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 ĐOÁN DOANH SỐ BÁN HÀNG TRONG TƯƠNG LAI TỪ DỮ LIỆU LỊCH SỬ GIAO DỊCH CỦA NHÀ BÁN LẺ TRỰC TUYẾN

Sinh viên thực hiện:

LÊ ĐÌNH TÙNG	DHKL16A1HN	22174600004
NGUYỄN THỊ THUYÊN	DHKL16A1HN	22174600032
HUỲNH NGỌC TƯỜNG VI	DHKL16A1HN	22174600005

Giáo viên giảng dạy: 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 ĐOÁN DOANH SỐ BÁN HÀNG TRONG TƯƠNG LAI TỪ DỮ LIỆU LỊCH SỬ GIAO DỊCH CỦA NHÀ BÁN LẺ TRỰC TUYẾN

Sinh viên thực hiện:

LÊ ĐÌNH TÙNG	DHKL16A1HN	22174600004
NGUYỄN THỊ THUYÊN	DHKL16A1HN	22174600032
HUỲNH NGỌC TƯỜNG VI	DHKL16A1HN	22174600005

Giáo viên giảng dạy: LÊ HẰNG ANH

Hà Nội - 2025

PHIẾU ĐĂNG KÝ ĐỀ TÀI

1. Tên đề tài: Dự đoán doanh số bán hàng trong tương lai từ dữ liệu lịch sử giao dịch của nhà bán lẻ trực tuyến

2. Thông tin nhóm sinh viên: (Nhóm TTV3)

Sinh viên 1 (Nhóm trưởng):

- **Họ và tên:** Lê Đình Tùng
- **Mã sinh viên:** 22174600004
- **Điện thoại:** 0396881736
- **Email:** ldtung.dhkl16a1hn@sv.uneti.edu.vn

Sinh viên 2:

- **Họ và tên:** Nguyễn Thị Thuyên
- **Mã sinh viên:** 22174600032
- **Điện thoại:** 0356491588
- **Email:** ntthuyen.dhkl16a1hn@sv.uneti.edu.vn

Sinh viên 3:

- **Họ và tên:** Huỳnh Ngọc Tường Vi
- **Mã sinh viên:** 22174600005
- **Điện thoại:** 0866701821
- **Email:** hntvi.dhkl16a1hn@sv.uneti.edu.vn

3. Tóm tắt nội dung đề tài:

Trong bối cảnh thương mại điện tử phát triển, dự báo doanh số đóng vai trò quan trọng trong hỗ trợ doanh nghiệp ra quyết định chiến lược. Đề tài này xây dựng các mô hình dự báo doanh số tương lai dựa trên dữ liệu lịch sử từ bộ dữ liệu Online Retail II, ghi nhận giao dịch của một công ty bán lẻ trực tuyến có trụ sở tại Anh.

Quy trình gồm: tiền xử lý dữ liệu (làm sạch, xử lý thiếu, loại bỏ lỗi), phân tích khám phá (EDA) để tìm hiểu xu hướng tiêu dùng, và xây dựng mô hình dự báo như Random Forest, Prophet, LSTM theo ngày, tuần, tháng.

Hiệu quả mô hình được đánh giá bằng RMSE, MAE, MAPE, R^2, kết hợp trực quan hóa kết quả để hỗ trợ ra quyết định.

Kết quả giúp doanh nghiệp tối ưu tồn kho, lập kế hoạch khuyến mãi và nâng cao năng lực đáp ứng thị trường. Đề tài cũng góp phần thúc đẩy chuyển đổi số và khai thác dữ liệu lớn trong bán lẻ.

Ngày 11 tháng 4 năm 2025

Nhóm trưởng

Tùng

Lê Đình Tùng

ĐỀ CƯƠNG CHI TIẾT ĐỀ TÀI

1. Tên đề tài: Dự đoán doanh số bán hàng trong tương lai từ dữ liệu lịch sử giao dịch của nhà bán lẻ trực tuyến

2. Mục tiêu đề tài:

Mục tiêu chính của đề tài là xây dựng mô hình dự báo doanh số bán hàng trong tương lai dựa trên dữ liệu lịch sử từ bộ dữ liệu **Online Retail II**. Thông qua việc ứng dụng các kỹ thuật học máy, đề tài hướng tới việc xác định xu hướng mua sắm của khách hàng, từ đó giúp doanh nghiệp bán lẻ dự đoán được lượng doanh thu theo các khoảng thời gian cụ thể như theo ngày, tuần hoặc tháng.

Các mục tiêu cụ thể bao gồm:

- **Tiền xử lý và phân tích dữ liệu:** Làm sạch, xử lý dữ liệu bị thiếu và dữ liệu không hợp lệ; phân tích hành vi tiêu dùng và xu hướng mua hàng.
- **Lựa chọn và huấn luyện mô hình:** Áp dụng các thuật toán học máy và mô hình chuỗi thời gian: Random Forest, Prophet, LSTM để dự báo doanh số.
- **Đánh giá hiệu quả mô hình:** Sử dụng các chỉ số đánh giá như RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) và R^2 (R-squared) để so sánh và lựa chọn mô hình tối ưu nhất.
- **Đưa ra khuyến nghị ứng dụng:** Đề xuất cách ứng dụng mô hình trong thực tế kinh doanh như quản lý tồn kho, kế hoạch tiếp thị và dự báo tài chính.

Việc hoàn thành đề tài sẽ góp phần nâng cao hiệu quả ra quyết định cho doanh nghiệp bán lẻ trong việc tối ưu hóa doanh thu và hoạt động vận hành.

3. Tổng quan tình hình nghiên cứu thuộc lĩnh vực đề tài:

Dự báo doanh số bán hàng là một chủ đề quan trọng trong lĩnh vực phân tích dữ liệu và khoa học dữ liệu, đặc biệt trong bối cảnh dữ liệu lớn và chuyển đổi số đang diễn ra mạnh mẽ. Các doanh nghiệp hiện đại ngày càng phụ thuộc vào dữ liệu lịch sử để đưa ra quyết định chiến lược như dự báo doanh thu, tối ưu hóa tồn kho và điều chỉnh chiến lược tiếp thị. Trong những năm gần đây, nhiều nghiên cứu đã ứng dụng kỹ thuật học máy và mô hình chuỗi thời gian để giải quyết bài toán này. Các phương pháp truyền thống như ARIMA hay Holt-Winters vẫn được sử dụng nhờ tính đơn giản và khả năng diễn giải, song thiếu sức mạnh khi dữ liệu có tính phi tuyến hoặc biến động phức tạp. Trong khi đó, các mô hình hiện đại như Prophet (mô hình chuỗi thời gian với tính năng tự động phát hiện xu hướng và mùa vụ) hay các mô hình học sâu như mạng nơ-ron hồi tiếp LSTM (Long Short-Term Memory) cho thấy

hiệu quả vượt trội trong việc dự báo các chuỗi dữ liệu phức tạp theo thời gian. Bên cạnh đó, các thuật toán học máy mạnh như Random Forest cũng cho thấy khả năng dự báo ổn định khi dữ liệu được tổng hợp theo đặc trưng phù hợp.

Bộ dữ liệu Online Retail II từ UCI Machine Learning Repository cung cấp thông tin giao dịch của một công ty bán lẻ trực tuyến có trụ sở tại Anh, bao gồm mã hóa đơn, mã sản phẩm, mô tả, số lượng, ngày giao dịch, đơn giá và quốc gia khách hàng. Đây là nguồn dữ liệu phổ biến trong các nghiên cứu về hành vi tiêu dùng, phân khúc khách hàng và tính điểm RFM. Tuy nhiên, hầu hết các nghiên cứu hiện tại mới chỉ tập trung vào phân tích mô tả hoặc khai thác đặc điểm khách hàng, trong khi ứng dụng mô hình học máy để dự báo doanh số một cách toàn diện vẫn còn hạn chế. Mục tiêu của đề tài này là áp dụng ba mô hình đại diện cho các hướng tiếp cận khác nhau – Prophet (mô hình chuỗi thời gian hiện đại), Random Forest (thuật toán học máy mạnh) và LSTM (mạng nơ-ron sâu) – nhằm đánh giá hiệu quả dự báo và tìm ra giải pháp phù hợp cho doanh nghiệp trong việc lập kế hoạch và ra quyết định.

4. Nội dung đề tài:

Đề tài sử dụng bộ dữ liệu Online Retail II từ UCI Machine Learning Repository, ghi nhận chi tiết giao dịch của một doanh nghiệp bán lẻ trực tuyến tại Anh, gồm: mã hóa đơn, mã sản phẩm, mô tả, số lượng, ngày giao dịch, đơn giá và quốc gia khách hàng. Sau giai đoạn làm sạch (loại bỏ bản ghi hoàn trả, lỗi nhập liệu, thông tin thiếu) và chuẩn hóa ngày tháng để trích xuất đặc trưng thời gian (ngày, tuần, tháng, quý), doanh thu được tính bằng tích giữa số lượng và đơn giá, sau đó tổng hợp theo tuần hoặc tháng để phù hợp với yêu cầu mô hình hóa. Trước khi áp dụng mô hình, đề tài kiểm chứng độ ổn định (stationarity) của chuỗi, xử lý ngoại lệ và thiếu hụt dữ liệu.

Phần chính của nghiên cứu tập trung vào triển khai ba mô hình:

- Prophet: xử lý tự động các thành phần xu hướng, mùa vụ, ngày lễ và dị thường, cho phép xây dựng dự báo linh hoạt trên chuỗi kinh doanh.
- LSTM: mạng nơ-ron hồi tiếp sâu, ưu thế trong nắm bắt các phụ thuộc dài hạn và quan hệ phi tuyến phức tạp trong chuỗi thời gian.
- Random Forest: thuật toán ensemble cây quyết định, thích hợp khi dữ liệu được biểu diễn qua nhiều đặc trưng kỹ thuật, có thể đánh giá độ quan trọng của biến và chịu nhiễu tốt, hỗ trợ dự báo phi tuyến mà không yêu cầu giả định hình thức phân phối dữ liệu.

Mỗi mô hình sẽ được huấn luyện trên tập dữ liệu lịch sử và đánh giá độc lập trên tập kiểm tra. Các chỉ số định lượng bao gồm RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) và R^2 (hệ số xác định). Bên cạnh đó, kết quả dự báo được trực quan hóa so sánh với giá trị thực tế, nhằm đánh giá toàn diện ưu-nhược điểm và tính ứng dụng của từng phương pháp.

5. Phương pháp thực hiện:

Để thực hiện đề tài, nhóm sẽ áp dụng một quy trình phân tích dữ liệu tiêu chuẩn, bao gồm các bước chính: làm sạch dữ liệu, phân tích khám phá dữ liệu (EDA), xây dựng mô hình dự báo, đánh giá mô hình và trực quan hóa kết quả. Đây là quy trình phổ biến trong khoa học dữ liệu nhằm đảm bảo tính logic, khả thi và hiệu quả trong việc phân tích và dự báo.

Trước hết, dữ liệu được xử lý nhằm loại bỏ các bản ghi không hợp lệ (ví dụ: số lượng âm, đơn giá bằng 0, hóa đơn bị hủy có tiền tố 'C'), xử lý trùng lặp và giá trị thiếu. Ngày giao dịch được chuẩn hóa và trích xuất các đặc trưng thời gian như ngày, tuần, tháng, quý… để phục vụ phân tích chu kỳ và dự báo theo thời gian.

