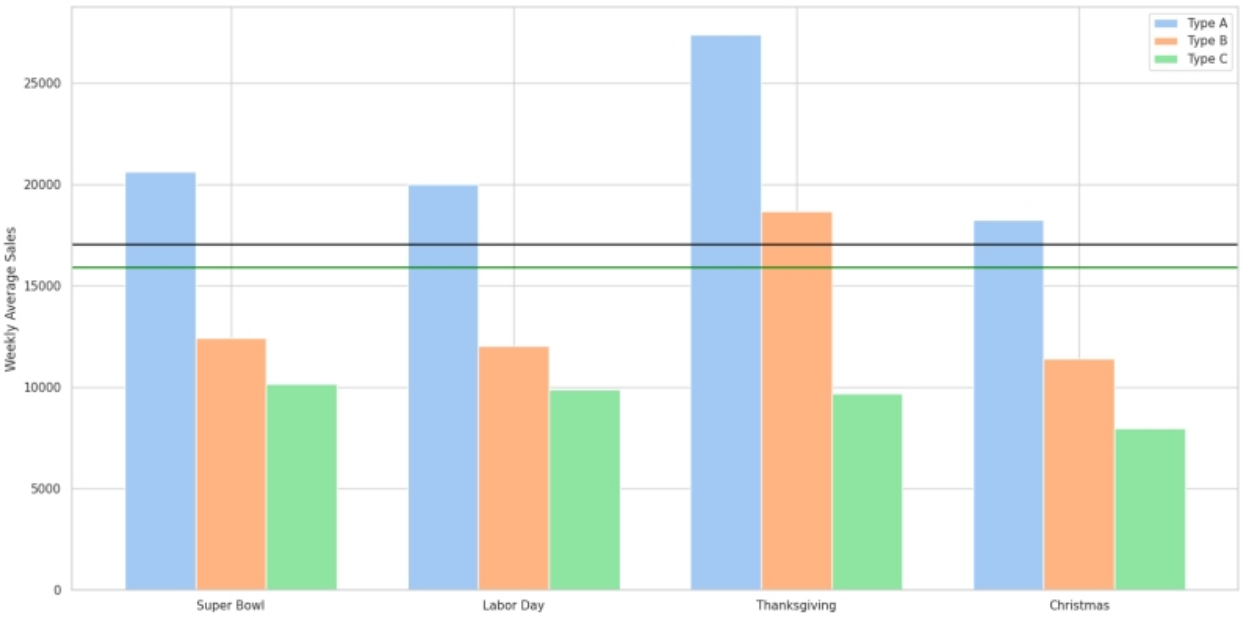
4.1.1. Phân phối doanh số

Ta có 3 size cửa hàng là:

- Loại A 11655.363907
- Loại B 9410.009499
- Loại C 6833.327180



Hình 4.1 : Biểu đồ tròn Doanh số với loại cửa hàng



Hình 4.2 : Hiệu suất bán hàng của từng loại cửa hàng vào dịp lễ

Hiệu suất bán hàng của từng loại cửa hàng vào dịp lễ

- Tính nhất quán về thứ hạng doanh số: Một quan sát nổi bật là sự duy trì nhất quán trong thứ hạng doanh số giữa ba loại cửa hàng qua tất cả bốn dịp lễ. Cụ thể, Type A luôn ghi nhận doanh số trung bình hàng tuần cao nhất, theo sau là Type B với mức doanh số trung bình, và Type C có doanh số trung bình thấp nhất. Điều này cho thấy có những yếu tố cơ bản, có thể liên quan đến mô hình kinh doanh, quy mô hoạt động, hoặc vị trí chiến lược, tạo ra sự khác biệt bền vững trong hiệu suất bán hàng giữa các loại cửa hàng này.
- Ảnh hưởng đáng kể của dịp lễ Thanksgiving: Dịp lễ Thanksgiving dường như có tác động đặc biệt lớn đến doanh số của cả ba loại cửa hàng, nhưng đáng chú ý nhất là đối với Type A, khi doanh số trung bình hàng tuần tăng vọt lên mức cao nhất so với bất kỳ dịp lễ nào khác. Type B cũng cho thấy sự tăng trưởng doanh số đáng kể trong dịp Thanksgiving. Ngược lại, Type C chỉ ghi nhận một sự tăng nhẹ, cho thấy mức độ ảnh hưởng của Thanksgiving đến loại cửa hàng này là không đáng kể so với Type A và Type B.
- Sự suy giảm doanh số sau Thanksgiving: Sau đỉnh điểm doanh số vào dịp Thanksgiving, doanh số trung bình hàng tuần của cả ba loại cửa hàng đều có xu hướng giảm vào dịp Christmas. Tuy nhiên, mức độ suy giảm này khác nhau giữa các loại:

- Type A vẫn duy trì được mức doanh số tương đối cao vào Christmas, mặc dù không bằng Thanksgiving. Type B chứng kiến sự sụt giảm đáng kể, quay trở lại mức doanh số tương đương hoặc thậm chí thấp hơn so với Super Bowl và Labor Day. Type C có mức giảm mạnh nhất, đạt doanh số thấp nhất trong tất cả các dịp lễ vào dịp Christmas. Hiệu suất tương đối ổn định trong Super Bowl và Labor Day: So với sự biến động rõ rệt trong Thanksgiving và Christmas, doanh số trung bình hàng tuần của cả ba loại cửa hàng có vẻ tương đối ổn định hơn trong hai dịp lễ Super Bowl và Labor Day. Sự khác biệt về doanh số giữa các loại cửa hàng vẫn duy trì tương tự như xu hướng chung, với Type A dẫn đầu, theo sau là Type B và Type C.

- So sánh với các mức trung bình:

Hai đường ngang màu đen và xanh lá cây trên biểu đồ là Trung bình doanh số bán hàng trong các tuần lễ có kỳ nghỉ và không có kỳ nghỉ lễ.

- Doanh số trung bình hàng tuần của Type A thường xuyên vượt qua cả hai mức trung bình này trong tất cả các dịp lễ, cho thấy hiệu suất vượt trội của loại cửa hàng này so với mức chuẩn. Type B dao động quanh các mức trung bình, đôi khi vượt qua (như vào dịp Thanksgiving) và đôi khi thấp hơn (như vào dịp Christmas). Type C luôn có doanh số trung bình hàng tuần nằm dưới cả hai mức trung bình, cho thấy hiệu suất kém hơn so với mức chuẩn chung.

- Những phát hiện từ biểu đồ này gợi ý rằng các yếu tố phân biệt giữa Type A, Type B và Type C (có thể bao gồm quy mô cửa hàng, vị trí, chiến lược sản phẩm, hoạt động marketing, hoặc đặc điểm khách hàng mục tiêu) có tác động sâu sắc đến khả năng tạo ra doanh số, đặc biệt là trong các dịp lễ lớn.

- Type A dường như đã xây dựng được một mô hình kinh doanh mạnh mẽ, có khả năng tận dụng tối đa các cơ hội mua sắm cao điểm. Type B cho thấy sự nhạy cảm hơn với từng dịp lễ cụ thể, với hiệu suất tăng đáng kể vào Thanksgiving nhưng lại giảm vào Christmas. Type C có thể đang gặp thách thức trong việc thu hút khách hàng và tạo ra doanh số, đặc biệt là trong bối cảnh cạnh tranh gia tăng vào các dịp lễ.

- Các cửa hàng có kích thước lớn hơn (Type A) nhất quán ghi nhận doanh số trung bình hàng tuần cao hơn đáng kể so với các cửa hàng có kích thước trung bình (Type B) và nhỏ (Type C) trong tất cả các dịp lễ được xem xét.

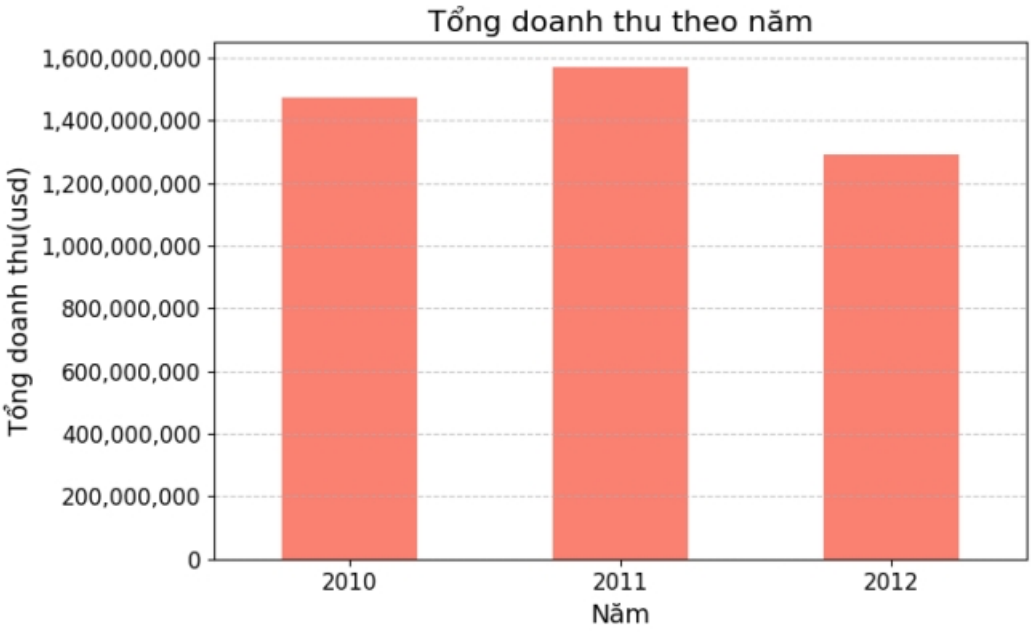
→ Điều này củng cố giả thuyết cho rằng kích thước cửa hàng có thể là một yếu tố quan trọng ảnh hưởng đến doanh số. Các cửa hàng lớn hơn, với lợi thế về không gian trưng bày rộng rãi, khả năng chứa nhiều hàng hóa và phục vụ đồng thời lượng lớn khách hàng hơn, có tiềm năng cao hơn để triển khai các chương trình khuyến mãi quy mô lớn, trưng bày sản phẩm hấp dẫn và đa dạng, từ đó thu hút và đáp ứng nhu cầu mua sắm đa dạng của khách hàng, đặc biệt là trong các dịp lễ mua sắm cao điểm.

Tuy nhiên, cần lưu ý rằng kích thước cửa hàng không phải là yếu tố duy nhất quyết định doanh số. Mô hình hoạt động tổng thể của từng loại cửa hàng (ví dụ: chiến lược marketing, quản lý hàng tồn kho, chất lượng dịch vụ khách hàng) và vị trí địa lý chiến lược cũng đóng vai trò then chốt trong việc thu hút khách hàng và tối đa hóa doanh thu. Các cửa hàng lớn (Type A) có thể đồng thời được hưởng lợi từ vị trí đắc địa, khả năng đầu tư vào các chương trình khuyến mãi quy mô lớn và một mô hình hoạt động hiệu quả, tất cả cùng nhau tạo nên hiệu suất bán hàng vượt trội.

Ngược lại, các cửa hàng có kích thước nhỏ hơn (Type C) có thể bị hạn chế về không gian trưng bày và khả năng phục vụ, đồng thời có thể không có đủ nguồn lực để triển khai các chương trình khuyến mãi lớn, dẫn đến doanh số thấp hơn. Tuy nhiên, điều này không loại trừ khả năng các cửa hàng nhỏ hơn có thể thành công nhờ vào một mô hình kinh doanh , tập trung vào một phân khúc khách hàng cụ thể hoặc một vị trí địa lý đặc biệt thuận lợi cho phân khúc đó.

Tóm lại, biểu đồ này cung cấp bằng chứng trực quan cho thấy có mối tương quan giữa kích thước cửa hàng (được đại diện bởi Type A, B, C) và doanh số bán hàng, đặc biệt trong các dịp lễ. Tuy nhiên, để có một kết luận toàn diện và sâu sắc hơn, cần xem xét thêm các yếu tố khác như mô hình hoạt động và vị trí địa lý của từng loại cửa hàng.

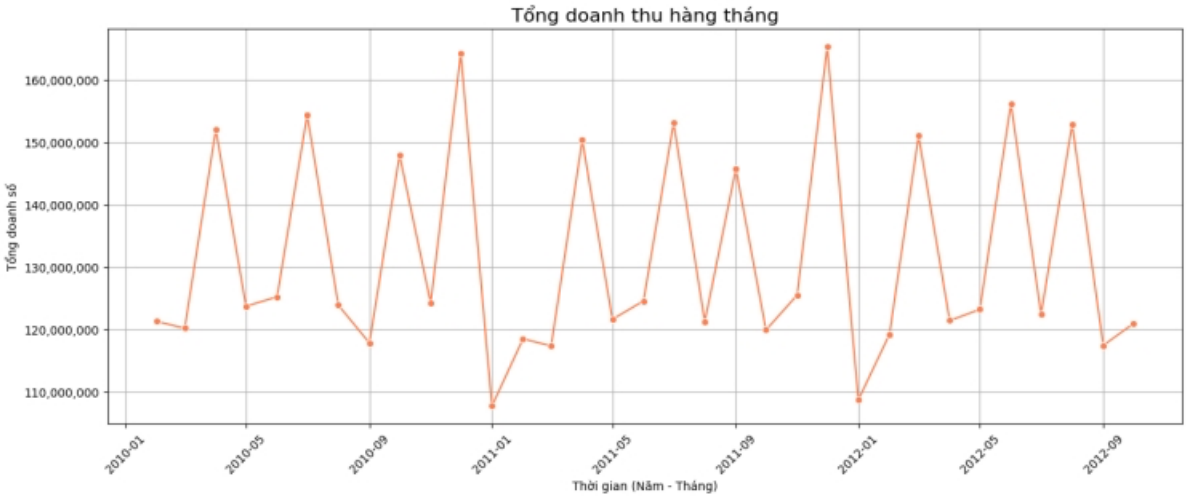
4.1.2. Phân bố doanh số theo thời gian



Hình 4.3 Tổng doanh thu theo năm

- Năm 2011 đạt doanh thu cao nhất: Tổng doanh thu đạt đỉnh vào năm 2011 với giá trị khoảng 1,58 tỷ USD (1,580,000,000 USD). Đây là năm có hiệu suất bán hàng tốt nhất trong giai đoạn khảo sát. Năm 2010 có doanh thu cao thứ hai: Năm 2010 ghi nhận tổng doanh thu khoảng 1,48 tỷ USD (1,480,000,000 USD), thấp hơn một chút so với năm 2011. Năm 2012 có doanh thu giảm: Tổng doanh thu giảm đáng kể vào năm 2012, xuống còn khoảng 1,29 tỷ USD (1,290,000,000 USD). Đây là năm có hiệu suất bán hàng thấp nhất trong 3 năm được khảo sát

→ Tóm lại: Trong giai đoạn từ năm 2010 đến 2012, bộ dữ liệu cho thấy một xu hướng tăng trưởng doanh thu từ năm 2010 đến năm 2011, sau đó lại giảm sút vào năm 2012. Năm 2011 là năm có tổng doanh thu cao nhất, trong khi năm 2012 chứng kiến sự sụt giảm đáng kể so với hai năm trước đó.