Sau tiền xử lý, nhóm tiến hành EDA bằng các biểu đồ trực quan như line chart, bar chart và heatmap để khám phá xu hướng doanh số theo thời gian, quốc gia, sản phẩm. Giai đoạn này giúp nhận diện rõ các yếu tố như tính mùa vụ, biến động doanh thu và xu hướng tăng/giảm bất thường.Trong bước xây dựng mô hình, ba phương pháp sẽ được triển khai và so sánh hiệu quả:

- Prophet (Facebook): Xử lý tốt tính mùa vụ và bất thường nhờ cơ chế phân tách xu hướng, mùa vụ và dị thường.
- LSTM (Long Short-Term Memory): Mạng nơ-ron hồi tiếp sâu có khả năng học các quan hệ dài hạn và phi tuyến trong chuỗi thời gian phức tạp.
- Random Forest: Tập hợp các cây quyết định ngẫu nhiên, phù hợp với dữ liệu có nhiều biến giải thích không tuyến tính và có thể khống chế quá khớp.

Mỗi mô hình sẽ được huấn luyện trên tập dữ liệu đã xử lý và đánh giá bằng các chỉ số RMSE (Root Mean Square Error), MAE (Mean Absolute Error) và R^2 (hệ số xác định). Mô hình có hiệu suất tối ưu nhất—cân bằng giữa độ chính xác dự báo và tính ổn định—sẽ được lựa chọn để áp dụng cho dự báo doanh số trong tương lai.

Cuối cùng, kết quả dự báo của mô hình được chọn sẽ được trực quan hóa thông qua dashboard hoặc biểu đồ tương tác, hỗ trợ doanh nghiệp theo dõi xu hướng kinh doanh theo thời gian thực và đưa ra các quyết định chiến lược một cách kịp thời và hiệu quả.

6. Phân công công việc (dự kiến):

Họ và tên	Mã sinh viên	Nội dung công việc được phân công
Lê Đình Tùng	22174600004	– Thu thập và tổng hợp bộ dữ liệu Online Retail II. – Tiền xử lý dữ liệu: loại bỏ bản ghi không hợp lệ, xử lý dữ liệu thiếu, định dạng lại ngày tháng, Tạo đặc trưng thời gian: ngày, tuần, tháng, năm. – Xây dựng mô hình Prophet. – Đánh giá và trực quan hóa kết quả Prophet. – So sánh hiệu quả các mô hình. – Tham gia sửa lỗi định dạng, chính tả báo cáo. – Thiết kế PowerPoint thuyết trình.
Nguyễn Thị Thuyên	22174600032	– Thực hiện phân tích EDA: trực quan hóa doanh số theo thời gian, quốc gia, sản phẩm. – Tạo biến mô tả (features) cho mô hình dự báo. – Xây dựng và huấn luyện mô hình Random Forest. – Đánh giá (RMSE, MAE, R^2) và trực quan hóa kết quả của Random Forest. – Tổng hợp nội dung và biên tập bản Word hoàn chỉnh.
Huỳnh Ngọc Tường Vi	22174600005	– Chuẩn hóa dữ liệu đầu vào,tạo đặc trưng thời gian, mô hình LSTM. – Xây dựng mô hình LSTM – Đánh giá và trực quan hóa kết quả của mô hình. – Phân tích dự đoán doanh số. – Tham gia sửa lỗi định dạng báo cáo. – Thiết kế PowerPoint thuyết trình.

7. Dự kiến kết quả đạt được:

Thông qua việc triển khai đề tài "Dự báo doanh số bán hàng dựa trên dữ liệu lịch sử từ bộ dữ liệu Online Retail II", nhóm kỳ vọng đạt được nhiều kết quả giá trị về cả học thuật và ứng dụng thực tiễn. Mục tiêu đầu tiên là xây dựng một mô hình dự báo doanh số chính xác, giúp doanh nghiệp dự đoán xu hướng bán hàng trong các tháng tiếp theo. Điều này sẽ hỗ trợ doanh nghiệp trong việc lập kế hoạch kinh doanh, quản lý hàng tồn kho và tối ưu hóa nguồn lực. Các mô hình dự báo sẽ được đánh giá thông qua các chỉ số như RMSE, MAE, MAPE và R^2, nhằm lựa chọn mô hình có độ chính xác cao và ổn định tốt nhất.

Nhóm cũng mong muốn xây dựng quy trình phân tích dữ liệu hoàn chỉnh, từ thu thập, tiền xử lý, phân tích khám phá (EDA), xây dựng mô hình đến đánh giá và trực quan hóa kết quả. Quy trình này không chỉ áp dụng cho bài toán này mà còn có thể tái sử dụng cho các bài toán dự báo trong các ngành khác như bán lẻ, tài chính, logistics. Việc thực hành trên bộ dữ liệu thực tế cũng giúp các thành viên nhóm nâng cao kỹ năng xử lý dữ liệu lớn và hiểu sâu về chuỗi thời gian, cùng với việc áp dụng các thuật toán học máy và học sâu.

Sản phẩm cuối cùng của đề tài sẽ là một báo cáo đầy đủ, kèm theo các biểu đồ trực quan, giúp người đọc dễ dàng tiếp cận và hiểu quy trình cùng kết quả nghiên cứu. Đề tài cũng củng cố kiến thức chuyên môn của nhóm, tạo nền tảng vững chắc cho các nghiên cứu sâu hơn trong tương lai hoặc ứng dụng vào công việc thực tế.

Hà Nội, ngày 11 tháng 4 năm 2025

Nhóm trưởng

Tùng

Lê Đình Tùng

MỤC LỤC

MỤC LỤC HÌNH VẼ III

MỤC LỤC BẢNG BIỂU IV

LỜI MỞ ĐẦU 1

CHƯƠNG 1. ĐẶT VẤN ĐỀ 2

1.1. LÝ DO CHỌN ĐỀ TÀI 2
1.2. MỤC TIÊU NGHIÊN CỨU 2
1.3. PHƯƠNG PHÁP NGHIÊN CỨU 2
1.4. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU 3

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 5

2.1. DỰ BÁO CHUỖI THỜI GIAN (TIME SERIES FORECASTING) 5
2.2. CÁC PHƯƠNG PHÁP DỰ BÁO CHUỖI THỜI GIAN 5
2.2.1. Mô hình Prophet 5
2.2.2. Mô hình Random Forest: 7
2.2.3. Mô hình LSTM (Long Short-Term Memory) 8
2.2.4. Các chỉ số đánh giá mô hình 10
2.2.4.1. Sai số trung bình tuyệt đối (Mean Absolute Error - MAE): 10
2.2.4.2. Căn bậc hai của sai số trung bình bình phương (Root Mean Squared Error - RMSE) 10
2.2.4.3. Hệ số xác định (R^2 Score) 11

CHƯƠNG 3. THỰC NGHIỆM 13

3.1. THU THẬP VÀ KHÁM PHÁ DỮ LIỆU 13
3.1.1. Tải và kiểm tra dữ liệu ban đầu 13
3.1.2. Kiểm tra chất lượng dữ liệu 14
3.1.2.1. Kiểm tra giá trị thiếu 14
3.1.2.2. Thống kê mô tả cho biến số 14
3.1.2.3. Kiểm tra giá trị trùng lặp 15
3.1.2.4 Kiểm tra dữ liệu không hợp lệ 15
3.1.2.5. Kiểm tra phân phối theo thời gian 15
3.2. TIỀN XỬ LÝ DỮ LIỆU 16

3.2.1. Làm sạch dữ liệu 16
3.2.2 Tạo các đặc trưng mới 17
3.3. EDA - PHÂN TÍCH KHÁM PHÁ DỮ LIỆU 17
3.3.1. Phân tích doanh số theo ngày 17
3.3.2. Phân tích doanh số theo tháng 18
3.3.3 Phân tích doanh số theo sản phẩm 20
3.3.4 Phân tích doanh số theo quốc gia 21
3.3.5. Phát hiện từ các phân tích EDA 22
3.4. XÂY DỰNG VÀ ĐÁNH GIÁ MÔ HÌNH 22
3.4.1. Chuẩn bị dữ liệu 22
3.4.2. Huấn luyện mô hình 24
3.4.2.1. Mô hình Prophet 24
3.4.2.2. Mô hình Random Forest 26
3.4.2.3. Mô hình LSTM 27
3.4.3. So sánh và đánh giá các mô hình 29
CHƯƠNG 4. KẾT QUẢ ĐẠT ĐƯỢC 32
4.1. KẾT QUẢ PHÂN TÍCH 32
4.2.DỰ ĐOÁN DOANH SỐ 34
CHƯƠNG 5. KẾT LUẬN,ƯU NHƯỢC ĐIỂM VÀ HƯỚNG PHÁT TRIỂN 37
5.1. KẾT LUẬN 37
5.2. ƯU ĐIỂM RÚT RA TỪ QUÁ TRÌNH DỰ BÁO DOANH SỐ 38
5.3. NHƯỢC ĐIỂM RÚT RA TỪ QUÁ TRÌNH DỰ BÁO DOANH SỐ 39
5.4. ĐỊNH HƯỚNG PHÁT TRIỂN VÀ ĐỀ XUẤT CHO DOANH NGHIỆP 40
TÀI LIỆU THAM KHẢO 41

MỤC LỤC HÌNH VẼ

Hình 2 - 1: Mô tả mô hình Prophet 5
Hình 2 - 2: Mô tả mô hình Random Foresst.....7
Hình 2 - 3: Mô tả mô hình LSTM 9
Hình 3 - 1: Tổng số lượng sản phẩm bán ra theo ngày 17
Hình 3 - 2: Tổng số sản phẩm bán ra theo từng tháng 18
Hình 3 - 3: Top 10 các sản phẩm bán chạy nhất 20
Hình 3 - 4: Tổng số lượng sản phẩm bán ra theo quốc gia 21
Hình 3 - 5: So sánh RMSE của 3 mô hình 29
Hình 3 - 6: So sánh MAE của 3 mô hình 30
Hình 3 - 7: So sánh R^2 của 3 mô hình 31
Hình 4 - 1: Dashboard cho kết quả phân tích.....32
Hình 4 - 2: Dự đoán doanh số các 3 tháng tiếp theo 34

MỤC LỤC BẢNG BIỂU

Bảng 3 - 1: Bảng thống kê mô tả biến số 14
Bảng 3 - 2: Kết quả đánh giá hiệu suất mô hình Prophet.....26
Bảng 3 - 3: Kết quả đánh giá hiệu suất mô hình Random Forest.....27
Bảng 3 - 4: Kết quả đánh giá hiệu suất mô hình LSTM.....29
Bảng 4 - 1: Kết quả so sánh của ba mô hình 32
Bảng 4 - 2: Kết quả dự đoán doanh số 3 tháng tiếp theo 34

LỜI MỞ ĐẦU

Trong thời đại chuyển đổi số, dữ liệu trở thành tài nguyên chiến lược, đóng vai trò then chốt trong sự phát triển bền vững của doanh nghiệp. Đặc biệt trong lĩnh vực thương mại điện tử, việc khai thác dữ liệu để phân tích hành vi tiêu dùng, xây dựng chiến lược và dự báo doanh số ngày càng trở nên cấp thiết. Các quyết định kinh doanh dựa trên dữ liệu không chỉ giúp tối ưu hóa vận hành mà còn nâng cao năng lực cạnh tranh và khả năng thích ứng với thị trường.

Tuy nhiên, nhiều doanh nghiệp vẫn gặp khó khăn trong việc xử lý và phân tích dữ liệu một cách hiệu quả, dù sở hữu lượng lớn dữ liệu giao dịch phản ánh xu hướng tiêu dùng và chu kỳ kinh doanh. Do đó, việc xây dựng mô hình dự báo doanh số dựa trên các kỹ thuật phân tích dữ liệu hiện đại và thuật toán học máy là một yêu cầu cấp bách.

Xuất phát từ thực tiễn đó, nhóm chúng em thực hiện đề tài "Dự đoán doanh số bán hàng trong tương lai từ dữ liệu lịch sử giao dịch của nhà bán lẻ trực tuyến", nhằm xây dựng quy trình phân tích dữ liệu toàn diện — từ thu thập, xử lý, khám phá đến xây dựng và đánh giá mô hình dự báo — qua đó hỗ trợ doanh nghiệp ra quyết định chính xác hơn.

Trong quá trình thực hiện, nhóm đã nhận được sự hướng dẫn tận tình và những góp ý quý báu từ cô Lê Hằng Anh – người đã truyền cảm hứng và kiến thức sâu sắc cho chúng em. Nhóm xin chân thành cảm ơn cô. Dù đã nỗ lực hoàn thiện, nhưng do kinh nghiệm thực tiễn còn hạn chế, bài làm khó tránh khỏi thiếu sót. Chúng em mong nhận được những góp ý chân thành từ cô và quý thầy cô để hoàn thiện đề tài tốt hơn.

Đồ án bao gồm các phần được phân chương như sau:

Chương 1: Đặt vấn đề (giới thiệu)

Chương 2: Cơ sở lý thuyết

Chương 3: Thực nghiệm

Chương 4: Kết quả đạt được

Chương 5: Kết luận, ưu điểm, nhược điểm, hướng phát triển

CHƯƠNG 1. ĐẶT VẤN ĐỀ

1.1. LÝ DO CHỌN ĐỀ TÀI

Trong bối cảnh thị trường bán lẻ ngày càng cạnh tranh khốc liệt, khả năng dự báo chính xác doanh số đóng vai trò then chốt trong việc tối ưu hóa vận hành, kiểm soát tồn kho và lập kế hoạch sản xuất – kinh doanh hiệu quả. Tuy nhiên, nhiều doanh nghiệp vẫn đang sử dụng các phương pháp dự báo truyền thống như hồi quy tuyến tính đơn giản hoặc trung bình trượt, vốn khó phản ánh các đặc điểm phức tạp của dữ liệu như tính mùa vụ, xu hướng thay đổi dài hạn hay những biến động bất thường.

Bộ dữ liệu Online Retail II từ UCI Machine Learning Repository là một nguồn dữ liệu thực tế có giá trị, ghi nhận các giao dịch thương mại điện tử tại châu Âu trong giai đoạn 2010–2011. Tuy được sử dụng rộng rãi trong các nghiên cứu về phân tích hành vi khách hàng và phân nhóm RFM, nhưng ứng dụng bộ dữ liệu này trong bài toán dự báo doanh số bằng các mô hình học máy hiện đại vẫn còn tương đối hạn chế. Chính vì vậy, đề tài lựa chọn khai thác bộ dữ liệu này để xây dựng và so sánh hiệu quả giữa các mô hình tiên tiến trong dự báo chuỗi thời gian, nhằm đóng góp thêm phương pháp và công cụ hỗ trợ doanh nghiệp trong lĩnh vực thương mại điện tử.

1.2. MỤC TIÊU NGHIÊN CỨU

Mục tiêu tổng quát của đề tài là xây dựng một quy trình dự báo doanh số bán hàng hiệu quả từ dữ liệu thực tế, đồng thời đánh giá hiệu suất của các mô hình học máy hiện đại trong phân tích chuỗi thời gian. Các mục tiêu cụ thể bao gồm:

- Thực hiện tiền xử lý và chuẩn hóa bộ dữ liệu Online Retail II để đảm bảo chất lượng đầu vào cho mô hình.
- Khám phá và phân tích các xu hướng doanh số theo thời gian, quốc gia và nhóm sản phẩm.
- Xây dựng, huấn luyện và đánh giá ba mô hình dự báo gồm: Prophet, Random Forest, và LSTM.
- So sánh hiệu suất dự báo của các mô hình dựa trên các chỉ số định lượng và đề xuất mô hình tối ưu cho ứng dụng thực tiễn.

1.3. PHƯƠNG PHÁP NGHIÊN CỨU

Phương pháp nghiên cứu được tổ chức thành các bước sau:

Thu thập và tiền xử lý dữ liệu: Dữ liệu được lọc bỏ các bản ghi không hợp lệ như hóa đơn hủy (InvoiceNo bắt đầu bằng “C”), đơn giá bằng 0, số lượng âm, các dòng bị

thiếu giá trị, và các dữ liệu trùng lặp. Đồng thời, trường thời gian sẽ được chuẩn hóa và trích xuất thêm các đặc trưng như ngày, tuần, tháng, quý nhằm phục vụ phân tích theo chu kỳ.

Phân tích dữ liệu khám phá (EDA): Sử dụng các kỹ thuật trực quan hóa dữ liệu như biểu đồ đường, histogram,… để nhận diện xu hướng tổng thể, tính mùa vụ, sự biến động doanh thu theo quốc gia và sản phẩm.

Xây dựng mô hình dự báo:

- Prophet: Mô hình do Facebook phát triển, phù hợp với dữ liệu chuỗi thời gian có yếu tố mùa vụ rõ rệt và khả năng điều chỉnh dị thường.
- Random Forest: Mô hình học máy phi tuyến dựa trên tập hợp các cây quyết định, có khả năng xử lý các quan hệ phức tạp giữa nhiều đặc trưng đầu vào và chống overfitting hiệu quả.
- LSTM: Mạng nơ-ron hồi tiếp sâu có khả năng học các quan hệ phụ thuộc dài hạn trong chuỗi thời gian, thích hợp với dữ liệu có tính phi tuyến và biến động mạnh.

Các mô hình được đánh giá bằng các chỉ số phổ biến trong dự báo chuỗi thời gian gồm RMSE, MAE và R^2, nhằm đo lường độ sai lệch và mức độ giải thích biến động doanh số. Kết quả dự báo cũng được so sánh trực quan với dữ liệu thực tế.

1.4. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU

Đối tượng nghiên cứu của đề tài là dữ liệu bán hàng theo hình thức thương mại điện tử, cụ thể là các giao dịch được ghi nhận trong bộ dữ liệu Online Retail II, được công bố bởi UCI Machine Learning Repository. Dữ liệu này bao gồm các thông tin chi tiết như mã hóa đơn, mã sản phẩm, mô tả mặt hàng, số lượng bán ra, giá đơn vị, ngày giờ giao dịch, mã khách hàng và quốc gia. Đối tượng trọng tâm mà nghiên cứu hướng đến là chuỗi thời gian doanh số, tức tổng giá trị tiền hàng được ghi nhận theo từng mốc thời gian cụ thể. Việc phân tích và dự báo tập trung vào biến động của doanh số theo ngày hoặc theo tháng nhằm hỗ trợ ra quyết định trong quản lý chuỗi cung ứng, lập kế hoạch bán hàng và kiểm soát tồn kho.

Phạm vi nghiên cứu được giới hạn trên nhiều khía cạnh nhằm đảm bảo tính khả thi và phù hợp với nguồn lực hiện có:

- Về thời gian: Phân tích dữ liệu từ tháng 1/2010 đến tháng 11/2011 – đủ dài để nhận diện xu hướng và mùa vụ.

- Về dữ liệu: Chỉ giữ lại các giao dịch hợp lệ (số lượng và đơn giá dương), loại bỏ hóa đơn bị hủy và các bản ghi thiếu thông tin.
- Về cấp độ phân tích: Không đi sâu vào từng khách hàng mà phân tích tổng doanh số theo mốc thời gian.
- Về biến đầu vào: Mô hình tập trung vào chuỗi thời gian đơn biến (tổng doanh số), không xét đến các yếu tố ngoại sinh như khuyến mãi hoặc marketing.
- Về sản phẩm/khu vực: Trong bước EDA có thể phân tích theo quốc gia hoặc nhóm sản phẩm, nhưng mô hình dự báo sẽ áp dụng ở cấp độ tổng doanh số.

Việc giới hạn phạm vi như trên giúp đảm bảo tính khả thi và tập trung nguồn lực vào việc phát triển mô hình dự báo có độ chính xác cao, ứng dụng thực tiễn trong hoạch định kinh doanh.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

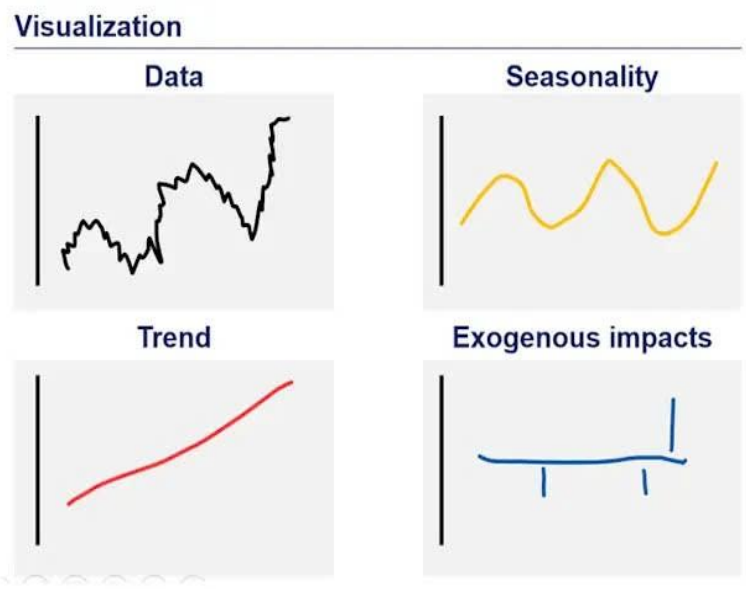
2.1. DỰ BÁO CHUỖI THỜI GIAN (TIME SERIES FORECASTING)

Dự báo chuỗi thời gian là quá trình xây dựng mô hình toán học nhằm ước lượng giá trị tương lai của một đại lượng biến thiên theo thời gian, dựa trên những quan sát trong quá khứ. Chuỗi thời gian thường thể hiện các yếu tố như xu hướng (trend), tính mùa vụ (seasonality), và nhiễu loạn ngẫu nhiên (random noise). Mục tiêu của phân tích chuỗi thời gian là bóc tách và mô hình hóa những thành phần này để đưa ra dự đoán chính xác, đặc biệt hữu ích trong các lĩnh vực tài chính, thương mại điện tử, sản xuất và hậu cần [1].