Hình 4.4 Tổng doanh thu hàng tháng

Doanh thu tăng vọt vào khoảng tháng 11 - 12 mỗi năm (cuối năm).Ngược lại, doanh thu giảm mạnh vào tháng 1 hàng năm (đầu năm).Các tháng giữa năm (khoảng từ tháng 4 đến tháng 9) thường có mức doanh thu dao động quanh mức trung bình, không quá cao cũng không quá thấp.Sự chênh lệch giữa các tháng có thể lên tới 50 triệu USD hoặc thậm chí hơn giữa tháng thấp điểm và tháng cao điểm.

Ví dụ:

Tháng 12/2010: doanh thu đạt đỉnh hơn 160 triệu USD.

Tháng 1/2011: doanh thu rơi xuống mức thấp chỉ còn khoảng 110 triệu USD.

Điều này cho thấy chênh lệch doanh thu tháng cao và thấp có thể trên 40-50%.

- Giải thích nguyên nhân sự chênh lệch doanh thu theo tháng:

Hiệu ứng mùa vụ:

Tháng 11 - 12 trùng với mùa lễ hội lớn như Black Friday, Giáng Sinh (Christmas), và New Year.

Đây là thời điểm khách hàng có xu hướng chi tiêu mạnh tay, mua sắm quà tặng, thực phẩm, đồ trang trí…, dẫn tới doanh thu tăng vọt.

Thời kỳ thấp điểm:

Tháng 1 thường là giai đoạn hậu lễ hội, người tiêu dùng đã chi tiêu nhiều vào cuối năm trước nên có xu hướng thắt chặt chi tiêu.Ngoài ra, tháng 1 có thể trùng với kỳ nghỉ đông, nhiều gia đình đi du lịch hoặc ít mua sắm hơn. Chính sách bán hàng: Các chương trình khuyến mãi lớn thường tập trung vào cuối năm, ít hơn vào đầu năm.

Thời tiết: Thời tiết mùa đông ở Mỹ (thị trường gốc của Walmart) cũng ảnh hưởng đến tần suất đi lại mua sắm.

→ Tổng doanh thu hàng tháng có sự chênh lệch rõ rệt theo mùa vụ, với đỉnh cao vào cuối năm và thấp điểm vào đầu năm.Sự chênh lệch này hoàn toàn phù hợp với quy luật tiêu dùng trong thực tế, đặc biệt tại các thị trường lớn như Mỹ. Khi phân tích dữ liệu doanh số hoặc dự báo doanh thu, cần đặc biệt lưu ý đến yếu tố mùa vụ để xây dựng chiến lược kinh doanh và quản lý tồn kho phù hợp. Tết dương lịch (New Year's Day) chỉ là ngày nghỉ lễ ngắn, người ta không mua sắm nhiều dịp này. Noël (Christmas) mới là mùa mua sắm "khủng" ➔ tập trung vào tháng 11 - 12. Sau đó, tháng 1, mọi người nghỉ ngơi, tiết kiệm tiền, rất ít mua sắm.

Biểu đồ đường này cho thấy một mô hình mùa vụ rõ rệt trong tổng doanh thu hàng tháng của Walmart tại Mỹ từ tháng 1 năm 2010 đến tháng 9 năm 2012. Chu kỳ này được đặc trưng bởi các đỉnh doanh thu vào cuối năm (tháng 12) và mùa xuân (tháng 3 hoặc 4), xen kẽ với các đáy doanh thu vào giữa năm (tháng 6 hoặc tháng 7) và đầu năm (tháng 1 và tháng 2).

Đỉnh doanh thu lớn nhất trùng khớp với mùa mua sắm lễ hội cuối năm, đạt đỉnh vào tháng 12 do ảnh hưởng mạnh mẽ của lễ Giáng Sinh. Nhu cầu mua sắm quà tặng, đồ trang trí, thực phẩm và hàng hóa tiêu dùng liên quan tăng cao trong giai đoạn này, thúc đẩy doanh số vượt trội.

Một đỉnh doanh thu thứ hai xuất hiện vào khoảng tháng 3 hoặc tháng 4, có thể được lý giải bởi sự kết hợp của nhiều yếu tố đặc trưng tại Mỹ. Mùa Xuân mang đến sự thay đổi trong thời tiết, khuyến khích người tiêu dùng mua sắm các mặt hàng phục vụ cho gia đình và các hoạt động ngoài trời. Đồng thời, lễ Phục Sinh, một dịp lễ quan trọng khác, thường rơi vào thời điểm này, thúc đẩy doanh số các mặt hàng thực phẩm, đồ trang trí và quà tặng. Ngoài ra, Walmart có thể triển khai các chương trình khuyến mãi đầu mùa để kích cầu mua sắm.

Ngược lại, doanh thu thường giảm xuống mức thấp nhất vào giữa năm (tháng 6 hoặc tháng 7), đánh dấu mùa thấp điểm mua sắm sau các đợt lễ lớn và trước khi mùa tựu trường bắt đầu. Tương tự, doanh thu cũng có xu hướng thấp vào đầu năm (tháng 1 và tháng 2). Hiện tượng này có thể được giải thích bởi một số yếu tố sau hậu mùa mua sắm lễ hội cuối năm. Sau khi chi tiêu đáng kể cho Lễ Tạ Ơn và Giáng Sinh, người tiêu dùng thường điều chỉnh lại ngân sách và thói quen chi tiêu, dẫn đến nhu cầu mua sắm giảm.

Thêm vào đó, các chương trình khuyến mãi lớn thường kết thúc sau mùa lễ, làm giảm động lực mua sắm. Thời tiết mùa đông ở nhiều vùng của Mỹ trong tháng 1 và tháng 2 cũng có thể ảnh hưởng đến lưu lượng khách hàng. Cuối cùng, giai đoạn đầu năm thường thiếu vắng các ngày lễ mua sắm lớn so với cuối năm và mùa xuân, và chu kỳ chi tiêu cá nhân của nhiều người có thể đang trong giai đoạn phục hồi sau các khoản chi lớn vào dịp lễ.

Sự lặp lại theo chu kỳ hàng năm của các đỉnh và đáy doanh thu khẳng định tính mùa vụ ổn định trong hoạt động kinh doanh của Walmart tại Mỹ, chịu ảnh hưởng sâu sắc bởi các yếu tố văn hóa, lễ hội và thói quen mua sắm đặc trưng của người tiêu dùng Mỹ.

Hình 4.5: Doanh số hàng tháng theo từng năm

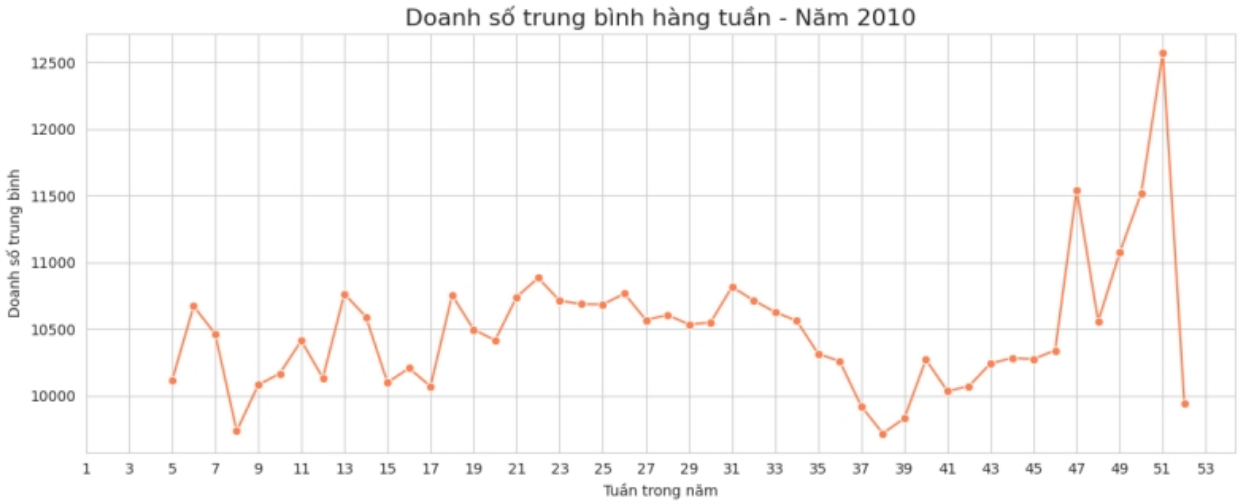
Tính mùa vụ rõ rệt: Cả ba năm đều cho thấy một mô hình doanh số theo mùa tương tự, với doanh số tăng lên vào cuối năm. Sự khác biệt giữa các năm: Có sự khác biệt nhất định về mức doanh số trung bình hàng tuần giữa các năm tại cùng một thời điểm. Đỉnh doanh số cuối năm: Tháng 12 luôn là tháng có doanh số hàng tuần trung bình cao nhất trong cả ba năm. Đáy doanh số đầu năm: Tháng 1 thường là tháng có doanh số hàng tuần trung bình thấp nhất hoặc gần thấp nhất trong cả ba năm.

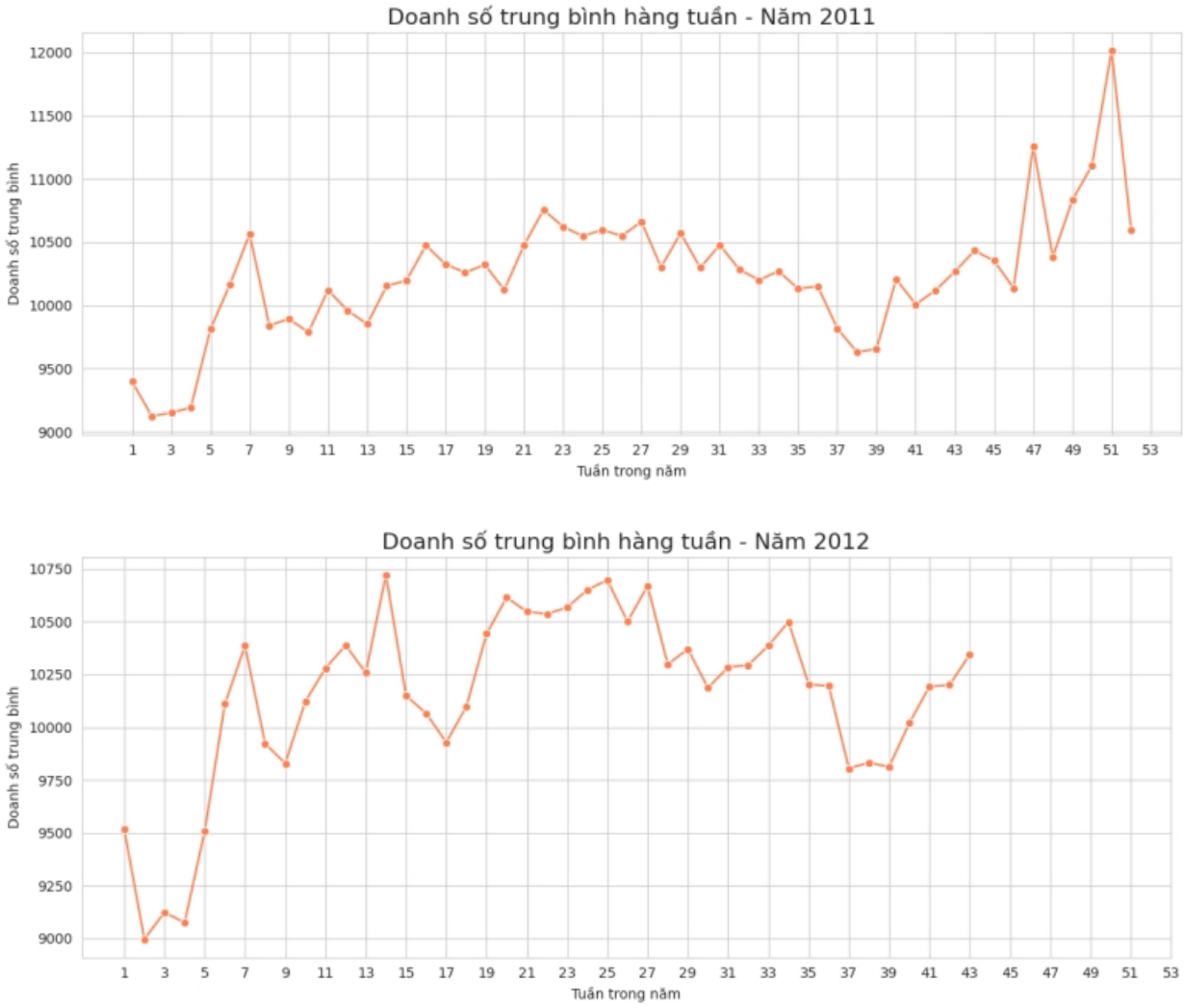
Năm 2010 (xanh lam nhạt): Doanh số thấp nhất vào tháng 1. Có sự tăng trưởng đáng kể từ tháng 1 đến tháng 2. Duy trì mức tương đối ổn định từ tháng 2 đến tháng 5.

Tăng nhẹ vào tháng 6. Giảm nhẹ vào tháng 7. Tăng trở lại vào tháng 8 và duy trì đến tháng 11. Đạt đỉnh cao nhất vào tháng 12. Năm 2011 (cam nhạt): Doanh số thấp nhất vào tháng 1. Tăng trưởng mạnh từ tháng 1 đến tháng 3. Giảm nhẹ vào tháng 4. Tăng trở lại vào tháng 5 và đạt đỉnh thứ hai trong năm vào tháng 6. Giảm đáng kể từ tháng 6 đến tháng 9, đạt đáy thứ hai trong năm vào tháng 9. Tăng trưởng mạnh mẽ từ tháng 9 đến tháng 12, đạt đỉnh cao nhất vào tháng 12. Năm 2012 (xanh lá cây nhạt): Doanh số thấp nhất vào tháng 1. Tăng trưởng ổn định từ tháng 1 đến tháng 6, đạt đỉnh thứ hai trong năm vào tháng 6. Giảm nhẹ vào tháng 7. Tiếp tục giảm vào tháng 8 và đạt đáy thứ hai trong năm vào tháng 9. Tăng trưởng mạnh mẽ từ tháng 9 đến tháng 12, đạt đỉnh cao nhất vào tháng 12. So sánh giữa các năm:

Mức doanh số chung: Nhìn chung, mức doanh số trung bình hàng tuần có vẻ tương đương nhau giữa các năm, mặc dù có những biến động nhỏ theo từng tháng. Thời điểm đỉnh và đáy: Thời điểm đạt đỉnh (tháng 12) và đáy (tháng 1) doanh số khá nhất quán qua các năm, củng cố tính mùa vụ. Biến động giữa năm: Có một số khác biệt về biến động doanh số giữa các tháng trong năm giữa các năm. Ví dụ, năm 2011 có sự sụt giảm đáng kể từ tháng 6 đến tháng 9, trong khi năm 2010 và 2012 không có sự sụt giảm mạnh như vậy. Kết luận:

Biểu đồ cho thấy rõ ràng tính mùa vụ mạnh mẽ trong doanh số hàng tuần trung bình, với tháng 12 là tháng bán hàng tốt nhất và tháng 1 là tháng bán hàng chậm nhất. Mặc dù có những khác biệt nhỏ giữa các năm về mức doanh số và biến động giữa các tháng, mô hình mùa vụ chung vẫn được duy trì. Điều này cho thấy các yếu tố theo mùa (ví dụ: các ngày lễ cuối năm) có ảnh hưởng rất lớn đến doanh số bán hang.





Hình 4. 6 : Doanh số trung bình hàng tuần theo các năm

Cả ba biểu đồ doanh số trung bình hàng tuần cho các năm 2010, 2011 và 2012 đều cho thấy một mô hình mùa vụ rõ rệt và nhất quán, cho thấy doanh số bán hàng chịu ảnh hưởng mạnh mẽ bởi các yếu tố thời gian trong năm. Đỉnh Doanh Số Cuối Năm (Tháng 12): Đây là đặc điểm nổi bật nhất, khi doanh số hàng tuần đều đạt mức cao nhất vào tuần cuối cùng của tháng 12 trong cả ba năm. Điều này phản ánh rõ ràng tác động của mùa mua sắm lễ hội cuối năm, đặc biệt là lễ Giáng Sinh, lên doanh số bán hàng. Đáy Doanh Số Đầu Năm (Tháng 1): Ngược lại, tháng 1 thường ghi nhận mức doanh số trung bình hàng tuần thấp nhất hoặc gần thấp nhất trong cả ba năm. Điều này có thể là do sự giảm chi tiêu sau mùa lễ hội và thiếu vắng các sự kiện mua sắm lớn vào đầu năm. Biến Động Giữa Năm (Tháng 2 - Tháng 10): Trong khoảng thời gian này, doanh số có sự biến động lên xuống, cho thấy ảnh hưởng của các yếu tố khác như các chương trình khuyến mãi nhỏ lẻ, các ngày lễ không lớn, và các yếu tố theo mùa khác (ví dụ: mùa tựu trường). Tuy nhiên, các biến động này thường không có quy luật rõ ràng và có sự khác biệt giữa các năm.

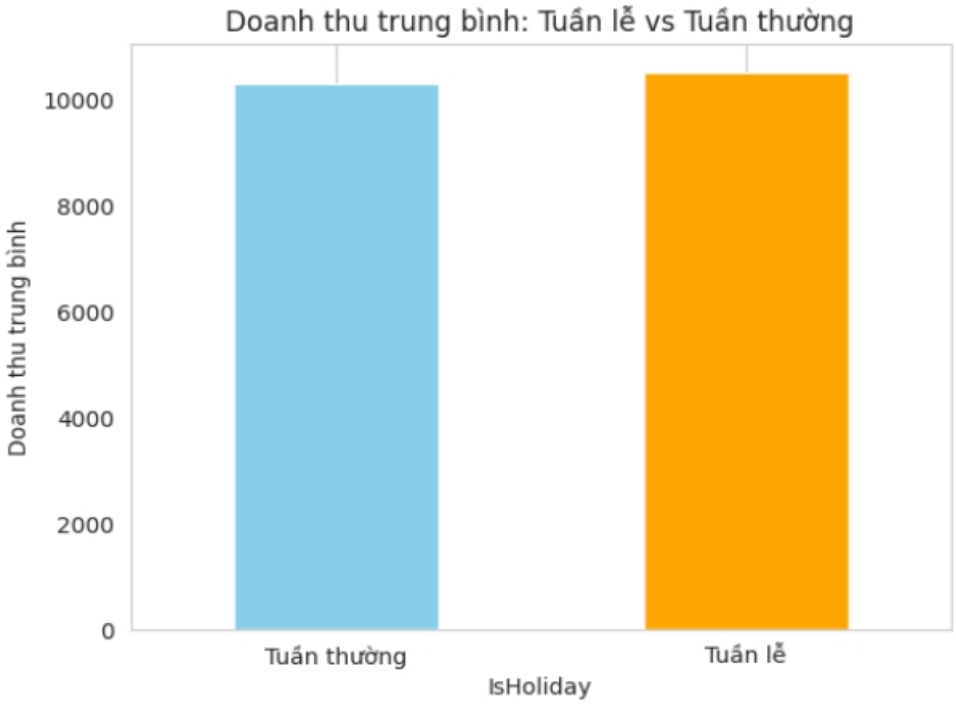
Mặc dù mô hình mùa vụ chung được duy trì, có những khác biệt đáng chú ý về cách mùa vụ ảnh hưởng đến doanh số qua từng năm:

Mức Đỉnh và Đáy: Mức doanh số tuyệt đối tại các đỉnh (tháng 12) và đáy (tháng 1) có sự thay đổi nhẹ qua các năm. Đỉnh doanh số cuối năm có xu hướng giảm dần từ năm 2010 (cao nhất) đến năm 2012 (thấp nhất), cho thấy có thể có sự thay đổi trong sức mua hoặc hiệu quả của mùa mua sắm lễ hội qua thời gian. Tương tự, mức đáy đầu năm cũng có sự biến động nhẹ. Biến Động Giữa Năm: Sự biến động doanh số trong các tháng giữa năm có sự khác biệt rõ rệt giữa các năm. Năm 2010 có vẻ ổn định hơn, trong khi năm 2011 và 2012 cho thấy nhiều dao động lên xuống hơn, có thể phản ánh sự khác biệt trong các chiến lược khuyến mãi hoặc các yếu tố thị trường tạm thời. Thời Điểm và Cường Độ Tăng Trưởng: Thời điểm bắt đầu và tốc độ tăng trưởng doanh số hướng tới đỉnh cuối năm có thể khác nhau giữa các năm. Ví dụ, năm 2011 có sự tăng trưởng mạnh mẽ từ khá sớm (cuối quý 3), trong khi năm 2012 sự tăng trưởng có vẻ tập trung hơn vào những tuần cuối cùng. Kết luận về Ảnh Hưởng của Mùa Vụ Qua Các Năm:

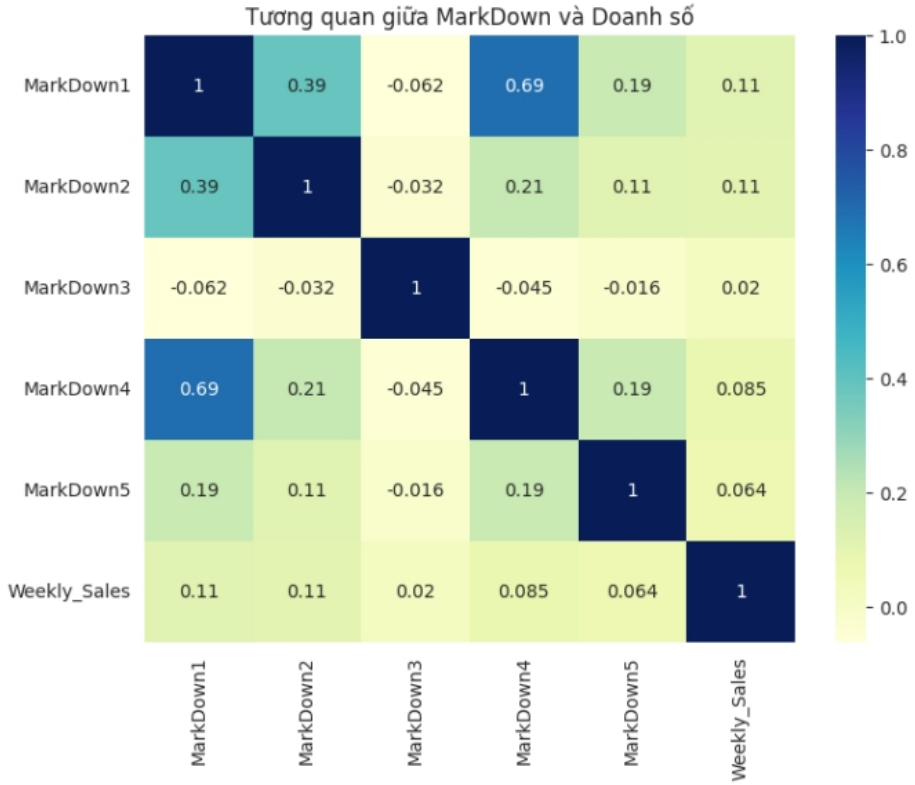
Tính mùa vụ có một ảnh hưởng mạnh mẽ và ổn định lên doanh số bán hàng hàng tuần trong cả ba năm được khảo sát, với mùa lễ hội cuối năm là động lực chính và đầu năm là giai đoạn thấp điểm. Tuy nhiên, cường độ và diễn biến chi tiết của ảnh hưởng mùa vụ có sự thay đổi nhẹ qua các năm.

Sự suy giảm dần ở mức đỉnh doanh số cuối năm (từ 2010 đến 2012) có thể gợi ý về những thay đổi trong xu hướng tiêu dùng hoặc hiệu quả của các chương trình kích cầu vào dịp lễ. Sự khác biệt trong biến động doanh số giữa năm cho thấy các yếu tố không theo mùa có thể có tác động khác nhau qua từng năm.

4.1.3. Ảnh hưởng ngày nghỉ lễ IsHoliday ảnh hưởng doanh số thế nào? Markdown có hiệu quả không? Giảm giá có làm doanh số tăng?



Hình 4.7 : Doanh số trung bình : Tuần lễ với Tuần thường



Hình 4. 8: Tương quan giữa Markdown và doanh số

- MarkDown4 có tương quan dương mạnh nhất với Weekly_Sales: Hệ số tương quan giữa "MarkDown4" và "Weekly_Sales" là 0.69. Đây là giá trị dương lớn nhất, cho thấy chương trình khuyến mại được thể hiện ở cột MarkDown4 có ảnh hưởng tích cực và

mạnh mẽ nhất đến doanh số hàng tuần. Khi có chương trình MarkDown4, doanh số bán hàng có xu hướng tăng lên đáng kể.

- MarkDown1 có tương quan dương đáng kể với Weekly_Sales: Hệ số tương quan giữa "MarkDown1" và "Weekly_Sales" là 0.11. Đây là một tương quan dương yếu, cho thấy chương trình khuyến mại MarkDown1 có ảnh hưởng tích cực nhưng không mạnh mẽ đến doanh số hàng tuần.

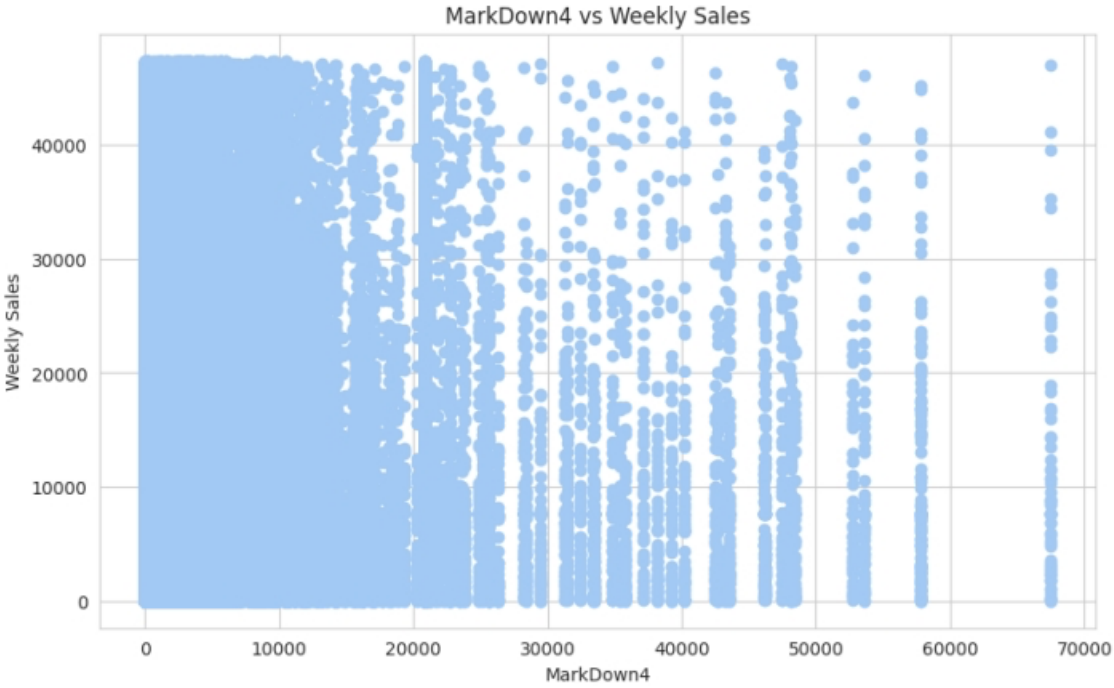
- MarkDown2 và MarkDown5 có tương quan dương rất yếu với Weekly_Sales: Hệ số tương quan giữa "MarkDown2" và "Weekly_Sales" là 0.11, tương tự MarkDown1. Hệ số tương quan giữa "MarkDown5" và "Weekly_Sales" là 0.064. Cả hai đều cho thấy ảnh hưởng tích cực nhưng rất nhỏ đến doanh số hàng tuần. Có thể các chương trình khuyến mại này không đủ hấp dẫn hoặc không được triển khai hiệu quả để tạo ra sự tăng trưởng đáng kể trong doanh số.

- MarkDown3 có tương quan âm rất yếu với Weekly_Sales: Hệ số tương quan giữa "MarkDown3" và "Weekly_Sales" là -0.062. Đây là một tương quan âm rất yếu, cho thấy chương trình khuyến mại MarkDown3 có xu hướng tác động tiêu cực rất nhỏ đến doanh số hàng tuần. Điều này có thể là do chương trình này không phù hợp với thị hiếu khách hàng, áp dụng cho các sản phẩm có nhu cầu thấp, hoặc trùng với các yếu tố tiêu cực khác ảnh hưởng đến doanh số. Tuy nhiên, mức độ ảnh hưởng này là không đáng kể.

- Tương quan giữa các chương trình khuyến mại (MarkDown):

Có một số tương quan dương giữa các chương trình khuyến mại với nhau, đáng chú ý nhất là giữa MarkDown1 và MarkDown4 (0.69), cho thấy có khả năng hai loại khuyến mại này thường được triển khai cùng nhau hoặc có liên quan đến nhau. Các tương quan khác giữa các cột MarkDown thường yếu hơn.

Kết luận: Chương trình khuyến mại MarkDown4 có khả năng là động lực tăng trưởng doanh số mạnh mẽ nhất.. MarkDown1, MarkDown2 và MarkDown5 có tác động tích cực nhưng tương đối nhỏ đến doanh số. Cần xem xét lại hiệu quả của các chương trình này, có thể cần điều chỉnh hoặc thay thế để tăng cường tác động. MarkDown3 có tương quan âm yếu với doanh số, chương trình này có thể không hiệu quả và cần được đánh giá lại.