Trong bối cảnh thương mại điện tử, việc dự đoán doanh số bán hàng theo thời gian đóng vai trò then chốt trong việc quản lý tồn kho, lập kế hoạch sản xuất và tối ưu hoá hoạt động marketing. Một chuỗi thời gian tốt cần được kiểm tra tính dừng (stationarity), nhận diện xu hướng hoặc mùa vụ, từ đó chọn lựa mô hình thích hợp để dự báo [2].

2.2. CÁC PHƯƠNG PHÁP DỰ BÁO CHUỖI THỜI GIAN

2.2.1. Mô hình Prophet



Hình 2 - 1: *Mô tả mô hình Prophet*

Prophet là một mô hình dự báo chuỗi thời gian được phát triển bởi Facebook, được thiết kế để xử lý các dữ liệu có tính mùa vụ mạnh, xu hướng phi tuyến và chứa nhiều điểm dị thường (outliers). Prophet được thiết kế để dễ sử dụng, khả năng tùy chỉnh cao và phù hợp với dữ liệu bán lẻ có tính chất phi tuyến và nhiều dị thường. Prophet giả định chuỗi thời gian bao gồm ba thành phần chính:

- Xu hướng: Mô hình hóa sự thay đổi dài hạn của chuỗi thời gian. Prophet hỗ trợ cả xu hướng tuyến tính và logistic.
- Mùa vụ: Mô hình hóa các thay đổi mang tính chu kỳ như theo tuần, tháng, năm.
- Sự kiện đặc biệt (Holidays): Cho phép người dùng chỉ định các dịp đặc biệt như lễ hội, khuyến mãi để mô hình hóa ảnh hưởng đến doanh số.

Công thức tổng quát của Prophet có dạng:

$$y(t)=g(t)+s(t)+h(t)+\varepsilon_t \qquad (1)$$

Trong đó:

- y(t): giá trị quan sát tại thời điểm ttt
- g(t): thành phần xu hướng
- s(t): thành phần mùa vụ
- h(t): hiệu ứng từ sự kiện đặc biệt
- ε_t: nhiễu trắng (white noise)

Ưu điểm:

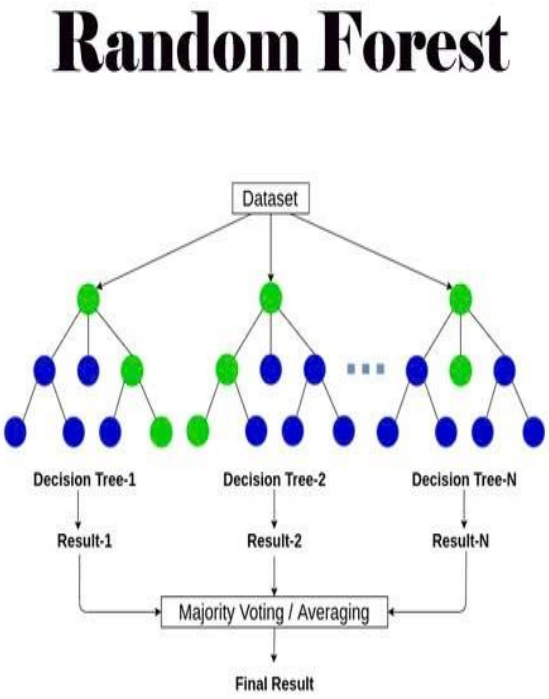
- Dễ sử dụng: Yêu cầu tối thiểu dữ liệu đầu vào với hai cột ds (thời gian) và y (giá trị).
- Tối ưu cho dữ liệu kinh doanh: Xử lý tốt các chuỗi có xu hướng thay đổi và mùa vụ phi tuyến.
- Tự động hóa cao:
  - Phát hiện điểm thay đổi xu hướng (changepoints)
  - Xử lý dữ liệu bị thiếu (missing values)
  - Mô hình hóa mùa vụ phi tuyến thông qua Fourier series
- Tùy chỉnh linh hoạt:
  - Thêm ngày lễ theo khu vực
  - Điều chỉnh độ nhạy phát hiện changepoint
  - Cấu hình riêng cho từng loại mùa vụ (tuần, tháng, năm)

Nhược điểm:

- Không xử lý tốt chuỗi đa biến (multivariate): Không thể tận dụng được thông tin bổ sung từ nhiều biến đầu vào.
- Giới hạn trong mô hình hóa phi tuyến phức tạp: Không phù hợp với các mối quan hệ phi tuyến hoặc có tính tương tác cao.

- Cần tinh chỉnh thủ công: Để đạt hiệu suất cao nhất, người dùng cần hiểu rõ các thành phần mô hình để điều chỉnh phù hợp [3].

2.2.2. Mô hình Random Forest



Hình 2 - 2: *Mô tả mô hình Random Foresst*

Random Forest là một phương pháp học máy thuộc nhóm mô hình ensemble (tổ hợp), hoạt động dựa trên nguyên lý xây dựng tập hợp nhiều cây quyết định (decision trees) và kết hợp kết quả của chúng để cải thiện độ chính xác dự báo. Trong bài toán dự báo doanh số, Random Forest có khả năng khai thác hiệu quả các đặc trưng đa chiều như thời gian (ngày, tháng, quý), số lượng đặt hàng, quốc gia, đơn giá và các đặc trưng được rút trích từ hành vi mua sắm, nhằm dự báo tổng doanh số trong tương lai.

Khác với hồi quy tuyến tính, Random Forest không yêu cầu giả định mối quan hệ tuyến tính giữa các biến đầu vào và biến mục tiêu, đồng thời có khả năng mô hình hóa tốt các mối quan hệ phi tuyến và tương tác phức tạp giữa các biến.

Random Forest được xây dựng bằng cách huấn luyện một tập hợp gồm nhiều cây quyết định trên các tập con ngẫu nhiên khác nhau của dữ liệu huấn luyện (kỹ thuật bagging), đồng thời tại mỗi nút phân chia, mô hình chỉ xem xét một tập con ngẫu nhiên của các đặc trưng. Dự báo cuối cùng là kết quả trung bình (trong bài toán hồi quy) của tất cả các cây con trong rừng.

Công thức tổng quát:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \qquad (2)$$

Trong đó:

- T: số lượng cây trong rừng.
- $h_t(x)$: kết quả dự báo của cây thứ t.
- $\hat{y}$: giá trị doanh số được dự báo trung bình từ tất cả các cây.

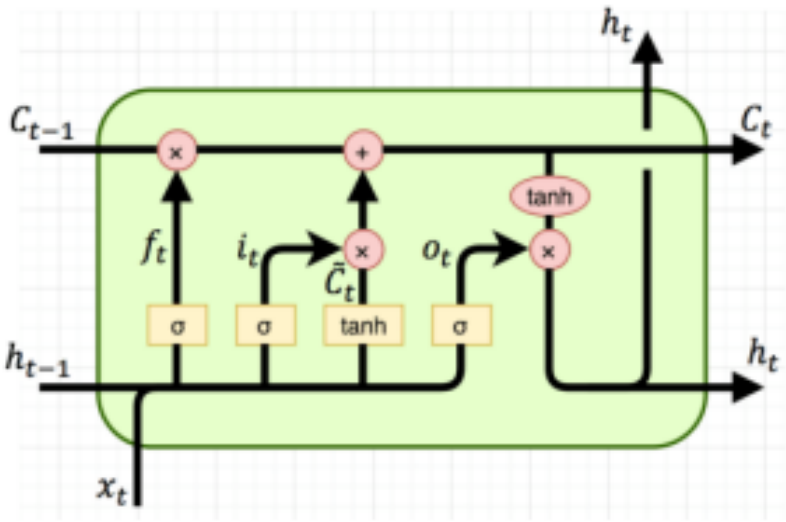
Ưu điểm:

- Mô hình hóa tốt quan hệ phi tuyến:
- Có khả năng nắm bắt các mối quan hệ phức tạp và phi tuyến giữa các biến đầu vào.
- Chống overfitting tốt: Nhờ việc kết hợp nhiều cây và áp dụng kỹ thuật bagging, Random Forest thường có độ tổng quát hóa cao hơn so với các mô hình cây đơn lẻ.
- Không yêu cầu chuẩn hóa dữ liệu: Mô hình không nhạy cảm với việc biến đổi tỷ lệ giữa các đặc trưng.
- Tự động đánh giá mức độ quan trọng của đặc trưng: Giúp nhà phân tích hiểu rõ yếu tố nào ảnh hưởng lớn đến kết quả dự báo.
- Ổn định trước dữ liệu nhiễu và ngoại lệ: Ít bị ảnh hưởng bởi outliers hoặc dữ liệu thiếu chính quy.

Nhược điểm:

- Thiếu tính diễn giải: Khó giải thích mối quan hệ giữa đầu vào và đầu ra do bản chất của mô hình tổ hợp.
- Chi phí tính toán cao: Cần nhiều tài nguyên hơn để huấn luyện và dự báo khi số lượng cây lớn hoặc dữ liệu có kích thước lớn[4].
- Không phù hợp với dữ liệu có thứ tự thời gian nếu không xử lý đặc trưng phù hợp: Mô hình không nắm bắt được thông tin tuần tự nếu không có đặc trưng chuỗi thời gian rõ ràng như “tháng”, “ngày trong tuần”, “quý”, v.v.

2.2.3. Mô hình LSTM (Long Short-Term Memory)



Hình 2 - 3: *Mô tả mô hình LSTM*

LSTM là một biến thể của mạng nơ-ron hồi tiếp (Recurrent Neural Networks - RNN), được thiết kế để giải quyết bài toán gradient biến mất và lưu giữ thông tin trong chuỗi dài. LSTM rất phù hợp với các bài toán có tính phụ thuộc theo thời gian dài hạn như chuỗi doanh số theo tuần/tháng.

Kiến trúc của một đơn vị LSTM được mô tả qua các phương trình sau:

$$f_t = \sigma(Wf \cdot [h_{t-1}, x_t] + b_f)$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$\widetilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (3)$$
$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t}$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$
$$h_t = o_t * \tanh(C_t)$$

Trong đó:

- xt: đầu vào tại thời điểm ttt
- ht: đầu ra ẩn
- Ct: bộ nhớ (cell state)
- σ: hàm sigmoid
- *: phép nhân phần tử

Ưu điểm:

- Khả năng học phi tuyến sâu: Mô hình hóa được các mối quan hệ phi tuyến phức tạp mà các mô hình thống kê không xử lý được.
- Ghi nhớ dài hạn: Có khả năng học và lưu giữ thông tin từ nhiều bước thời gian trước đó.

- Linh hoạt: Phù hợp cho cả dữ liệu chuỗi đơn biến (univariate) và đa biến (multivariate).
- Không yêu cầu kỹ thuật đặc trưng hóa: Trích xuất đặc trưng tự động từ dữ liệu gốc, không cần xử lý thủ công nhiều.

Nhược điểm:

- Yêu cầu dữ liệu lớn: Cần nhiều mẫu dữ liệu để huấn luyện hiệu quả và tránh overfitting.
- Khó huấn luyện và tối ưu: Việc lựa chọn siêu tham số như look_back, số tầng LSTM, dropout, batch_size, v.v., ảnh hưởng lớn đến kết quả.
- Tốn tài nguyên tính toán: Thời gian huấn luyện dài hơn so với các mô hình thống kê.
- Giải thích hạn chế: Là một "hộp đen" (black-box), khó giải thích so với các mô hình tuyến tính.