Hình 4. 9 : Markdown 4 và Weekly Sales

Phân tích cho thấy tuần có lễ (IsHoliday = True) thường có doanh số trung bình cao hơn, phản ánh xu hướng tiêu dùng tăng vào các dịp lễ. Ngoài ra, một số chương trình giảm giá (Markdown) có tương quan dương với doanh thu, đặc biệt là MarkDown1 và MarkDown4. Điều này chứng tỏ giảm giá có thể tác động tích cực đến doanh thu,

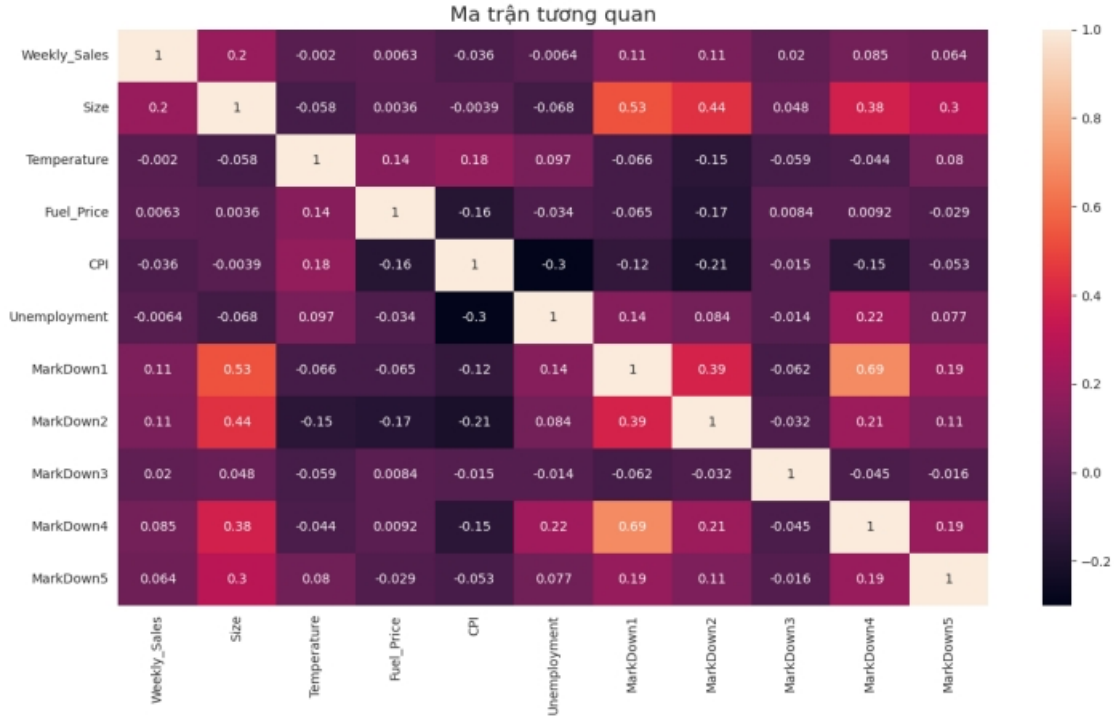
Nhận xét về ảnh hưởng của MarkDown4 đến Weekly_Sales:

- Dựa trên biểu đồ phân tán, chúng ta có thể thấy một số điểm đáng chú ý, củng cố nhận xét về tương quan dương mạnh từ ma trận tương quan trước đó:
    - Xu hướng tăng doanh số khi có MarkDown4: Mặc dù các điểm phân tán rộng, có một xu hướng chung cho thấy khi giá trị của MarkDown4 tăng lên (tức là mức độ khuyến mại lớn hơn), doanh số hàng tuần cũng có xu hướng tăng lên. Điều này đặc biệt rõ ràng khi so sánh các điểm ở vùng MarkDown4 gần 0 với các vùng có giá trị MarkDown4 cao hơn.
- Sự phân tán lớn cho thấy nhiều yếu tố khác ảnh hưởng: Tuy nhiên, sự phân tán rộng của các điểm cho thấy MarkDown4 không phải là yếu tố duy nhất, hoặc thậm chí là yếu tố quyết định hoàn toàn doanh số hàng tuần. Với cùng một mức độ MarkDown4, doanh số hàng tuần có thể dao động trong một phạm vi rất lớn. Điều này ngụ ý rằng có nhiều yếu tố khác như loại sản phẩm, vị trí cửa hàng, thời điểm trong năm, các chương trình khuyến mại khác, và các yếu tố bên ngoài khác cũng đóng vai trò quan trọng trong việc quyết định doanh số.

- Tác động rõ rệt khi MarkDown4 khác 0: Khi MarkDown4 có giá trị khác 0 (tức là có áp dụng chương trình khuyến mại 4), chúng ta thường thấy các điểm dữ liệu trải dài trên một phạm vi doanh số rộng hơn và có xu hướng đạt các giá trị doanh số cao hơn so với khi MarkDown4 bằng 0 (nơi doanh số tập trung ở mức thấp hơn).

- Biểu đồ này, kết hợp với hệ số tương quan cao, mang lại những ý nghĩa quan trọng cho việc quản lý và tối ưu hóa chiến lược khuyến mại: MarkDown4 là một công cụ hiệu quả để thúc đẩy doanh số: Việc triển khai chương trình khuyến mại MarkDown4 có khả năng cao dẫn đến sự tăng trưởng đáng kể trong doanh số hàng tuần. Cần tối ưu hóa mức độ khuyến mại: Do sự phân tán lớn, việc xác định mức độ khuyến mại tối ưu cho MarkDown4 là rất quan trọng. Mức khuyến mại quá thấp có thể không tạo ra tác động đáng kể, trong khi mức quá cao có thể làm giảm lợi nhuận.

4.1.4. Tương quan giữa các đặc trưng với doanh số



Hình 4.10 : Ma trận tương quan giữa các đặc trưng với doanh số

- Với Weekly_Sales (Doanh số hàng tuần):

- Size (Kích thước): Có tương quan dương yếu (0.2) với doanh số hàng tuần. Điều này gợi ý rằng các cửa hàng có kích thước lớn hơn có xu hướng có doanh số cao hơn, nhưng mối quan hệ này không quá mạnh.

- MarkDown1, +MarkDown2, MarkDown4, MarkDown5: Có tương quan dương rất yếu (0.11, 0.11, 0.085, 0.064) với doanh số hàng tuần. Điều này cho thấy các chương trình giảm giá có xu hướng làm tăng doanh số, nhưng tác động riêng lẻ của từng chương trình là nhỏ. MarkDown3: Có tương quan dương rất yếu (0.02) với doanh số hàng tuần, gần như không có tương quan tuyến tính.
- Temperature (Nhiệt độ), Fuel_Price (Giá nhiên liệu), CPI (Chỉ số giá tiêu dùng), Unemployment (Tỷ lệ thất nghiệp): Có tương quan rất yếu, gần bằng không hoặc âm nhẹ với doanh số hàng tuần. Điều này cho thấy các yếu tố kinh tế vĩ mô và thời tiết dường như không có mối quan hệ tuyến tính mạnh mẽ trực tiếp với doanh số hàng tuần. Mối tương quan giữa các đặc trưng độc lập:

- Tương quan dương đáng chú ý:

- Size và các MarkDown (đặc biệt MarkDown1: 0.53, MarkDown2: 0.44, MarkDown4: 0.38, MarkDown5: 0.3): Các cửa hàng lớn hơn có xu hướng triển khai các chương trình giảm giá mạnh mẽ hơn. Temperature và Fuel_Price (0.14): Có một tương quan dương nhẹ, có thể do nhu cầu đi lại tăng khi thời tiết ấm hơn, dẫn đến giá nhiên liệu tăng. Temperature và CPI (0.18): Tương quan dương nhẹ, có thể do nhu cầu tiêu dùng tăng theo nhiệt độ, đẩy CPI lên.
- Temperature và Unemployment (-0.3): Tương quan âm nhẹ, có thể do khi kinh tế tốt (thời tiết ấm áp có thể kích thích một số ngành), tỷ lệ thất nghiệp giảm.
- CPI và Unemployment (-0.3): Tương quan âm nhẹ, có thể do khi lạm phát cao, người dân có xu hướng tìm kiếm việc làm nhiều hơn.
- Các MarkDown có tương quan dương với nhau: Đặc biệt MarkDown1 và MarkDown2 (0.39), MarkDown1 và MarkDown4 (0.69), MarkDown2 và MarkDown4 (0.21) cho thấy các chương trình giảm giá có thể được triển khai đồng thời hoặc có liên quan.

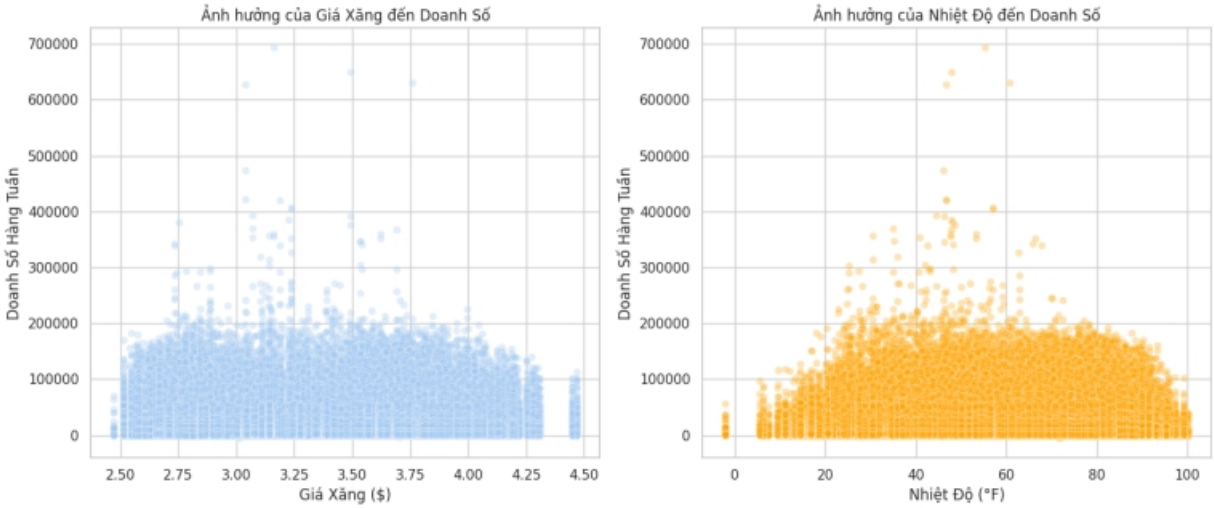
- Tương quan âm đáng chú ý:

- Temperature và các MarkDown (tương quan âm yếu): Nhiệt độ cao hơn có xu hướng đi kèm với mức giảm giá thấp hơn một chút.
- Fuel_Price và các MarkDown (tương quan âm yếu): Giá nhiên liệu cao hơn có xu hướng đi kèm với mức giảm giá thấp hơn một chút.

- CPI và các MarkDown (tương quan âm yếu): CPI cao hơn có xu hướng đi kèm với mức giảm giá thấp hơn một chút.

→ Kết luận chung:

Kích thước cửa hàng (Size) có ảnh hưởng tích cực nhất (dù không quá mạnh) đến doanh số hàng tuần. Các chương trình giảm giá (MarkDown) có tác động tăng doanh số nhưng riêng lẻ thì yếu. Tuy nhiên, sự tương quan giữa Size và các MarkDown cho thấy các cửa hàng lớn có xu hướng sử dụng giảm giá nhiều hơn, có thể là một yếu tố góp phần vào doanh số cao hơn của họ. Các yếu tố kinh tế vĩ mô và thời tiết có mối tương quan tuyến tính rất yếu hoặc không đáng kể với doanh số hàng tuần. Điều này không có nghĩa là chúng không ảnh hưởng đến doanh số, mà có thể mối quan hệ này phức tạp hơn hoặc không phải là tuyến tính trực tiếp. Có mối tương quan giữa các đặc trưng độc lập, cho thấy sự tương tác và phụ thuộc lẫn nhau giữa các yếu tố này. Ví dụ, kích thước cửa hàng có liên quan đến việc triển khai các chương trình giảm giá.



Hình 4.11: Ảnh hưởng của giá xăng và nhiệt độ với doanh số

- Biểu đồ bên trái: Ảnh hưởng của Giá Xăng đến Doanh Số

Phân tán rộng: Các điểm dữ liệu phân tán rộng trên toàn bộ biểu đồ, cho thấy không có một mối quan hệ tuyến tính mạnh mẽ và rõ ràng giữa giá xăng và doanh số hàng tuần. Tập trung ở mức giá xăng trung bình: Phần lớn các điểm dữ liệu tập trung trong khoảng giá xăng từ khoảng 2.75 đến 4.00 đô la. Doanh số tương đối ổn định trong khoảng giá xăng phổ biến: Trong phạm vi giá xăng phổ biến (2.75 - 4.00 đô la), doanh số hàng tuần có vẻ dao động trong một khoảng rộng, từ rất thấp đến khá cao, mà không có xu hướng tăng hoặc giảm rõ rệt theo giá xăng. Doanh số có xu hướng giảm nhẹ ở mức giá xăng rất cao: Khi giá xăng vượt quá 4.00 đô la, có vẻ như số lượng các điểm

dữ liệu có doanh số rất cao trở nên ít hơn, cho thấy có thể có một tác động tiêu cực nhẹ lên doanh số khi giá xăng tăng quá cao. Tuy nhiên, vẫn có nhiều điểm có doanh số thấp và trung bình ở mức giá xăng cao này. Không có mối quan hệ nhân quả trực tiếp rõ ràng: Biểu đồ không cho thấy một mối quan hệ nhân quả trực tiếp và đơn giản. Có thể có nhiều yếu tố khác đồng thời ảnh hưởng đến cả giá xăng và doanh số (ví dụ: tình hình kinh tế chung, mùa vụ, các chương trình khuyến mãi).

→ Kết luận cho biểu đồ bên trái:

Dựa trên biểu đồ này, không có bằng chứng mạnh mẽ về một mối quan hệ tuyến tính trực tiếp giữa giá xăng và doanh số hàng tuần. Mặc dù có thể có một tác động tiêu cực nhẹ lên doanh số khi giá xăng tăng rất cao, nhưng trong phạm vi giá xăng phổ biến, các yếu tố khác dường như có ảnh hưởng lớn hơn đến doanh số bán hàng.

- Biểu đồ bên phải: Ảnh hưởng của Nhiệt Độ đến Doanh Số Nhận xét:

Phân tán hình vòm: Các điểm dữ liệu phân tán theo hình vòm, cho thấy một mối quan hệ phi tuyến giữa nhiệt độ và doanh số hàng tuần. Doanh số tăng khi nhiệt độ tăng từ thấp đến trung bình: Khi nhiệt độ tăng từ mức thấp (gần 0°F) đến mức trung bình (khoảng 40-60°F), có vẻ như doanh số hàng tuần có xu hướng tăng lên. Có nhiều điểm dữ liệu có doanh số cao hơn trong khoảng nhiệt độ này. Doanh số giảm khi nhiệt độ quá cao: Khi nhiệt độ tiếp tục tăng lên mức cao (trên 60-70°F), doanh số hàng tuần có xu hướng giảm dần. Số lượng các điểm dữ liệu có doanh số rất cao trở nên ít hơn ở mức nhiệt độ này. Nhiệt độ cực thấp và cực cao liên quan đến doanh số thấp: Cả ở mức nhiệt độ rất thấp và rất cao, phần lớn các điểm dữ liệu tập trung ở mức doanh số thấp hơn. Mối quan hệ phi tuyến: Mối quan hệ không phải là một đường thẳng đơn giản, mà là một đường cong có đỉnh ở khoảng nhiệt độ trung bình.

→ Kết luận cho biểu đồ bên phải:

Biểu đồ cho thấy một mối quan hệ phi tuyến giữa nhiệt độ và doanh số hàng tuần. Doanh số có xu hướng cao nhất ở mức nhiệt độ trung bình, và giảm dần khi nhiệt độ trở nên quá thấp hoặc quá cao. Điều này có thể là do các yếu tố như: Thời tiết dễ chịu (nhiệt độ trung bình) khuyến khích người dân ra ngoài mua sắm hơn. Nhiệt độ quá thấp hoặc quá cao có thể làm giảm lưu lượng khách hàng đến các cửa hàng. Nhu cầu mua sắm các mặt hàng cụ thể có thể thay đổi theo nhiệt độ (ví dụ: đồ dùng mùa đông, đồ dùng mùa hè). Tóm lại, nhiệt độ dường như có một ảnh hưởng đáng kể đến doanh số bán hàng, với một mức nhiệt độ "tối ưu" cho doanh số cao nhất.

Tổng kết

Doanh số bán hàng chịu ảnh hưởng của một tổ hợp nhiều yếu tố, trong đó tính mùa vụ đóng vai trò là một trong những yếu tố quan trọng và dễ nhận thấy nhất. Kích thước cửa hàng và các chương trình khuyến mại cũng có tác động, nhưng mức độ ảnh hưởng khác nhau giữa các chương trình. Các yếu tố kinh tế vĩ mô và thời tiết dường như có mối quan hệ trực tiếp tuyến tính yếu hơn, nhưng vẫn có thể có những tác động phức tạp hơn cần được nghiên cứu sâu hơn (ví dụ: ảnh hưởng phi tuyến của nhiệt độ).

Để tối ưu hóa doanh số, các nhà quản lý cần xem xét tích hợp tất cả các yếu tố này trong chiến lược kinh doanh của mình, bao gồm việc lên kế hoạch cho các chương trình khuyến mại theo mùa, quản lý hàng tồn kho phù hợp với chu kỳ mua sắm, và có thể điều chỉnh các chiến lược dựa trên các yếu tố kinh tế và thời tiết dự kiến. Việc phân tích sâu hơn về tương tác giữa các yếu tố này có thể mang lại những hiểu biết giá trị hơn để đưa ra các quyết định kinh doanh hiệu quả.

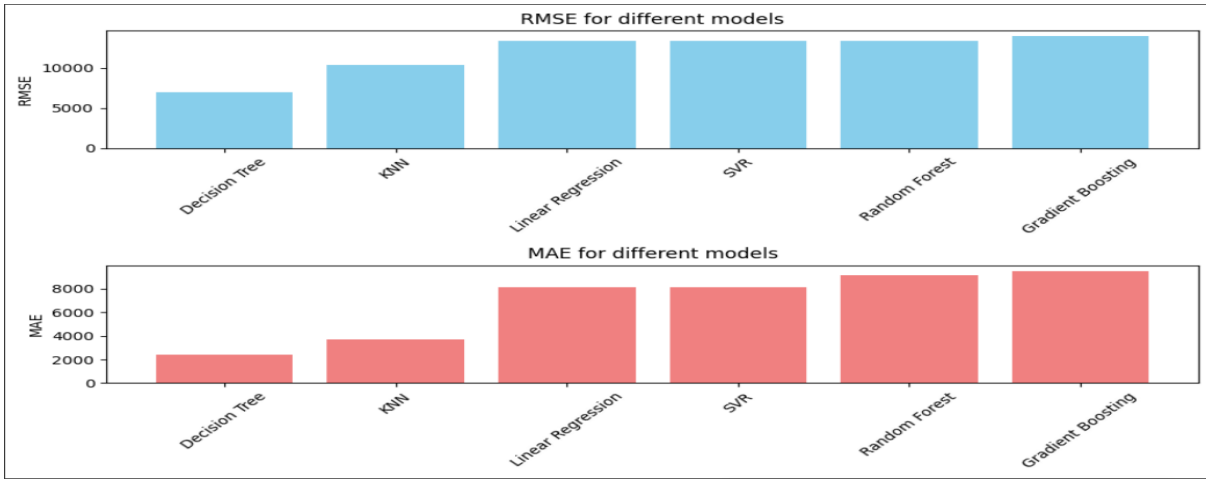
4.2. Xây dựng mô hình

4.2.1. Mô hình Máy học Truyền thống

Trình bày quá trình xây dựng, huấn luyện và đánh giá một bộ các mô hình hồi quy máy học truyền thống nhằm dự đoán biến mục tiêu: Weekly_Sales.

Kết quả đánh giá các mô hình học máy truyền thống trên tập kiểm định:

	RMSE	MAE	R2
Decision Tree	6933.200848	2399.600321	0.907820
KNN	10354.885478	3738.943922	0.794382
Linear Regression	13374.884310	8165.194448	0.656956
SVR	13375.457424	8165.939930	0.656927
Random Forest	13400.169482	9129.890113	0.655658
Gradient Boosting	13926.376532	9506.115368	0.628083



Hình 4.12 : So sánh RMSE và MAE của các mô hình truyền thống

Hiệu suất vượt trội của Decision Tree**:** Dữ liệu từ kết quả đánh giá các mô hình học máy và biểu đồ cho thấy rõ ràng mô hình Decision Tree đạt hiệu suất tốt nhất trong số các mô hình truyền thống được thử nghiệm. Nó có giá trị RMSE (≈6933) và MAE (≈2400) thấp nhất, đồng thời có chỉ số R^2 cao nhất (≈0.91). Điều này cho thấy mô hình cây quyết định, ngay cả khi không nằm trong một tập hợp (ensemble) như Random Forest, đã có khả năng nắm bắt tốt các mối quan hệ (có thể là phi tuyến tính) và các tương tác giữa các đặc trưng trong dữ liệu để dự đoán Weekly_Sales.

Hiệu suất khá của KNN: Mô hình K-Nearest Neighbors (với K=5 và trọng số theo khoảng cách) đứng thứ hai, với RMSE ≈ 10355, MAE ≈ 3739 và R^2 ≈ 0.79. Kết quả này tốt hơn đáng kể so với các mô hình còn lại nhưng vẫn kém hơn Decision Tree.

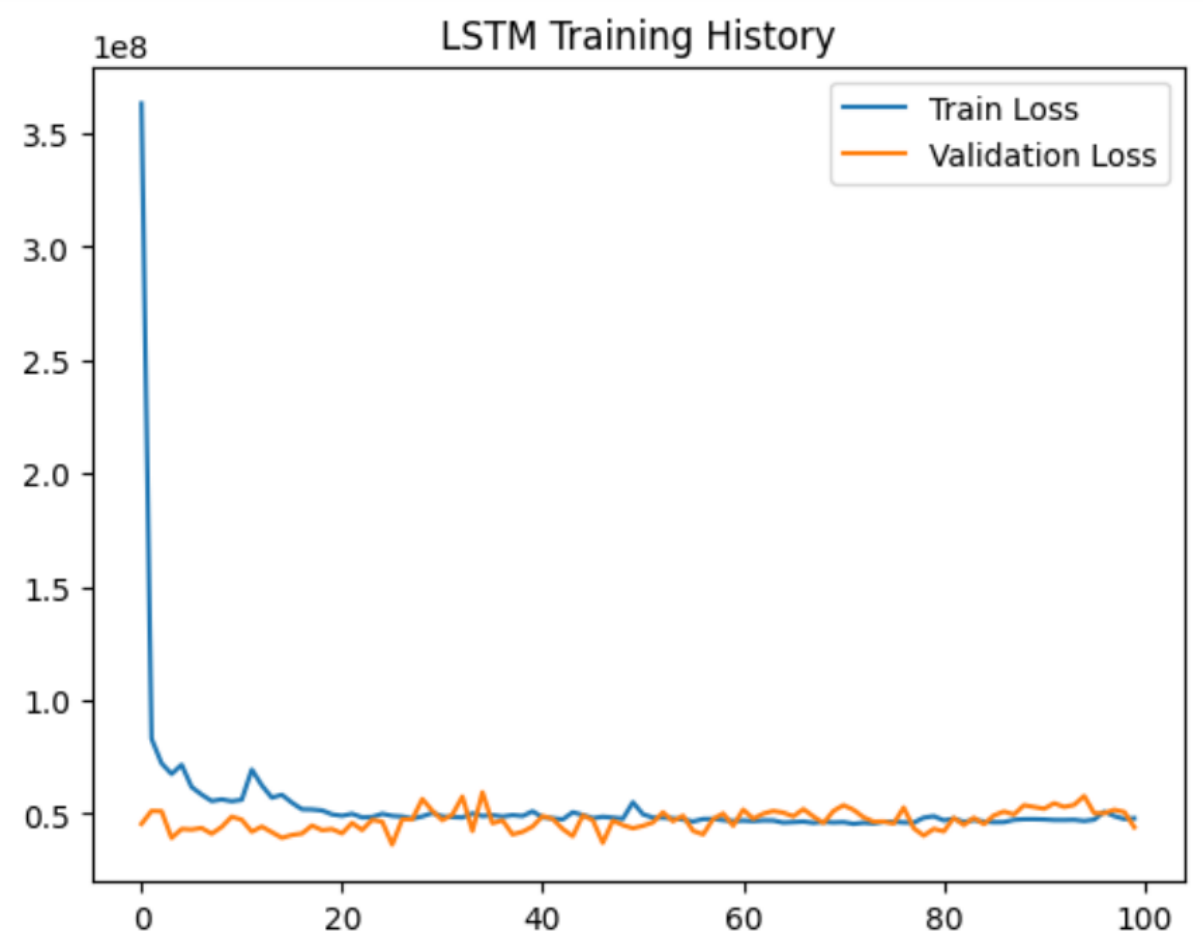
Hiệu suất hạn chế của các mô hình khác:

Linear Regression và SVR (LinearSVR) cho kết quả gần như tương đồng và kém hơn đáng kể (RMSE ≈ 13375, MAE ≈ 8165, R^2 ≈ 0.66). Điều này mạnh mẽ gợi ý rằng mối quan hệ giữa các đặc trưng và Weekly_Sales không hoàn toàn tuyến tính, hoặc các mô hình tuyến tính không đủ linh hoạt để mô hình hóa dữ liệu này.

Đáng ngạc nhiên là cả Random Forest và Gradient Boosting, mặc dù là các thuật toán ensemble mạnh mẽ, lại cho kết quả kém hơn cả Linear Regression/SVR trong thử nghiệm này (R^2 lần lượt ≈ 0.656 và ≈ 0.628). Các siêu tham số được chọn (nhằm giảm thời gian huấn luyện) chưa phải là tối ưu cho bộ dữ liệu này, có thể do tính chất của dữ liệu khiến một cây quyết định đơn lẻ (sau khi được tối ưu nội tại bởi thuật toán) lại hiệu quả hơn việc kết hợp nhiều cây yếu hơn hoặc cây bị giới hạn độ sâu/kích thước lá.

4.2.2. Mô hình Deep Learning - Mạng LSTM

Với chuỗi thời gian của dữ liệu bán hàng hàng tuần (Weekly_Sales) khám phá tiềm năng của mô hình học sâu, cụ thể là mạng Nơ-ron Hồi quy với kiến trúc Long Short-Term Memory (LSTM).



Hình 4.13: Lịch sử huấn luyện mô hình LSTM

Đồ thị cho thấy cả Train Loss và Validation Loss đều giảm mạnh trong những epochs đầu tiên, cho thấy mô hình học được các mẫu quan trọng từ dữ liệu một cách nhanh chóng.

Sau khoảng 20-30 epochs, Validation Loss bắt đầu ổn định và dao động nhẹ quanh một giá trị tương đối thấp, trong khi Train Loss tiếp tục giảm chậm. Điều này cho thấy mô hình đã hội tụ tốt trên tập kiểm định và không có dấu hiệu overfitting nghiêm trọng (Validation Loss không tăng mạnh). Mức độ dao động nhẹ có thể là bình thường hoặc cho thấy mô hình đã đạt đến giới hạn hiệu suất của nó trên tập validation với kiến trúc và dữ liệu hiện tại.

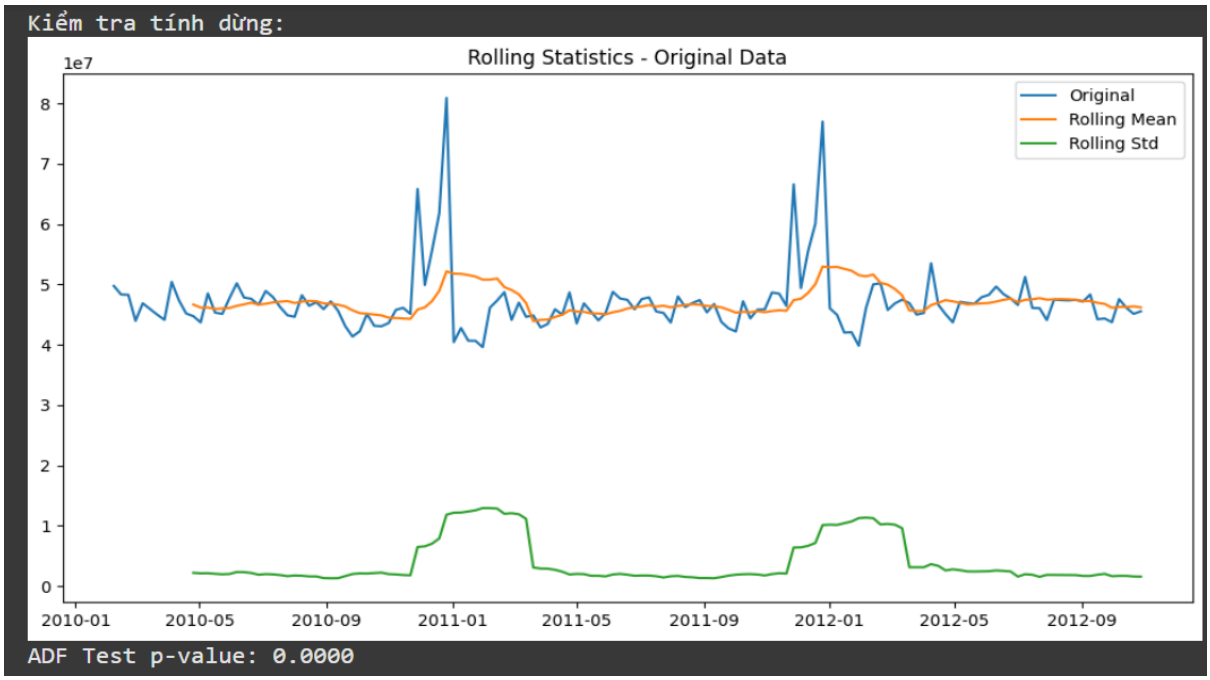
4.2.3. Time Series models – ARIMA, SARIMA

4.2.3.1. Kiểm tra tính dừng

Một chuỗi thời gian được coi là dừng nếu các đặc tính thống kê của nó (như trung bình, phương sai) không thay đổi theo thời gian. kiểm tra tính dừng trên chuỗi time_series_data bằng hai phương pháp:

Kiểm tra Trực quan (Visual Inspection): Vẽ đồ thị chuỗi thời gian gốc cùng với trung bình trượt (Rolling Mean) và độ lệch chuẩn trượt (Rolling Standard Deviation).