Với dữ liệu Online Retail II được tổng hợp thành chuỗi doanh số theo tháng hoặc tuần, LSTM có thể học các mô hình phức tạp để đưa ra dự báo chính xác [5].

2.2.4. Các chỉ số đánh giá mô hình

2.2.4.1. Sai số trung bình tuyệt đối (Mean Absolute Error - MAE):

Sai số trung bình tuyệt đối là một thang đo được sử dụng phổ biến trong đánh giá mô hình hồi quy. Nó đo lường sai số trung bình giữa giá trị thực tế (y_i) và giá trị dự đoán ($\hat{y}_i$), bằng cách lấy giá trị tuyệt đối của chênh lệch đó. Công thức tính MAE được biểu diễn như sau:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \qquad (4)$$

Ưu điểm của Sai số tuyệt đối: 1 Đơn giản và dễ hiểu: MAE trực tiếp biểu thị sai số trung bình bằng đơn vị của biến dự đoán, giúp dễ dàng giải thích. 46 Ít nhạy cảm với ngoại lệ: Vì không bình phương sai số, MAE ít bị ảnh hưởng bởi các giá trị ngoại lệ hơn so với MSE hoặc RMSE.

Nhược điểm của Sai số tuyệt đối: Không phạt nặng sai số lớn: MAE không khuếch đại ảnh hưởng của sai số lớn, có thể không phù hợp nếu ta muốn tập trung vào việc giảm thiểu các sai số nghiêm trọng.

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

Căn bậc hai của sai số trung bình bình phương là một thước đo phổ biến để đánh giá chất lượng của mô hình hồi quy. RMSE là căn bậc hai của sai số trung bình bình phương (MSE), giúp đưa đơn vị của thước đo này trở về cùng đơn vị với dữ liệu thực tế, giúp trực quan hóa dễ dàng hơn. RMSE mang lại trọng số cao hơn cho các sai số lớn, làm cho nó đặc biệt hữu ích trong các bài toán đòi hỏi độ chính xác cao.

Công thức tính RMSE:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)} \quad (5)$$

Trong đó:

- n: Số lượng điểm dữ liệu.
- y_i : Giá trị thực tế.
- $\hat{y}_i$: Giá trị dự đoán.

Ưu điểm của RMSE: Dễ diễn giải: Do có cùng đơn vị với dữ liệu đầu ra, RMSE dễ hiểu hơn so với MSE. Phạt sai số lớn: RMSE nhạy cảm với các ngoại lệ, giúp nhận diện và điều chỉnh các dự đoán sai lệch nghiêm trọng.

Nhược điểm của RMSE: Nhạy cảm với ngoại lệ: Giống như MSE, RMSE có thể bị ảnh hưởng lớn bởi các giá trị ngoại lệ, làm sai lệch kết quả.

2.2.4.3. Hệ số xác định (R^2 Score)

R^2 Score còn được gọi là hệ số xác định của mô hình, đo lường mức độ giải thích được của mô hình đối với biến mục tiêu. R^2 được định nghĩa như tỷ lệ giữa tổng phương sai được mô hình giải thích và tổng phương sai tổng thể trong dữ liệu. Công thức tính R^2:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SS_{RES}}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

Trong đó:

- SSR: Tổng bình phương hồi quy, đo lường phần biến động của biến phụ thuộc được giải thích bởi mô hình.
- SSE : Tổng bình phương sai số, đo lường sự khác biệt giữa giá trị dự đoán của mô hình và giá trị thực tế.
- SST: Tổng bình phương toàn phần, đo lường tổng biến động của biến phụ thuộc quanh giá trị trung bình của nó.
- y_i : Giá trị thực tế.
- $\hat{y}_i$: Giá trị dự đoán.

Ý nghĩa của R^2:

- $R^2=1$: Mô hình giải thích hoàn toàn phương sai của dữ liệu.
- $R^2=0$: Mô hình không giải thích được gì, tương đương với dự đoán giá trị trung bình y.
- $R^2<0$: Mô hình tệ hơn việc dự đoán giá trị trung bình.

Việc sử dụng kết hợp các chỉ số này giúp đảm bảo đánh giá toàn diện về độ chính xác và tính ổn định của mô hình dự báo [6].

CHƯƠNG 3. THỰC NGHIỆM

Các đoạn code giải quyết bài toán được trình bày tại trang, đường link: https://github.com/Tung0004/DO-AN-2---Du-doan-doanh-so-ban-hang-trong-tuong-lai-tu-du-lieu-lich-su-giao-dich

3.1. THU THẬP VÀ KHÁM PHÁ DỮ LIỆU

3.1.1. Tải và kiểm tra dữ liệu ban đầu

```
===== THÔNG TIN CƠ BẢN =====
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1067371 entries, 0 to 1067370
Data columns (total 8 columns):
 #   Column       Non-Null Count    Dtype
---  ------       --------------    -----
 0   Invoice      1067371 non-null  object
 1   StockCode    1067371 non-null  object
 2   Description  1062989 non-null  object
 3   Quantity     1067371 non-null  int64
 4   InvoiceDate  1067371 non-null  datetime64[ns]
 5   Price        1067371 non-null  float64
 6   Customer ID  824364 non-null   float64
 7   Country      1067371 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 65.1+ MB
None
```

Bộ dữ liệu này gồm 1,067,371 dòng và 8 cột, với dung lượng khoảng 65 MB, phản ánh chi tiết các giao dịch bán hàng. Các cột dữ liệu bao gồm: mã hóa đơn (Invoice), mã sản phẩm (StockCode), mô tả sản phẩm (Description), số lượng bán ra (Quantity), thời gian giao dịch (InvoiceDate), giá sản phẩm (Price), mã khách hàng (Customer ID) và quốc gia (Country). Các kiểu dữ liệu (dtypes) nhìn chung phù hợp, đặc biệt InvoiceDate đã ở định dạng datetime64[ns], thuận lợi cho các bài toán chuỗi thời gian. Tuy nhiên, bộ dữ liệu có một số vấn đề về chất lượng: giá trị ở cột Description và Customer ID bị thiếu. Điều này có thể ảnh hưởng đến các phân tích liên quan đến sản phẩm hoặc khách hàng nếu không xử lý thích hợp. Nhìn chung, bộ dữ liệu có quy mô lớn, đa dạng, phù hợp cho nhiều bài toán phân tích dữ liệu, dự báo bán hàng hoặc khai thác hành vi người tiêu dùng.

3.1.2. Kiểm tra chất lượng dữ liệu

3.1.2.1. Kiểm tra giá trị thiếu

```
===== KIỂM TRA GIÁ TRỊ THIẾU =====
             Số lượng  Phần trăm (%)
Description      4382       0.410541
Customer ID    243007      22.766873
```

Bộ dữ liệu có hai cột chứa giá trị thiếu đáng chú ý. Cột Description thiếu 4,382 dòng, chiếm khoảng 0.41% tổng số dữ liệu, tỷ lệ này khá nhỏ và nhìn chung sẽ không ảnh hưởng lớn đến phân tích nếu được xử lý phù hợp (ví dụ như loại bỏ hoặc thay thế). Tuy nhiên, cột Customer ID lại thiếu tới 243,007 dòng, tương đương khoảng 22.77% dữ liệu. Đây là tỷ lệ thiếu rất cao và có thể gây ảnh hưởng nghiêm trọng đến các phân tích liên quan đến hành vi khách hàng, phân khúc khách hàng hoặc xây dựng mô hình dự đoán theo cá nhân. Vì vậy, cần đặc biệt lưu ý khi xử lý biến này: có thể phải cân nhắc loại bỏ các dòng thiếu hoặc áp dụng kỹ thuật suy đoán ID nếu có thể.

3.1.2.2. Thống kê mô tả cho biến số

Bảng 3 - 1: *Bảng thống kê mô tả biến số*

	Quantity	**Price**
count	1.067371e+06	1.067371e+06
mean	9.938898e+00	4.649388e+00
std	1.727058e+02	1.235531e+02
min	-8.099500e+04	-5.359436e+04
25%	1.000000e+00	1.250000e+00
50%	3.000000e+00	2.100000e+00
75%	1.000000e+01	4.150000e+00
max	8.099500e+04	3.897000e+04

Phân tích thống kê cho thấy biến Quantity có trung bình khoảng 9.93 sản phẩm mỗi giao dịch, nhưng độ lệch chuẩn rất cao (172.70), phản ánh sự phân tán dữ liệu mạnh và khả năng tồn tại nhiều giá trị ngoại lai. Giá trị Quantity dao động từ -80,995 đến

80,995, trong đó giá trị âm thể hiện các giao dịch hoàn trả hoặc lỗi nhập liệu cần được kiểm tra kỹ lưỡng. Đa số các giao dịch có số lượng sản phẩm nhỏ, thể hiện qua các mốc tứ phân vị: 25% dữ liệu có Quantity ≤ 1, 50% ≤ 3 và 75% ≤ 10.

Về biến Price, giá trung bình mỗi sản phẩm là khoảng 4.64 đơn vị tiền tệ, với độ lệch chuẩn lớn (123.55), cho thấy sự chênh lệch rất lớn giữa các mức giá sản phẩm. Các mức giá dao động từ -53,594 đến 38,970, trong đó giá trị âm không hợp lý với bản chất của giá sản phẩm, có thể do lỗi dữ liệu hoặc liên quan tới các nghiệp vụ trả hàng. Các giá trị phân vị cho thấy phần lớn sản phẩm có giá khá thấp: 25% sản phẩm có giá dưới 1.25, 50% dưới 2.1 và 75% dưới 4.15 đơn vị tiền tệ, cho thấy phần lớn sản phẩm giao dịch có giá trị nhỏ, trong khi một số sản phẩm cao cấp kéo theo sự biến động lớn về giá.

Tóm lại, cả hai biến Quantity và Price đều có sự phân tán rất mạnh, nhiều ngoại lệ, và có dấu hiệu cần xử lý bất thường (đặc biệt là các giá trị âm) để đảm bảo chất lượng phân tích sau này.

3.1.2.3. Kiểm tra giá trị trùng lặp

```
===== GIÁ TRỊ TRÙNG LẶP =====
Tổng số dòng trùng lặp: 34335
```

Quan sát cho thấy có nhiều dòng trong bộ dữ liệu có cùng Invoice, InvoiceDate, Customer ID, và Country. Tuy nhiên, các dòng này lại có StockCode và Description khác nhau. Điều này cho thấy các giá trị trùng ở đây là hợp lý, bởi một hóa đơn (Invoice) thường bao gồm nhiều sản phẩm khác nhau được mua trong cùng một giao dịch. Đây là hiện tượng "nhiều dòng cho một hóa đơn" (multi-line invoice), rất phổ biến trong dữ liệu bán lẻ. Không có dấu hiệu về trùng lặp hoàn toàn, vì các dòng khác nhau ít nhất ở StockCode, Description, hoặc Quantity.