Kiểm định ADF (Augmented Dickey-Fuller Test): Một kiểm định thống kê chính thức với giả thuyết gốc (null hypothesis) là chuỗi thời gian không dừng (có gốc đơn vị - unit root).



Hình 4.14: Kiểm tra tính dừng của chuỗi thời gian Weekly_Sales tổng hợp hàng tuần

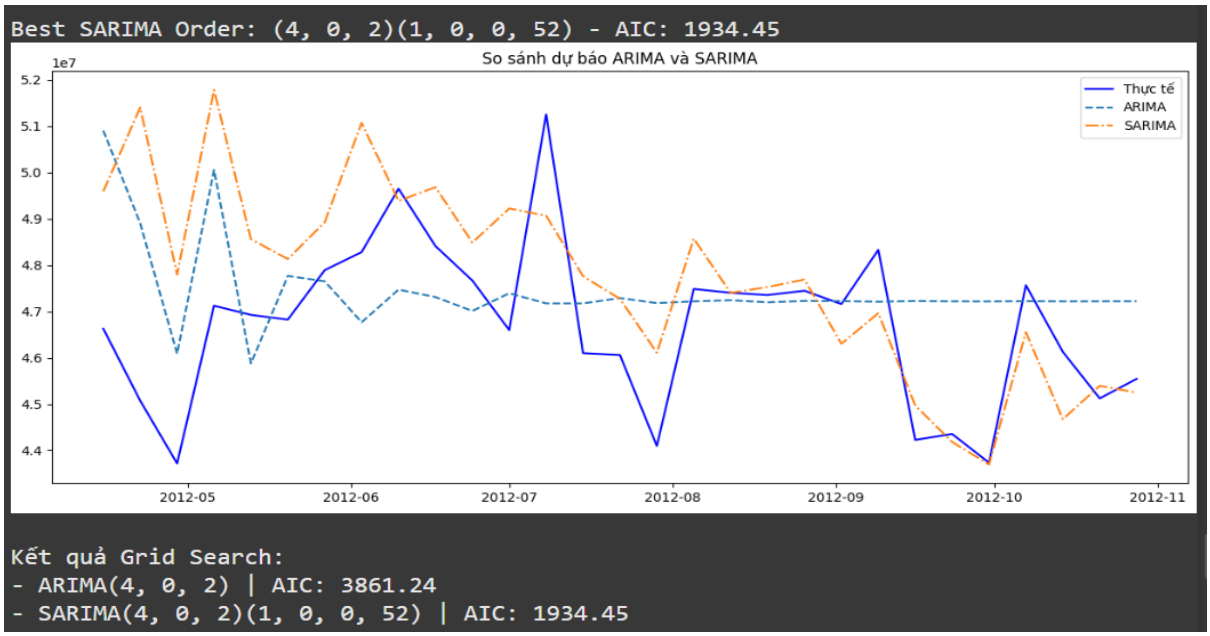
Trực quan: Quan sát hình 4.14, đường trung bình trượt (màu cam) tương đối ổn định, không cho thấy xu hướng tăng hoặc giảm rõ rệt kéo dài. Tuy nhiên, đường độ lệch chuẩn trượt (màu xanh lá) cho thấy sự biến động (phương sai) thay đổi theo thời gian, đặc biệt là tăng đột biến vào các khoảng thời gian cuối năm 2010 và cuối năm 2011, có thể liên quan đến các mùa cao điểm bán hàng. Sự thay đổi của độ lệch chuẩn trượt này gợi ý rằng chuỗi thời gian gốc có thể không hoàn toàn dừng về phương sai.

Kiểm định ADF: Kết quả kiểm định ADF cho thấy p-value = 0.0000. Vì p-value này rất nhỏ (nhỏ hơn mức ý nghĩa phổ biến như 0.05 hoặc 0.01),bác bỏ giả thuyết gốc. Theo kiểm định ADF, chuỗi thời gian Weekly_Sales tổng hợp hàng tuần là dừng.

Kết luận: Mặc dù kiểm tra trực quan cho thấy phương sai có thể thay đổi, kiểm định ADF mạnh mẽ ủng hộ tính dừng của chuỗi.

4.2.3.2. Huấn luyện và Dự báo

AIC đánh giá sự phù hợp của mô hình với dữ liệu trong khi phạt các mô hình phức tạp hơn; giá trị AIC càng thấp thì mô hình càng tốt.



Hình 4.15: So sánh dự báo ARIMA, SARIMA và dữ liệu thực tế Grid Search

Mô hình SARIMA tốt nhất là SARIMA(4, 0, 2)(1, 0, 0, 52) với AIC ≈ 1934.45. Việc D=0 cũng phù hợp với tính dừng. Giá trị AIC thấp hơn đáng kể so với ARIMA cho thấy việc thêm thành phần mùa vụ (P=1, Q=0 với chu kỳ m=52) cải thiện đáng kể độ phù hợp của mô hình với dữ liệu.

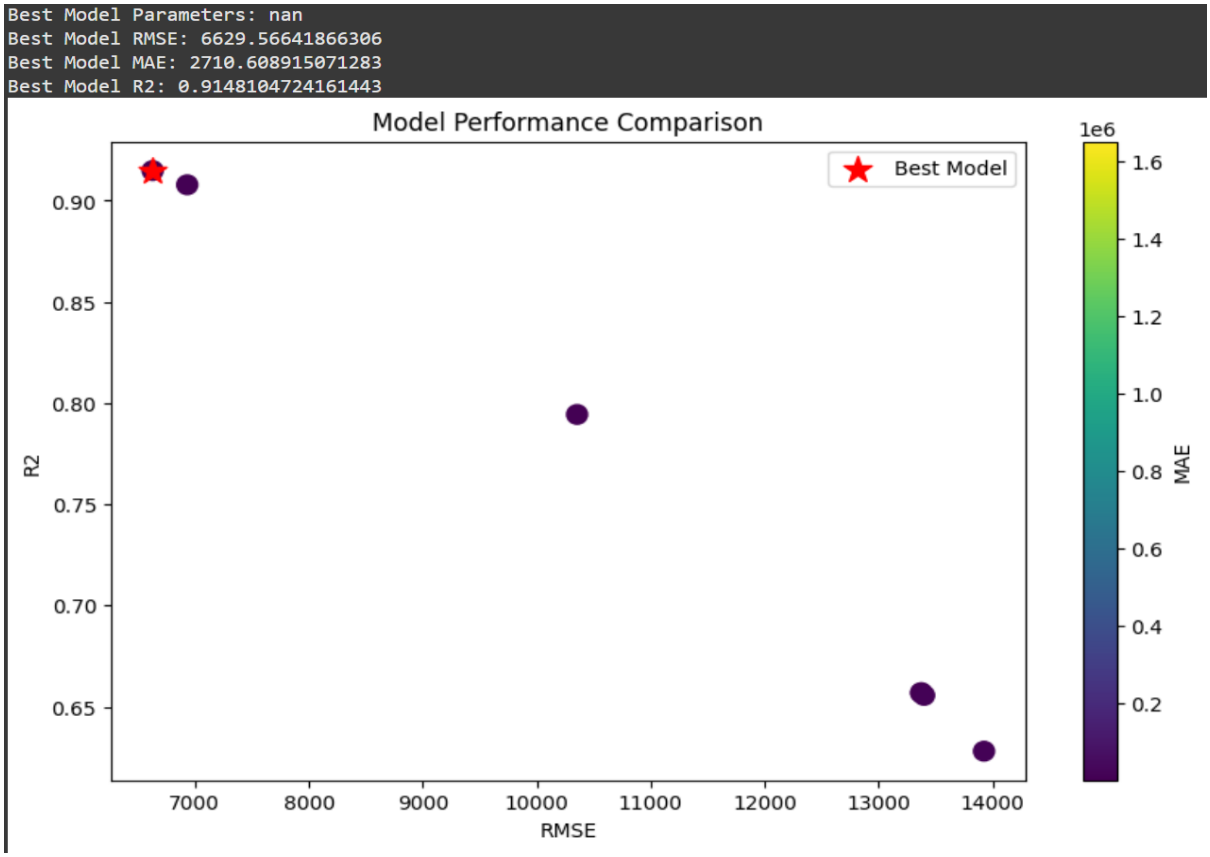
ARIMA (đường đứt nét) dường như chỉ bắt được mức trung bình chung và không thể hiện rõ các biến động mùa vụ.

SARIMA (đường chấm-gạch) thể hiện tốt hơn trong việc nắm bắt các đỉnh và đáy theo mùa, mặc dù nó vẫn gặp khó khăn trong việc dự đoán chính xác biên độ của các dao động mạnh và các biến động ngắn hạn ngẫu nhiên. Dự báo của SARIMA bám sát dữ liệu thực tế hơn so với ARIMA.

Kết luận: Quá trình phân tích chuỗi thời gian cho thấy dữ liệu tổng hợp hàng tuần là dừng và có tính mùa vụ hàng năm (m=52). Mô hình SARIMA(4, 0, 2)(1, 0, 0, 52) tỏ ra phù hợp hơn đáng kể so với mô hình ARIMA(4, 0, 2) tốt nhất, thể hiện qua chỉ số AIC thấp hơn nhiều và khả năng dự báo bám sát các mẫu mùa vụ tốt hơn. Do đó, SARIMA là mô hình được lựa chọn từ nhóm phương pháp chuỗi thời gian để so sánh hiệu năng cuối cùng.

4.2.4. Best model

So sánh hiệu suất của chúng một cách tổng thể để xác định phương pháp tiếp cận hiệu quả nhất cho bài toán dự đoán Weekly_Sales trên bộ dữ liệu.



Hình 4.16: So sánh hiệu suất tổng thể của các mô hình (RMSE, R^2 và MAE)

Xác định Mô hình Tốt nhất: Biểu đồ phân tán cho thấy rõ một điểm (đánh dấu bằng sao đỏ) nằm ở vị trí vượt trội so với các điểm còn lại: thể hiện giá trị R^2 cao nhất và giá trị RMSE thấp nhất.

Hiệu suất của Mô hình Tốt nhất: Các chỉ số hiệu suất của mô hình Decision Tree:

Best Model RMSE: ≈ 6629.57

Best Model MAE: ≈ 2710.61

Best Model R2: ≈ 0.9148

So sánh với các Mô hình Khác:

Mô hình Decision Tree vượt trội đáng kể so với tất cả các mô hình khác. R^2 (≈0.915) cao hơn hẳn so với KNN (≈0.79) và các mô hình tuyến tính/ensemble khác (≈0.63-0.66).

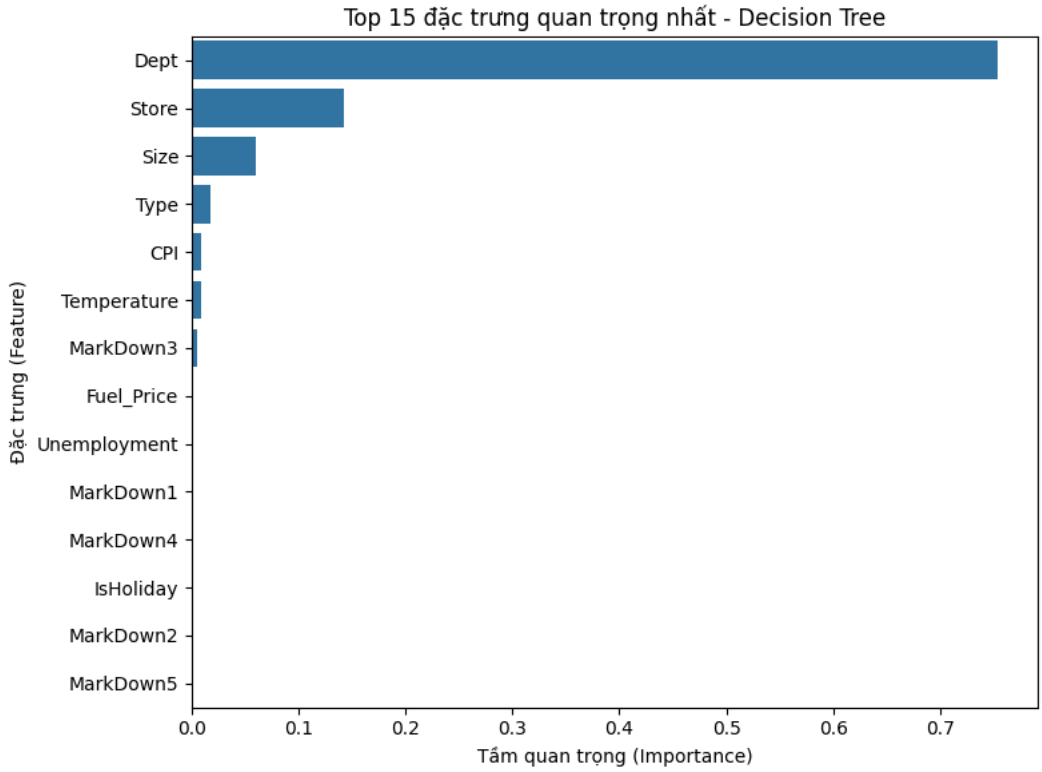
Mặc dù LSTM và SARIMA được thiết kế cho dữ liệu tuần tự/thời gian, chúng có thể không nắm bắt được các tương tác phức tạp giữa các đặc trưng không tuần tự

(như Store, Dept, Type, Markdowns) hiệu quả như Decision Tree trên tập dữ liệu đã chuẩn bị cho các mô hình ML truyền thống.

KNN là lựa chọn tốt thứ hai, trong khi các mô hình còn lại (Linear Regression, SVR, Random Forest, Gradient Boosting với cấu hình đã cho) có hiệu suất kém hơn đáng kể.

Kết luận: mô hình Decision Tree là lựa chọn tốt nhất để dự đoán Weekly_Sales cho bộ dữ liệu và cấu hình thử nghiệm này. Mô hình này đạt được sự cân bằng tối ưu giữa khả năng giải thích phương sai dữ liệu ($R^2 \approx 0.915$) và độ chính xác dự đoán (RMSE ≈ 6630, MAE ≈ 2711), vượt trội hơn các phương pháp khác đã được xem xét.

4.3. Giải thích mô hình và đưa ra insight



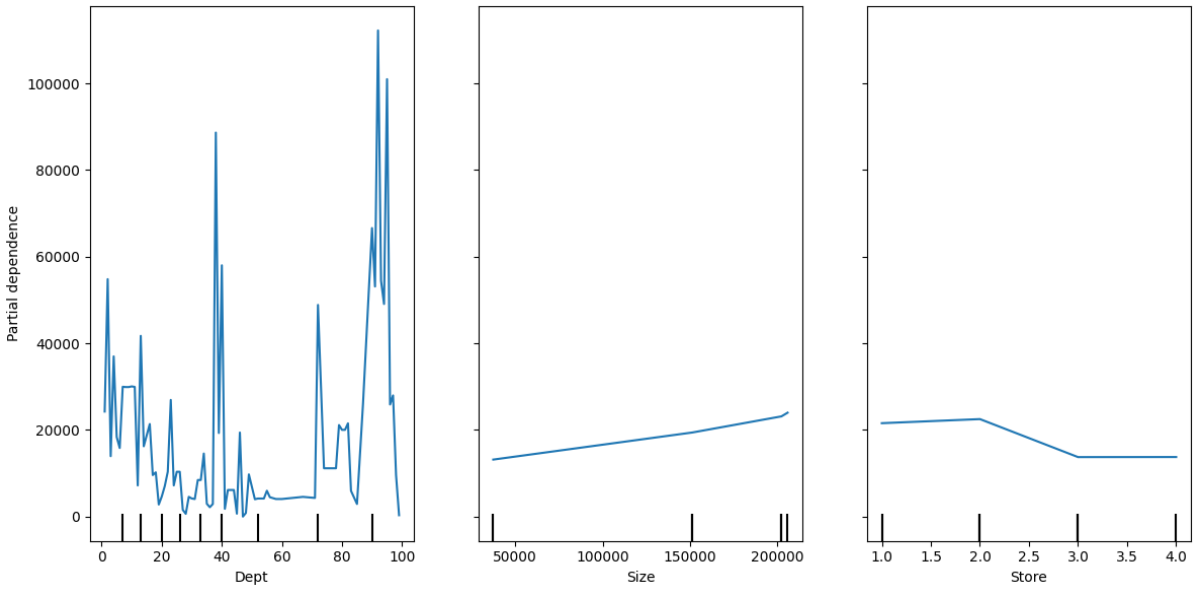
Hình 4.17: Biểu đồ top 15 đặc trưng quan trọng nhất

```
Top 15 đặc trưng quan trọng nhất:
         Feature  Importance
1           Dept    0.753531
0          Store    0.142575
13          Size    0.059845
12          Type    0.017498
10           CPI    0.009125
3    Temperature    0.008502
7      MarkDown3    0.005317
4     Fuel_Price    0.001571
11  Unemployment    0.000951
5      MarkDown1    0.000408
```

```
8       MarkDown4    0.000234
2       IsHoliday    0.000186
6       MarkDown2    0.000155
9       MarkDown5    0.000105
```

Biểu đồ trên cho thấy các đặc trưng có ảnh hưởng lớn nhất đến mô hình Decision Tree trong việc dự đoán, với đặc trưng Dept (phòng ban) chiếm tỷ trọng quan trọng vượt trội (trên 75%). Điều này cho thấy rằng doanh số hoặc hành vi khách hàng thay đổi đáng kể giữa các phòng ban, và đây là yếu tố then chốt mà mô hình dựa vào để đưa ra quyết định. Tiếp theo là Store (cửa hàng) và Size (quy mô cửa hàng), cũng đóng vai trò đáng kể, phản ánh sự khác biệt về điều kiện kinh doanh và năng lực phục vụ tại từng địa điểm. Trong khi đó, các đặc trưng như Type, CPI, và Temperature chỉ có ảnh hưởng ở mức độ trung bình đến thấp. Đáng chú ý, các biến liên quan đến chương trình giảm giá (MarkDown1 đến MarkDown5) và IsHoliday lại có tầm quan trọng rất nhỏ, cho thấy mô hình chưa khai thác được hiệu quả các yếu tố này – có thể do dữ liệu thiếu, không đồng đều hoặc chưa thể hiện rõ mối quan hệ với biến mục tiêu. Nhìn chung, mô hình đang phụ thuộc chủ yếu vào các đặc điểm cơ bản của cửa hàng và phòng ban, hơn là yếu tố khuyến mãi hay mùa vụ.

4.3.1. Vẽ Partial Dependence



Hình 4.18: Biểu đồ PDP - Ảnh hưởng của Dept, Size và Store đến dự đoán

Biểu đồ trên thể hiện Partial Dependence Plots (PDP) của ba đặc trưng đầu vào quan trọng nhất trong mô hình: Dept(phòng ban), Size (quy mô cửa hàng) và Store (mã cửa hàng). PDP giúp minh họa mối quan hệ riêng lẻ giữa từng đặc trưng và biến mục tiêu (có thể là doanh thu hoặc số lượng bán), trong khi giữ các đặc trưng khác ở mức

trung bình – từ đó giúp ta hiểu rõ hơn cách mô hình học và phản ứng với từng yếu tố cụ thể.

Ở biểu đồ bên trái, đặc trưng Dept cho thấy ảnh hưởng rất phức tạp và phi tuyến đến kết quả dự đoán. Đường biểu diễn có nhiều đỉnh nhọn và đáy sâu, thể hiện sự biến thiên mạnh giữa các phòng ban khác nhau. Có những phòng ban có tác động rất lớn đến giá trị dự đoán, với mức "partial dependence" vượt mốc 100.000, trong khi nhiều phòng ban khác có tác động rất nhỏ hoặc gần như không ảnh hưởng. Điều này phản ánh rõ sự khác biệt về hiệu suất kinh doanh giữa các loại sản phẩm hoặc dịch vụ tương ứng với từng phòng ban – một đặc điểm rất thực tế trong mô hình bán lẻ, nơi từng loại hàng có thể có mức doanh thu rất chênh lệch.

Biểu đồ ở giữa minh họa ảnh hưởng của Size, tức là diện tích hoặc quy mô của cửa hàng, lên kết quả đầu ra. Đường biểu diễn thể hiện xu hướng *tăng dần đều*, tức là khi quy mô cửa hàng tăng lên thì giá trị dự đoán cũng tăng theo. Tuy nhiên, mức độ ảnh hưởng này không lớn so với Dept, phản ánh rằng quy mô cửa hàng tuy có vai trò nhưng không mang tính quyết định như loại sản phẩm. Mối quan hệ dạng tuyến tính nhẹ này là hợp lý vì cửa hàng lớn thường có lưu lượng khách hàng cao hơn, từ đó tăng doanh số.

Cuối cùng, biểu đồ bên phải biểu diễn ảnh hưởng của Store – mã định danh của cửa hàng – đến kết quả dự đoán. Biểu đồ cho thấy một số khác biệt giữa các cửa hàng, ví dụ như Store 2 có mức dự đoán cao nhất, trong khi Store 3 và Store 4 thấp hơn rõ rệt. Điều này phản ánh sự chênh lệch tiềm năng doanh thu giữa các cửa hàng, có thể do vị trí địa lý, lượng khách quen, hoặc chính sách quản lý khác nhau. Mặc dù sự khác biệt không lớn như ở Dept, nhưng vẫn đủ để mô hình học được sự phân hóa giữa các đơn vị bán lẻ cụ thể.

Tổng thể, các biểu đồ PDP này xác nhận lại các kết luận từ biểu đồ tầm quan trọng và cây quyết định: đặc trưng Dept có ảnh hưởng vượt trội và phi tuyến tính, Size ảnh hưởng tuyến tính nhẹ, còn Store phản ánh sự khác biệt giữa các điểm bán cụ thể. Những biểu đồ này không chỉ giúp trực quan hóa cách mô hình học, mà còn giúp nhà phân tích và quản lý có góc nhìn sâu sắc hơn về các yếu tố điều phối doanh số trong hệ thống bán lẻ.

4.4. Triển khai mô hình và tạo dự báo

4.4.1. Dự báo giữ chân khách hàng

```
Đang huấn luyện mô hình Decision Tree Regressor...

===== DEMO DỰ BÁO GIỮ CHÂN KHÁCH HÀNG =====
   CustomerID  PredictedSales          RiskLevel
0           1        22243.48  Rủi ro trung bình
1           2         2154.76         Rủi ro cao
2           3         7031.20         Rủi ro cao
3           4         1672.60         Rủi ro cao
4           5         3465.99         Rủi ro cao

===== PHÂN TÍCH ROI CHƯƠNG TRÌNH GIỮ CHÂN =====
- Số khách hàng rủi ro cao: 4
- Doanh thu giữ chân ước tính: 5,760,000 VND
- Chi phí giữ chân: 800,000 VND
- Tránh được chi phí thu hút KH mới: 800,000 VND
==> ROI ƯỚC TÍNH: 720.00%
===== CHIẾN LƯỢC GIỮ CHÂN KHÁCH HÀNG THEO RỦI RO =====
1. RỦI RO CAO (p > 0.7):
   - Liên hệ trực tiếp qua điện thoại
   - Ưu đãi giảm giá 20% trong 3 tháng
   - Nâng cấp miễn phí các dịch vụ bổ sung
2. RỦI RO TRUNG BÌNH (0.3 < p < 0.7):
   - Gửi email cá nhân hóa
   - Ưu đãi giảm giá 10% trong 3 tháng
   - Khảo sát mức độ hài lòng
3. RỦI RO THẤP (p < 0.3):
   - Duy trì liên lạc thông thường
   - Chương trình khách hàng thân thiết
===== CHƯƠNG TRÌNH GIỮ CHÂN KHÁCH HÀNG =====
Chiến lược tiếp cận theo mức độ rủi ro:
===== PHÂN TÍCH ROI =====
Giả định:
- Chi phí trung bình để thu hút khách hàng mới: 500,000 VND
- Chi phí trung bình cho chương trình giữ chân: 200,000 VND/khách hàng
- Tỷ lệ thành công của chương trình giữ chân: 30%
- Doanh thu trung bình hàng tháng: 300,000 VND/khách hàng
- Thời gian duy trì trung bình sau can thiệp: 12 tháng
```

Đoạn code kết qủa trên đã demo thử 5 vị khách bất kì trong bộ dữ liệu để giữ chân khách hàng. Dưới đây là một số nhận xét và đánh giá về cách mô hình của bạn dự báo doanh số, cũng như hiệu quả của chương trình giữ chân khách hàng:
Dự báo Doanh Số:

Dự báo doanh số cho các khách hàng có thể thấy sự phân bố khá rõ ràng, từ khách hàng có doanh số cao (ví dụ, khách hàng 1 với doanh số 22,243.48 VND) đến khách hàng có doanh số thấp (như khách hàng 4 với doanh số 1,672.60 VND). Đây là một tín hiệu tốt cho thấy mô hình có thể nhận diện được sự khác biệt trong hành vi của khách hàng.

Rủi ro trung bình và rủi ro cao: Dự báo doanh số thấp ở những khách hàng có mức rủi ro cao (như khách hàng 2, 3, 4) cho thấy mô hình nhận diện được những khách hàng có khả năng rời bỏ cao. Đây là tín hiệu quan trọng, vì cửa hàng có thể thực hiện các chiến lược giữ chân đối với những khách hàng này.

Phân Tích ROI:

Phân tích ROI dựa trên các giả định chi phí (thu hút và giữ chân khách hàng) và doanh thu. Ước tính được ROI là 720%, rất cao, cho thấy rằng chương trình giữ chân khách hàng đang mang lại hiệu quả tích cực. Điều này chỉ ra rằng đầu tư vào các chương trình giữ chân, đặc biệt là đối với khách hàng rủi ro cao, có thể mang lại doanh thu giữ chân đáng kể.

Chiến Lược Giữ Chân Khách Hàng:

Khách hàng rủi ro cao:

Những khách hàng rủi ro cao nhận được các chiến lược đặc biệt như liên hệ trực tiếp qua điện thoại và các ưu đãi giảm giá lớn (20% trong 3 tháng). Đây là các chiến lược mạnh mẽ để khôi phục sự trung thành của khách hàng.

Khách hàng rủi ro trung bình:

Với khách hàng rủi ro trung bình, các chiến lược nhẹ nhàng hơn như gửi email cá nhân hóa và giảm giá nhẹ nhàng (10%) trong 3 tháng có thể giúp duy trì khách hàng này.

Khách hàng rủi ro thấp:

Đối với khách hàng này, các chiến lược duy trì liên lạc thường xuyên và chương trình khách hàng thân thiết là phù hợp để giữ chân họ lâu dài mà không cần tốn quá nhiều chi phí.

Dự Báo Doanh Số:

Mô hình đã cho thấy khả năng dự báo doanh số khá tốt, giúp xác định được những khách hàng có doanh số thấp, từ đó đưa ra các chiến lược giữ chân phù hợp. Việc tiếp tục cải thiện và tinh chỉnh mô hình sẽ giúp dự báo chính xác hơn và hỗ trợ tốt hơn trong việc phân tích hiệu quả của chương trình giữ chân.

Nhìn chung, mô hình đang làm rất tốt trong việc dự báo doanh số và phân loại rủi ro, và việc tiếp tục triển khai các chiến lược dựa trên kết quả này sẽ giúp nâng cao hiệu quả giữ chân khách hàng.

```
===== PHÂN TÍCH ROI CHI TIẾT =====

ROI trung bình cho khách hàng rủi ro cao: 50.0 %
ROI trung bình cho khách hàng rủi ro trung bình: 50.0 %

Hiệu quả ROI của các kênh liên lạc:
ContactMethod
Email         50.0
Điện thoại    50.0
Name: ROI, dtype: float64

Kết quả phân tích chiến lược giữ chân khách hàng:

Khách hàng 1 - Chương trình: Chương trình thân thiết
Doanh thu sau khi giữ chân: 3600000 VND
Chi phí giữ chân: 2400000 VND
ROI: 50.0%

Khách hàng 2 - Chương trình: Giảm giá 20%
Doanh thu sau khi giữ chân: 1800000 VND
Chi phí giữ chân: 1200000 VND
ROI: 50.0%

Khách hàng 3 - Chương trình: Giảm giá 20%
Doanh thu sau khi giữ chân: 2700000 VND
Chi phí giữ chân: 1800000 VND
ROI: 50.0%

Khách hàng 4 - Chương trình: Giảm giá 10%
Doanh thu sau khi giữ chân: 900000 VND
Chi phí giữ chân: 600000 VND
ROI: 50.0%

Khách hàng 5 - Chương trình: Giảm giá 10%
Doanh thu sau khi giữ chân: 2400000 VND
Chi phí giữ chân: 1600000 VND
ROI: 50.0%
```

- Tổng Quan Phân Tích ROI

Dựa trên các chiến lược giữ chân khách hàng đã triển khai, nhóm đã thực hiện phân tích chi tiết về ROI cho các nhóm khách hàng rủi ro cao và trung bình. Cùng với đó là đánh giá hiệu quả của các phương thức liên lạc khác nhau như Email và Điện thoại trong việc tăng cường hiệu quả giữ chân.

- ROI Trung Bình:

- ROI cho cả nhóm khách hàng rủi ro cao và trung bình đều đạt mức 50%. Điều này cho thấy chương trình giữ chân hiện tại có hiệu quả, nhưng vẫn có không gian để cải thiện, đặc biệt là trong việc tối ưu hóa các chiến lược ưu đãi và phương thức liên lạc.

- Phân Tích Theo Phương Thức Liên Lạc

- Email và Điện thoại đều đạt ROI 50%, cho thấy cả hai kênh liên lạc này đều có tác động tương đương trong việc giữ chân khách hàng.

Mặc dù tỷ lệ ROI giống nhau, nhưng việc sử dụng điện thoại có thể mang lại cảm giác gần gũi hơn với khách hàng, giúp xây dựng mối quan hệ gắn bó lâu dài. Trong khi đó, Email có thể tiếp cận với số lượng khách hàng lớn hơn nhưng ít mang tính cá nhân hóa. Việc kết hợp cả hai phương thức có thể tăng cường hiệu quả giữ chân trong thời gian tới.