3.1.2.4 Kiểm tra dữ liệu không hợp lệ

```
======Dữ liệu không hợp lệ======
Tổng: 19494
```

Có 19,494 giá trị không hợp lệ gồm những sản phẩm có cột "Quantity", "Price" < 0 và cột "Invoice" bị hủy

3.1.2.5. Kiểm tra phân phối theo thời gian

```
('2009-12-01 07:45:00', '2011-12-09 12:50:00')
```

Khoảng thời gian giữa hai mốc thời gian, từ ngày 1 tháng 12 năm 2009 lúc 07:45 sáng đến ngày 9 tháng 12 năm 2011 lúc 12:50 chiều, kéo dài tổng cộng 2 năm, 8 ngày

3.2. TIỀN XỬ LÝ DỮ LIỆU

3.2.1. Làm sạch dữ liệu

Đầu tiên, cột "InvoiceDate" của DataFrame sang dạng datetime thông qua hàm `df.to_datetime()` để đưa cột "InvoiceDate" về dạng thời gian tiêu chuẩn: YYYY-MM-DD HH:MM:SS. Điều này giúp dữ liệu về thời gian trở nên nhất quán và dễ dàng xử lý, chẳng hạn như trích xuất các thành phần năm, tháng, ngày, giờ, phút và giây từ mỗi giá trị trong cột

Do dữ liệu trong cột "InvoiceDate" trải dài từ tháng 12/2009 đến tháng 12/2011, tuy nhiên hai tháng đầu và cuối (12/2009 và 12/2011) không đầy đủ số ngày, điều này có thể gây sai lệch trong quá trình huấn luyện mô hình. Vì vậy, loại bỏ toàn bộ dữ liệu tương ứng với hai tháng này để đảm bảo chất lượng dữ liệu đầu vào.

Đối với bộ dữ liệu có giá trị thiếu ở hai cột "Customer" và "Description", việc xử lý các giá trị thiếu này là rất quan trọng để đảm bảo tính chính xác của mô hình dự đoán doanh số. Cột "Customer ID" không đóng vai trò quan trọng trong quá trình dự đoán và không ảnh hưởng nhiều đến kết quả cuối cùng, vì vậy sẽ được loại bỏ khỏi bộ dữ liệu để làm sạch và đơn giản hóa quá trình xử lý. Đối với cột "Description", thay vì loại bỏ các dòng dữ liệu có giá trị thiếu, chúng ta sẽ thay thế các giá trị thiếu bằng từ "Unknown". Cách làm này giúp giữ lại toàn bộ số lượng dữ liệu mà không làm giảm thông tin hữu ích, đồng thời vẫn đảm bảo tính toàn vẹn và khả năng phân tích dữ liệu.

Sử dụng phương pháp IQR với phân vị 1% và 99% xác định ngưỡng trên ngưỡng, dưới, sau đó các giá trị vượt quá ngưỡng sẽ được thay thế bằng chính các ngưỡng này nhằm giảm ảnh hưởng tiêu cực đến quá trình phân tích hoặc huấn luyện mô hình. Kết quả là một bộ dữ liệu sạch, ổn định.

- Loại bỏ các giao dịch bị hủy được nhận diện qua mã hóa đơn cột " Invoice " chứa ký tự "C"
- Loại bỏ các bản ghi có số lượng và đơn giá không hợp lệ, cụ thể là những dòng có Quantity ≤ 0 hoặc Price ≤ 0
- Tiến hành kiểm tra lại giá trị thiếu sau khi xử lý:

```
Invoice        0
StockCode      0
Description    0
Quantity       0
```

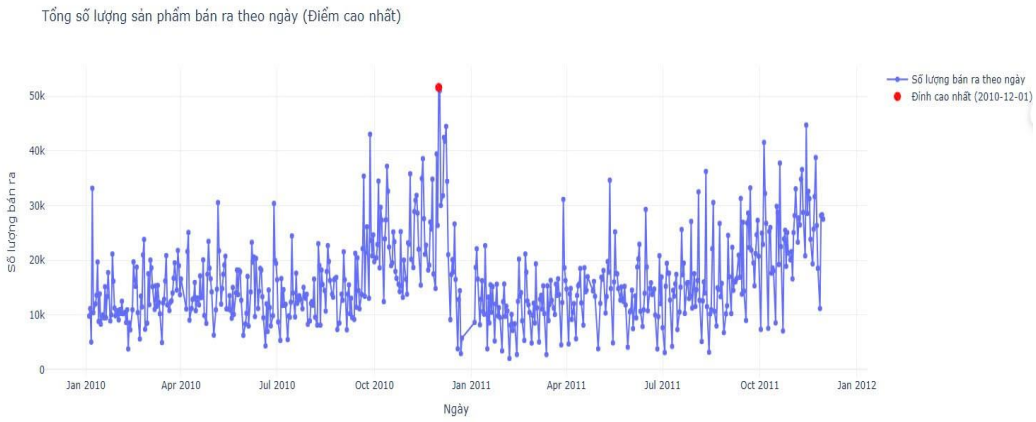
```
InvoiceDate    0
Price          0
Country        0
dtype: int64
```

3.2.2 Tạo các đặc trưng mới

- Date: Lấy phần ngày từ InvoiceDate, loại bỏ giờ, phút, giây — phục vụ cho phân tích theo từng ngày cụ thể
- Month: Chuyển InvoiceDate sang định dạng chu kỳ tháng (Period), giúp nhóm dữ liệu và phân tích theo tháng.
- Month_int: Trích xuất số thứ tự của tháng (1–12), hỗ trợ phân tích theo mùa vụ trong năm.
- Quarter: Xác định quý của năm (từ 1 đến 4) mà giao dịch diễn ra, phục vụ cho phân tích xu hướng theo quý.
- Year: Trích xuất năm từ ngày giao dịch, hỗ trợ theo dõi sự thay đổi hoặc xu hướng theo năm.
- DayofWeek: Lấy tên ngày trong tuần (ví dụ: Monday, Tuesday…), dùng để phân tích hành vi mua hàng theo ngày.
- Hour: Lấy khung giờ giao dịch (0–23), giúp phân tích thời điểm trong ngày có lượng mua sắm cao.
- TotalAmount: Tính tổng doanh thu của mỗi dòng giao dịch bằng cách nhân Quantity với Price, dùng để đánh giá doanh số.

3.3. EDA - PHÂN TÍCH KHÁM PHÁ DỮ LIỆU

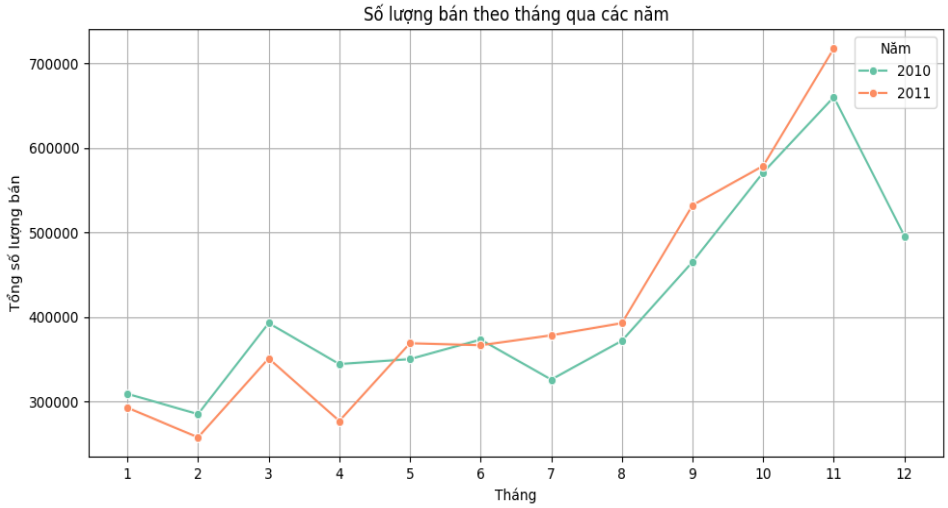
3.3.1. Phân tích doanh số theo ngày



Hình 3 - 1: *Tổng số lượng sản phẩm bán ra theo ngày*

- Xu hướng chung tăng dần theo thời gian
  - Từ đầu năm 2010 đến cuối năm 2011, mức độ biên động của doanh số ngày càng lớn, với nhiều đỉnh và đáy rõ rệt. Điều này cho thấy quy mô thị trường và khối lượng giao dịch có xu hướng gia tăng, có thể do sự mở rộng khách hàng, quảng bá thương hiệu hay các chiến dịch khuyến mãi.
- Tính mùa vụ (seasonality)
  - Đa số các ngọn đỉnh cao nhất tập trung vào các tháng cuối năm (cuối quý 3, quý 4), đặc biệt quanh tháng 9–12 của năm 2010 và 2011. Ví dụ:
  - Ngày 2/12/2010 và 1/12/2010 là hai ngày có doanh số lớn nhất, có thể liên quan đến các chương trình khuyến mãi mùa đông hoặc chuẩn bị cho lễ Giáng Sinh.
  - Ngày 27/9/2010 và 8/12/2010 cũng nằm trong nhóm "đỉnh" cuối năm, càng khẳng định mùa cao điểm mua sắm rơi vào thời gian này.
- Các điểm bất thường (outliers)
  - Một vài ngày giữa năm cũng có đột biến (ví dụ ngày 27/9/2010 hay ngày 10/10/2010), có thể do chiến dịch marketing đặc biệt, sự kiện flash sale hoặc đơn hàng lớn bị gộp vào một ngày. Biên độ dao động lớn
  - Cột "Số lượng bán ra theo ngày" có khuynh hướng từ khoảng 5000–10000 mặt hàng/ngày ở giai đoạn đầu 2010, đến trên 20000 – 30000/ngày vào cuối 2011. Biên độ và tần suất các đỉnh càng dày đặc cho thấy hoạt động kinh doanh ngày càng sôi động.

3.3.2. Phân tích doanh số theo tháng



Hình 3 - 2: *Tổng số sản phẩm bán ra theo từng tháng*

Giai đoạn ổn định ban đầu (01/2010 – 04/2010): Số lượng bán dao động nhẹ, nằm trong khoảng 280,000 – 300,000 đơn vị. Giai đoạn này thể hiện mức tiêu thụ ổn định, không có đột biến đáng kể. Có thể đây là giai đoạn chưa có chiến dịch tiếp thị lớn, hoặc đang trong mùa thấp điểm.

Giai đoạn dao động nhẹ và phục hồi (05/2010 – 08/2010): Tháng 5/2010 ghi nhận mức tăng nhẹ (~ 350,000), sau đó tiếp tục tăng đến khoảng 370,000 vào tháng 6, rồi giảm xuống mức thấp hơn vào tháng 7 (~320,000).- Sự dao động này gợi ý tính chu kỳ hoặc ảnh hưởng của các chiến dịch bán hàng ngắn hạn (ví dụ: khuyến mãi mùa hè), nhưng chưa có xu hướng rõ ràng.

Giai đoạn tăng trưởng mạnh (09/2010 – 11/2010): Tháng 9–11/2010 chứng kiến một sự gia tăng rõ rệt, từ ~460,000 lên ~600,000 đơn vị trong 3 tháng liên tiếp. Đây là điểm bùng nổ doanh số, có thể do các chương trình khuyến mãi lớn, nhu cầu mua sắm cao dịp cuối năm hoặc sự ra mắt sản phẩm mới. Đây là mùa cao điểm đầu tiên trong chuỗi thời gian.

Sự suy giảm sau cao điểm (12/2010 – 02/2011): Tháng 12/2010 và đầu năm 2011 chứng kiến sự giảm mạnh về lượng bán (~430,000 vào tháng 12, xuống khoảng 260,000 vào tháng 2/2011). Đây có thể là hiệu ứng tất yếu sau mùa cao điểm, đồng thời phản ánh tính mùa vụ thường thấy: sau dịp lễ người tiêu dùng chi tiêu ít lại.