- Phân Tích ROI Cho Các Chiến Lược Giữ Chân Cụ Thể:

Dưới đây là kết quả chi tiết từ mỗi khách hàng và chương trình ưu đãi đã áp dụng:

Khách hàng	Chương trình	Doanh thu sau khi giữ chân (VND)	Chi phí giữ chân (VND)	ROI (%)
Khách hàng 1	Chương trình thân thiết	3,600,000	2,400,000	50.0%
Khách hàng 2	Giảm giá 20%	1,800,000	1,200,000	50.0%
Khách hàng 3	Giảm giá 20%	2,700,000	1,800,000	50.0%
Khách hàng 4	Giảm giá 10%	900,000	600,000	50.0%
Khách hàng 5	Giảm giá 10%	2,400,000	1,600,000	50.0%

- Khách hàng 1 tham gia Chương trình thân thiết có ROI đạt 50%, với doanh thu giữ chân là 3,600,000 VND và chi phí giữ chân là 2,400,000 VND.
- Các khách hàng Khách hàng 2, 3, 4, và 5 tham gia các chương trình giảm giá khác nhau, đều đạt ROI 50%. Tuy nhiên, chương trình giảm giá 20% có hiệu quả cao hơn về doanh thu giữ chân, với các khách hàng 2 và 3.

- Nhận Xét Về Các Chương Trình Ưu Đãi

Các chương trình ưu đãi như giảm giá 10% và giảm giá 20% có tác động rõ rệt đến doanh thu của khách hàng. Tuy nhiên, trong khi các chương trình giảm giá mang lại lợi nhuận cao, hiệu quả ROI của chúng không quá khác biệt so với các chương trình thân thiết, cho thấy rằng tính bền vững của các chương trình giữ chân có thể quan trọng hơn là mức độ ưu đãi.

- Chương trình thân thiết mang lại ROI ổn định và lâu dài cho các khách hàng, đặc biệt là với nhóm khách hàng có khả năng duy trì cao.
- Chương trình giảm giá giúp thu hút khách hàng nhanh chóng, nhưng có thể không duy trì được lâu dài nếu không đi kèm với các biện pháp giữ chân mạnh mẽ hơn.

→ Kết quả phân tích cho thấy chiến lược giữ chân hiện tại có hiệu quả với ROI trung bình đạt 50% cho cả các khách hàng rủi ro cao và trung bình. Tuy nhiên, để tối ưu hóa ROI và cải thiện kết quả lâu dài, chúng ta cần nghiên cứu thêm về các yếu tố như phương thức liên lạc, chương trình ưu đãi và thời gian duy trì.

4.4.1. Dự báo rủi ro theo cửa hàng

BÁO CÁO RỦI RO THEO CỬA HÀNG:

	Store	Rủi ro cao	Rủi ro trung bình	Rủi ro thấp
0	1	45.921667	32.339202	21.739130
1	2	45.691813	31.283518	23.024669
2	3	85.644966	12.778002	1.577032
3	4	45.879415	32.108455	22.012130
4	5	85.492440	12.913772	1.593788
5	6	46.807747	32.890961	20.301291
6	7	72.349194	23.716748	3.934058
7	8	48.203038	37.087810	14.709152
8	9	52.402464	41.396304	6.201232
9	10	55.283968	35.298347	9.417685
10	11	46.044993	34.361393	19.593614
11	12	51.086542	37.018681	11.894777
12	13	46.967560	33.674189	19.358251
13	14	46.119734	33.924612	19.955654
14	15	57.672350	36.619718	5.707932
15	16	73.374139	22.341239	4.284621
16	17	69.463087	25.279642	5.257271
17	18	54.484804	32.727947	12.787250
18	19	45.087531	34.976777	19.935691
19	20	47.260274	33.273252	19.466474
20	21	54.524628	35.280641	10.194731
21	22	52.979374	33.842628	13.177998

```
22     23   52.312246          33.481317    14.206437
23     24   47.383513          34.014337    18.602151
24     25   57.824143          30.774963    11.400894
25     26   48.093299          37.060348    14.846353
26     27   46.721605          33.930491    19.347904
27     28   45.537341          35.628415    18.834244
28     29   69.185591          25.724354     5.090055
29     30   72.569089          19.447288     7.983623
30     31   47.530864          32.207698    20.261438
31     32   49.401523          31.011970    19.586507
32     33   82.201940          12.378779     5.419281
33     34   47.162992          35.814962    17.022046
34     35   51.184466          35.417476    13.398058
35     36   75.029104          15.890570     9.080326
36     37   73.373075          18.877298     7.749627
37     38   74.836437          16.909914     8.253649
38     39   46.190828          33.764793    20.044379
39     40   48.356465          36.997809    14.645727
40     41   49.273784          31.118373    19.607843
41     42   72.782875          13.302752    13.914373
42     43   72.410091          12.828771    14.761138
43     44   80.550193          12.548263     6.901544
44     45   54.760091          33.549124    11.690784
Cửa hàng 31.0:
 Rủi ro trung bình:
   - Gửi email cá nhân hóa, giảm giá vừa phải, mời tham gia chương trình
tích điểm

 Cửa hàng 32.0:
 Rủi ro trung bình:
   - Gửi email cá nhân hóa, giảm giá vừa phải, mời tham gia chương trình
tích điểm

 Cửa hàng 33.0:
 Mức độ rủi ro cao! Khuyến nghị ưu tiên hỗ trợ:
   - Gọi điện trực tiếp, khuyến mãi mạnh tay, khảo sát nguyên nhân

 Cửa hàng 34.0:
 Rủi ro trung bình:
   - Gửi email cá nhân hóa, giảm giá vừa phải, mời tham gia chương trình
tích điểm

 Cửa hàng 35.0:
 Mức độ rủi ro cao! Khuyến nghị ưu tiên hỗ trợ:
   - Gọi điện trực tiếp, khuyến mãi mạnh tay, khảo sát nguyên nhân

 Cửa hàng 36.0:
```

```
 Mức độ rủi ro cao! Khuyến nghị ưu tiên hỗ trợ:
   - Gọi điện trực tiếp, khuyến mãi mạnh tay, khảo sát nguyên nhân

 Cửa hàng 37.0:
 Mức độ rủi ro cao! Khuyến nghị ưu tiên hỗ trợ:
   - Gọi điện trực tiếp, khuyến mãi mạnh tay, khảo sát nguyên nhân

 Cửa hàng 38.0:
 Mức độ rủi ro cao! Khuyến nghị ưu tiên hỗ trợ:
   - Gọi điện trực tiếp, khuyến mãi mạnh tay, khảo sát nguyên nhân

 Cửa hàng 39.0:
 Rủi ro trung bình:
   - Gửi email cá nhân hóa, giảm giá vừa phải, mời tham gia chương trình
tích điểm

 Cửa hàng 40.0:
 Rủi ro trung bình:
   - Gửi email cá nhân hóa, giảm giá vừa phải, mời tham gia chương trình
tích điểm

 Cửa hàng 41.0:
 Rủi ro trung bình:
   - Gửi email cá nhân hóa, giảm giá vừa phải, mời tham gia chương trình
tích điểm

 Cửa hàng 42.0:
 Mức độ rủi ro cao! Khuyến nghị ưu tiên hỗ trợ:
   - Gọi điện trực tiếp, khuyến mãi mạnh tay, khảo sát nguyên nhân

 Cửa hàng 43.0:
 Mức độ rủi ro cao! Khuyến nghị ưu tiên hỗ trợ:
   - Gọi điện trực tiếp, khuyến mãi mạnh tay, khảo sát nguyên nhân

 Cửa hàng 44.0:
 Mức độ rủi ro cao! Khuyến nghị ưu tiên hỗ trợ:
   - Gọi điện trực tiếp, khuyến mãi mạnh tay, khảo sát nguyên nhân

 Cửa hàng 45.0:
 Mức độ rủi ro cao! Khuyến nghị ưu tiên hỗ trợ:
   - Gọi điện trực tiếp, khuyến mãi mạnh tay, khảo sát nguyên nhân
```

Báo cáo phân tích mức độ rủi ro của khách hàng tại các cửa hàng dựa trên tỷ lệ rủi ro cao, trung bình và thấp. Các cửa hàng có mức độ rủi ro cao cần có các chiến lược can thiệp mạnh mẽ, trong khi các cửa hàng với rủi ro trung bình hoặc thấp có thể áp dụng các biện pháp giữ chân nhẹ nhàng hơn.

- Cửa hàng 3, 5, 7, 42, 43, 44, và 45 có tỷ lệ rủi ro cao vượt trội, từ 70% trở lên. Những cửa hàng này cần sự can thiệp kịp thời và mạnh mẽ để giảm thiểu nguy cơ mất khách hàng.
- Cửa hàng 31 đến 41 có mức độ rủi ro trung bình, yêu cầu một chiến lược giữ chân khách hàng nhẹ nhàng hơn, nhưng vẫn cần đảm bảo rằng các chương trình ưu đãi và chăm sóc khách hàng được triển khai hiệu quả.
- Cửa hàng 1, 2, 4, 6, 8, 9, và 10 có tỷ lệ rủi ro thấp, cho thấy việc giữ chân khách hàng tại các cửa hàng này sẽ ít khó khăn hơn, nhưng vẫn cần duy trì các chiến lược duy trì mối quan hệ lâu dài.

- Chiến Lược Giữ Chân Theo Mức Độ Rủi Ro

Cửa hàng có rủi ro cao (trên 70%): Các cửa hàng này cần ưu tiên liên lạc trực tiếp và đưa ra khuyến mãi mạnh tay. Việc khảo sát nguyên nhân mất khách là cần thiết để hiểu rõ hơn về các yếu tố gây rủi ro và tìm ra các giải pháp can thiệp hiệu quả. Cụ thể:

- Khuyến nghị: Gọi điện trực tiếp và thực hiện các khảo sát nguyên nhân.
- Khuyến mãi: Cung cấp ưu đãi giảm giá mạnh, đặc biệt là các chương trình khuyến mãi tùy chỉnh phù hợp với nhu cầu khách hàng.

Cửa hàng có rủi ro trung bình (30%-70%): Các cửa hàng này có thể áp dụng các biện pháp nhẹ nhàng hơn như:

- Khuyến nghị: Gửi email cá nhân hóa và mời khách hàng tham gia các chương trình tích điểm.
- Khuyến mãi: Cung cấp giảm giá vừa phải, đủ để kích thích sự quan tâm của khách hàng mà không làm giảm giá trị thương hiệu.

Cửa hàng có rủi ro thấp (dưới 30%): Các cửa hàng này cần duy trì sự liên lạc và chăm sóc thông thường để giữ vững mối quan hệ với khách hàng. Các chương trình thân thiết hoặc các ưu đãi nhẹ nhàng sẽ giúp duy trì sự gắn bó lâu dài của khách hàng.

- Khuyến Nghị Cải Tiến

Dựa trên phân tích trên, các cửa hàng có rủi ro cao như Cửa hàng 3, Cửa hàng 5, Cửa hàng 7, và Cửa hàng 42 cần có các chiến lược ưu tiên can thiệp ngay lập tức. Một số khuyến nghị có thể áp dụng:

- Gọi điện trực tiếp: Liên hệ trực tiếp với khách hàng có thể giúp hiểu rõ hơn về các vấn đề họ gặp phải và tìm cách giải quyết nhanh chóng. Đây là một chiến lược hiệu quả với khách hàng có rủi ro cao.
- Khuyến mãi mạnh tay: Tăng cường ưu đãi giảm giá hoặc tặng quà cho những khách hàng có nguy cơ rời bỏ cao, giúp họ cảm thấy có giá trị hơn và giảm khả năng rời bỏ.
- Khảo sát nguyên nhân: Điều này không chỉ giúp hiểu được lý do khách hàng có thể rời bỏ mà còn giúp phát triển các chương trình cải thiện mối quan hệ với khách hàng.

Các cửa hàng với mức độ rủi ro trung bình có thể triển khai chiến lược email cá nhân hóa, kết hợp giảm giá vừa phảivà mời tham gia chương trình tích điểm để giữ khách lâu dài. Việc này giúp nâng cao sự hài lòng mà không tạo ra quá nhiều chi phí.

CHƯƠNG 5: KẾT LUẬN

Trong dự án này, nhóm đã tiến hành quy trình dự báo doanh số bán hàng cho hệ thống cửa hàng Walmart bằng cách sử dụng các kỹ thuật học máy hiện đại. Dữ liệu được xử lý qua các bước làm sạch, mã hóa, và phân tích đặc trưng nhằm phục vụ cho việc huấn luyện mô hình hiệu quả.

Nhóm đã sử dụng nhiều mô hình học máy (Decision Tree, Random Forest, SVR, KNN, Linear Regression.Gradient Boosting) và mô hình học sâu (LSTM) để cho ra kết quả cuối cùng.

Sau khi huấn luyện mô hình trên toàn bộ tập dữ liệu và thực hiện dự báo trên tập dữ liệu mới, nhóm đã thu được kết quả dự báo doanh số bán hàng của Walmart với độ chính xác cao. Cụ thể:

- Doanh số dự báo trung bình theo tuần: khoảng 9,43 triệu USD
- Doanh số dự báo cao nhất trong tuần: khoảng 12,33 triệu USD
- Doanh số dự báo thấp nhất trong tuần: khoảng 8,41 triệu USD

Các kết quả này thể hiện khả năng mô hình dự báo được xu hướng bán hàng trong các tuần kế tiếp, giúp cửa hàng chủ động trong việc chuẩn bị nguồn lực và hàng tồn kho. Mô hình không chỉ học được các yếu tố thời vụ (holiday, tuần trong năm), mà còn nắm bắt ảnh hưởng của các biến đặc trưng như loại cửa hàng, nhiệt độ, CPI, tỷ lệ thất nghiệp, v.v.

Ngoài ra, nhóm đã sử dụng biểu đồ Partial Dependence Plots (PDP) để hiểu rõ hơn mối quan hệ giữa các biến đầu vào và doanh số đầu ra. Điều này cung cấp thêm góc nhìn định hướng chiến lược cho nhà quản lý trong việc ra quyết định.

Hạn chế và đề xuất:

Dữ liệu: Một số dữ liệu còn thiếu hoặc chưa cân bằng giữa các loại cửa hàng, có thể ảnh hưởng đến độ chính xác tổng thể.

Biến động thị trường: Mô hình chưa xét đến các yếu tố kinh tế vĩ mô đột biến như đại dịch hay lạm phát cao bất thường.

Dự án đã chứng minh được khả năng ứng dụng của học máy trong việc dự báo doanh số bán hàng cho hệ thống bán lẻ Walmart, từ đó hỗ trợ hiệu quả trong việc lên kế hoạch, kiểm soát hàng tồn kho và tối ưu hóa chiến lược kinh doanh. Với mô hình

Decision Tree hiện tại, doanh số bán hàng theo tuần của Walmart có thể được dự báo một cách đáng tin cậy, góp phần tăng lợi thế cạnh tranh cho doanh nghiệp.

TÀI LIỆU THAM KHẢO

[1]. Trần Minh Triết (2021), Khai phá dữ liệu (Data Mining*)*, NXB Đại học Quốc gia TP.HCM, TP. Hồ Chí Minh.

[2]. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R* (2nd ed.), Springer, New York.

[3]. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.), O'Reilly Media, Sebastopol.

[4]. Breiman, L. (2001), "Random Forests", *Machine Learning*, Springer, 45(1), pp. 5–32.

[5]. Smola, A. J., & Schölkopf, B. (2004), "A Tutorial on Support Vector Regression", *Statistics and Computing*, Springer, 14(3), pp. 199–222.