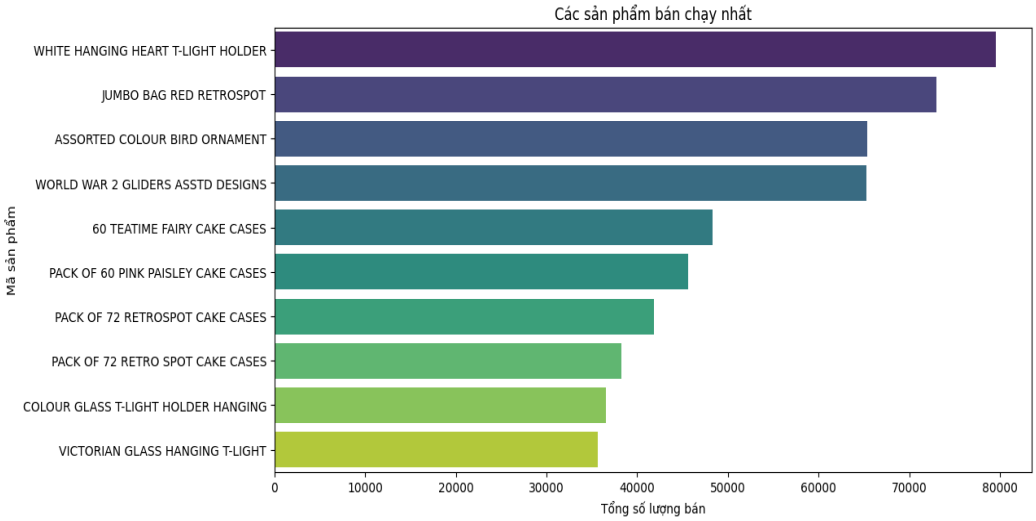
Phục hồi và ổn định giữa năm (03/2011 – 08/2011): Từ tháng 3 đến tháng 8/2011, lượng bán dao động từ 330,000 đến 380,000 đơn vị. Mặc dù không tăng đột biến, nhưng duy trì mức tiêu thụ ổn định cho thấy thị trường đã phục hồi và chuẩn bị bước vào mùa cao điểm tiếp theo.

Tăng trưởng đột phá lần thứ hai (09/2011 – 11/2011): Tăng vọt từ ~520,000 vào tháng 9 lên tới đỉnh 640,000 vào tháng 11/2011. Đáng chú ý là mức tăng trưởng này mạnh hơn và sớm hơn so với cùng kỳ năm 2010, cho thấy năm 2011 đã đạt được kết quả kinh doanh vượt trội. Điều này phản ánh sự cải thiện trong chiến lược bán hàng, năng lực cung ứng hoặc sự mở rộng thị trường.

Kết luận: Cần định kỳ hóa các chiến dịch khuyến mãi vào (tháng 9–11), vì đây là giai đoạn có tiềm năng bán hàng cao nhất mỗi năm. Tạo các chương trình giữ chân khách hàng sau dịp lễ, đặc biệt trong tháng 12 và các tháng đầu năm sau – nơi mức tiêu thụ sụt giảm mạnh. Năm 2011 có xu hướng tăng trưởng tốt hơn năm 2010, cho thấy các điều chỉnh chiến lược trong năm này có hiệu quả – cần phân tích sâu yếu tố thành công

để nhân rộng. Tối ưu tồn kho & cung ứng từ tháng 8 để chuẩn bị cho nhu cầu cao từ tháng 9 trở đi.

3.3.3 Phân tích doanh số theo sản phẩm



Hình 3 - 3: *Top 10 các sản phẩm bán chạy nhất*

Biểu đồ thể hiện top 10 sản phẩm bán chạy nhất của nhà bán lẻ trực tuyến theo tổng số lượng bán ra. Phân tích mô tả sản phẩm cho thấy phần lớn các mặt hàng thuộc về hai nhóm chính: đồ trang trí và dụng cụ làm bánh.

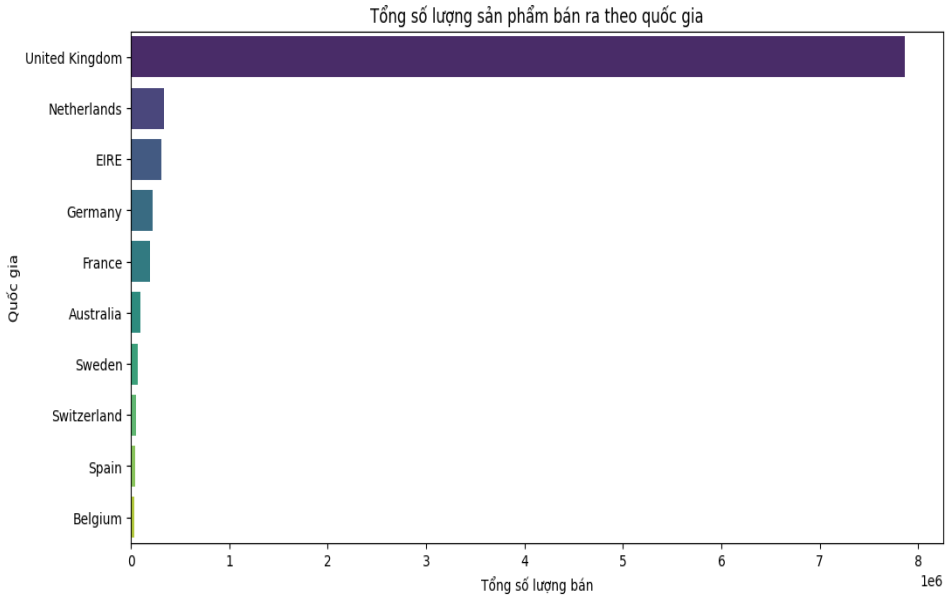
Nhóm đồ trang trí gồm các sản phẩm như WHITE HANGING HEART T-LIGHT HOLDER, ASSORTED COLOUR BIRD ORNAMENT và COLOUR GLASS T-LIGHT HOLDER HANGING. Các sản phẩm này thường mang yếu tố thẩm mỹ, được thiết kế để trang trí nhà cửa, không gian tiệc hoặc làm quà tặng. Việc các sản phẩm này xuất hiện nhiều trong danh sách bán chạy cho thấy xu hướng tiêu dùng thiên về nhu cầu làm đẹp không gian sống, đặc biệt trong các dịp lễ, tiệc tại gia.

Nhóm thứ hai là dụng cụ làm bánh, với các sản phẩm như PACK OF 60 PINK PAISLEY CAKE CASES, PACK OF 72 RETROSPOT CAKE CASES, 60 TEATIME FAIRY CAKE CASES. Đây là các mặt hàng giá trị nhỏ, tiêu hao theo lần sử dụng và thường được mua theo lô. Sự xuất hiện của nhóm này phản ánh hành vi tiêu dùng hướng đến tổ chức tiệc tại nhà, nấu nướng cá nhân hoặc kinh doanh hộ gia đình nhỏ lẻ.

Ngoài ra, một số sản phẩm thuộc các nhóm riêng lẻ như túi đựng (JUMBO BAG RED RETROSPOT) và bộ đồ thủ công (MINI PAINT SET VINTAGE) cho thấy có sự hiện diện của các mặt hàng phục vụ sinh hoạt và giải trí nhẹ nhàng. Sự đa dạng của các nhóm sản phẩm trong top bán chạy, tuy nhiên, vẫn xoay quanh nhu cầu cá nhân hóa không gian sống và hoạt động thủ công mang tính cá nhân.

Tóm lại, biểu đồ cho thấy khách hàng tập trung vào các sản phẩm nhỏ, giá trị thấp, mang tính thẩm mỹ và tiện lợi cao. Nhóm đối tượng mục tiêu có thể là phụ nữ, gia đình trẻ, hoặc các cá nhân yêu thích trang trí và hoạt động thủ công tại nhà. Đây là những insight quan trọng có thể định hướng chiến lược tiếp thị và quản lý hàng tồn kho trong các giai đoạn cao điểm nhu cầu như lễ hội, mùa cưới, hoặc Giáng sinh.

3.3.4 Phân tích doanh số theo quốc gia



Hình 3 - 4: *Tổng số lượng sản phẩm bán ra theo quốc gia*

- United Kingdom (Anh) vượt trội hoàn toàn so với các quốc gia khác:
    - Với tổng số lượng bán ra gần 8 triệu đơn vị, UK chiếm phần lớn hoạt động bán lẻ trong toàn bộ cơ sở dữ liệu.
    - Điều này không bất ngờ vì công ty trong bộ dữ liệu có trụ sở tại UK, do đó lượng giao dịch nội địa thường nhiều hơn và ít chi phí vận chuyển hơn.
- Khoảng cách rất lớn giữa UK và phần còn lại:
    - Quốc gia đứng thứ hai là Netherlands với lượng bán chỉ khoảng vài trăm nghìn đơn vị — chênh lệch hơn 20 lần so với UK.
    - Điều này cho thấy thị trường chính của công ty tập trung gần như tuyệt đối ở trong nước.
- Nhóm các quốc gia châu Âu như Netherlands, EIRE (Ireland), Germany, France chiếm các vị trí tiếp theo:
    - Đây là các quốc gia gần địa lý với UK, có chi phí vận chuyển thấp và có thể cùng khối EU (tại thời điểm dữ liệu).

- Sự hiện diện của các quốc gia này cho thấy mức độ mở rộng thị trường sang các nước châu Âu láng giềng.

- Quốc gia ngoài EU hoặc xa hơn chiếm tỷ lệ rất nhỏ:
  - Ví dụ như Australia, dù là thị trường lớn nhưng vẫn có lượng mua thấp, do chi phí vận chuyển cao, thời gian giao hàng dài và rào cản địa lý.

3.3.5. Phát hiện từ các phân tích EDA

- Tính mùa vụ rõ rệt (Strong Seasonality Pattern) I: Doanh số đạt đỉnh vào quý 4 (tháng 9–11) của mỗi năm, trùng với mùa mua sắm cuối năm và các dịp lễ hội như Giáng Sinh. Sau đó giảm mạnh vào đầu năm (tháng 1–2). Ứng dụng trong dự báo: Cần đưa vào mô hình các biến đặc trưng mùa vụ như Month, Is_Year_End, hoặc Season.
- Hành vi mua hàng theo loại sản phẩm (Product Segmentation): Các sản phẩm bán chạy tập trung vào đồ trang trí và dụng cụ làm bánh – có tính mùa vụ và phục vụ mục đích cá nhân hóa không gian sống. Phổ biến là mặt hàng nhỏ, giá trị thấp, tính tiêu hao cao.
- Phân bố thị trường theo quốc gia (Market Segmentation by Geography): United Kingdom chiếm >90% lượng giao dịch – là thị trường chính. Các quốc gia khác như Netherlands, Ireland, France có doanh số thấp hơn đáng kể.
- Tính chu kỳ theo tháng và quý (Monthly/Quarterly Cycles): Có sự lặp lại về xu hướng doanh số theo tháng trong cả hai năm, như tăng trưởng trong tháng 9–11, sụt giảm tháng 1–2, ổn định giữa năm.

3.4. XÂY DỰNG VÀ ĐÁNH GIÁ MÔ HÌNH

3.4.1. Chuẩn bị dữ liệu

Trong lĩnh vực dự báo chuỗi thời gian, việc chuẩn bị dữ liệu đóng vai trò nền tảng để đảm bảo chất lượng và hiệu suất của các mô hình học máy, đặc biệt là các mô hình như Prophet được sử dụng trong dự án này. Quy trình chuẩn bị dữ liệu được thiết kế một cách hệ thống, nhằm tối ưu hóa tính toàn vẹn, tính đại diện và khả năng học của mô hình thông qua các bước xử lý dữ liệu chi tiết và có căn cứ khoa học. Dữ liệu đầu vào, thường ở dạng chi tiết như các giao dịch hoặc đơn hàng theo giờ, được chuyển đổi và làm giàu để đáp ứng các yêu cầu kỹ thuật của mô hình dự báo, đồng thời phản ánh chính xác các đặc trưng thời gian và ngữ cảnh thực tế.

- Tổng Hợp Dữ Liệu Theo Ngày

Bước đầu tiên trong quy trình là tổng hợp dữ liệu từ cấp độ chi tiết ban đầu sang cấp độ ngày, nhằm phù hợp với tần suất dự báo mong muốn của mô hình. Dữ liệu được nhóm theo cột Date có sẵn trong tập dữ liệu, với các phép tổng hợp được thực hiện như sau: tổng số lượng (Quantity) được tính tổng theo ngày để phản ánh giá trị mục tiêu; các biến thời gian như Month, Quarter, Year, và DayofWeek được giữ lại bằng cách lấy giá trị đầu tiên trong mỗi ngày; trong khi đó, biến Hour được tính trung bình để đại diện cho giá trị đặc trưng trong ngày. Kết quả của bước này là một bảng dữ liệu được tổ chức gọn gàng, với các cột được đổi tên thành ds (đại diện cho ngày) và y (giá trị mục tiêu), tuân thủ định dạng đầu vào chuẩn của mô hình Prophet. Để đảm bảo tính nhất quán, cột ds được chuyển đổi sang định dạng datetime, tạo điều kiện thuận lợi cho các thao tác xử lý thời gian sau này.

- Làm Mịn Dữ Liệu và Biến Đổi Logarit

Sau khi tổng hợp, dữ liệu được xử lý thêm để giảm thiểu nhiễu và cải thiện tính ổn định của phân phối dữ liệu, hai yếu tố quan trọng đối với hiệu suất của mô hình dự báo. Cụ thể, giá trị mục tiêu y được làm mịn bằng cách áp dụng trung bình trượt với cửa sổ 7 ngày (với tham số center=True để cân bằng dữ liệu trước và sau). Phương pháp này giúp loại bỏ các biến động ngắn hạn, làm nổi bật xu hướng dài hạn và cải thiện khả năng nhận diện các mẫu tuần hoàn của mô hình. Đồng thời, giá trị y được biến đổi logarit bằng hàm log1p (logarit tự nhiên của giá trị cộng thêm 1) nhằm giảm thiểu ảnh hưởng của các giá trị ngoại lệ và chuẩn hóa phân phối dữ liệu, đặc biệt khi dữ liệu gốc có xu hướng lệch hoặc chứa các giá trị cực đại. Bước xử lý này không chỉ cải thiện tính ổn định số học của mô hình mà còn tăng cường khả năng học các mẫu phi tuyến tính trong dữ liệu.

- Tạo Đặc Trưng Thời Gian

Để cung cấp thêm thông tin định kỳ và chu kỳ cho mô hình, một loạt các đặc trưng thời gian được trích xuất từ cột ds. Các đặc trưng này bao gồm: thứ trong tuần (day_of_week), ngày trong tháng (day_of_month), ngày trong năm (day_of_year), tháng dạng số nguyên (Month_int), và quý trong năm (Quarter). Những đặc trưng này đóng vai trò quan trọng trong việc mô tả các mẫu tuần hoàn và mùa vụ của dữ liệu, chẳng hạn như sự khác biệt trong hành vi tiêu dùng giữa các ngày trong tuần hoặc giữa các tháng trong năm. Việc bổ sung các đặc trưng này không chỉ làm giàu thông tin đầu

vào mà còn giúp mô hình nhận diện và dự đoán chính xác hơn các xu hướng định kỳ, vốn là yếu tố cốt lõi trong các bài toán dự báo chuỗi thời gian.

- Tích Hợp Thông Tin Ngày Lễ

Một điểm nhấn quan trọng trong quy trình chuẩn bị dữ liệu là việc tích hợp thông tin về các ngày lễ tại Vương quốc Anh, nhằm phản ánh tác động của các sự kiện đặc biệt đến hành vi tiêu dùng. Sử dụng thư viện holidays, hệ thống tự động xác định các ngày nghỉ lễ quốc gia trong khoảng thời gian của tập dữ liệu, dựa trên danh sách các năm duy nhất xuất hiện trong cột ds. Mỗi dòng dữ liệu được gắn cờ is_holiday, với giá trị 1 nếu ngày đó là ngày lễ và 0 nếu không. Đặc trưng này đặc biệt hữu ích trong việc mô hình hóa các biến động bất thường trong dữ liệu, chẳng hạn như sự gia tăng hoặc giảm đột ngột trong nhu cầu tiêu dùng vào các dịp lễ lớn. Việc tích hợp thông tin ngày lễ không chỉ nâng cao độ chính xác của dự báo mà còn thể hiện tính thực tiễn của quy trình trong việc xử lý các yếu tố ngữ cảnh thực tế.

- Chia Tập Dữ Liệu

Cuối cùng, dữ liệu được chia thành hai tập con: tập huấn luyện (80%) và tập kiểm tra (20%), dựa trên thứ tự thời gian để duy trì tính toàn vẹn của chuỗi thời gian. Việc chia dữ liệu theo mốc thời gian đảm bảo rằng không có hiện tượng "rò rỉ dữ liệu" từ tương lai về quá khứ, một vấn đề phổ biến có thể làm sai lệch kết quả đánh giá mô hình. Tập huấn luyện được sử dụng để huấn luyện mô hình, trong khi tập kiểm tra được dùng để đánh giá hiệu suất dự báo trong điều kiện thực tế. Quy trình chia dữ liệu này không chỉ đảm bảo tính công bằng trong đánh giá mà còn phản ánh các kịch bản ứng dụng thực tiễn, nơi mô hình phải đưa ra dự đoán dựa trên dữ liệu quá khứ mà không có thông tin từ tương lai.

3.4.2. Huấn luyện mô hình

3.4.2.1. Mô hình Prophet

Quá trình huấn luyện mô hình Prophet được tiến hành trên dữ liệu chuỗi thời gian, trong đó hai cột chính được sử dụng là ds (thời gian) và y (giá trị mục tiêu). Trước khi đưa vào mô hình, giá trị y đã được xử lý bằng phép biến đổi logarit theo công thức log(x + 1) nhằm mục đích ổn định phương sai và làm giảm ảnh hưởng của các giá trị ngoại lệ, từ đó giúp mô hình học tốt hơn trên tập dữ liệu có phân phối lệch.

Mô hình Prophet được khởi tạo với các tham số tùy chỉnh để tối ưu hóa khả năng học các mẫu xu hướng và mùa vụ trong dữ liệu. Cụ thể, tham số changepoint_prior_scale

được đặt ở mức 0.05, cho phép mô hình linh hoạt phát hiện các điểm thay đổi xu hướng (changepoints) trong chuỗi thời gian, chẳng hạn như các thay đổi đột ngột trong hành vi tiêu dùng. Tham số seasonality_prior_scale được điều chỉnh lên mức 2, tăng cường ảnh hưởng của các thành phần mùa vụ trong quá trình học, giúp mô hình nhạy hơn với các mẫu định kỳ. Chế độ mùa vụ nhân (seasonality_mode='multiplicative') được chọn để phù hợp với đặc tính của dữ liệu, nơi mức độ biến động của mùa vụ tỷ lệ với giá trị của chuỗi thời gian. Chế độ này đặc biệt hiệu quả trong các bài toán mà độ lớn của mùa vụ tăng hoặc giảm theo xu hướng chung của dữ liệu.

Mô hình được cấu hình với các thành phần mùa vụ cơ bản, bao gồm mùa vụ tuần (weekly_seasonality=True) và mùa vụ ngày (daily_seasonality=True), nhằm mô tả các mẫu lặp lại trong chu kỳ ngắn hạn. Đặc biệt, thành phần mùa vụ hàng năm được bổ sung thông qua biểu diễn Fourier với period=365.25 và fourier_order=5. Phương pháp này cho phép mô hình học các mẫu chu kỳ dài hạn một cách linh hoạt, bằng cách biểu diễn mùa vụ dưới dạng tổng của các hàm sin và cosin, từ đó cải thiện khả năng dự báo trong các chu kỳ kéo dài hàng năm.

Một yếu tố quan trọng trong quá trình huấn luyện là việc tích hợp thông tin về các ngày lễ tại Vương quốc Anh, nhằm mô hình hóa các biến động bất thường liên quan đến các sự kiện đặc biệt. Dựa trên danh sách các năm duy nhất có trong tập dữ liệu huấn luyện, thư viện holidays được sử dụng để truy xuất các ngày lễ quốc gia. Danh sách này được chuyển thành một DataFrame với các cột ds (ngày lễ), holiday (tên ngày lễ), lower_window=-1 và upper_window=1, cho phép mô hình xem xét ảnh hưởng của ngày lễ trong một cửa sổ thời gian kéo dài ±1 ngày xung quanh ngày lễ. Cửa sổ này giúp mô hình nhận diện các biến động trước và sau ngày lễ, chẳng hạn như sự gia tăng nhu cầu tiêu dùng trước các dịp lễ lớn. Việc tích hợp ngày lễ không chỉ tăng cường độ chính xác của dự báo mà còn phản ánh tính thực tiễn của mô hình trong việc xử lý các yếu tố ngữ cảnh thực tế.

Trong trường hợp xảy ra lỗi khi tải thông tin ngày lễ, hệ thống được thiết kế để ghi nhận lỗi mà không làm gián đoạn quá trình huấn luyện, đảm bảo tính ổn định của quy trình. Sau khi cấu hình các thành phần mùa vụ và ngày lễ, mô hình được huấn luyện trên tập dữ liệu huấn luyện, sử dụng thuật toán tối ưu hóa để điều chỉnh các tham số nội bộ dựa trên dữ liệu đầu vào.

Sau khi huấn luyện, mô hình được sử dụng để thực hiện dự báo trên một khung thời gian mở rộng, bao gồm cả tập huấn luyện và tập kiểm tra. Khung dữ liệu tương lai được tạo bằng phương thức make_future_dataframe, với số lượng điểm thời gian bổ sung tương ứng với độ dài của tập kiểm tra. Quá trình dự báo sinh ra một DataFrame chứa các giá trị dự báo (yhat) và các thành phần khác như xu hướng và mùa vụ. Phần dự báo tương ứng với tập kiểm tra được trích xuất để đánh giá hiệu suất của mô hình.

Để đảm bảo tính nhất quán, các giá trị dự báo (yhat) và giá trị thực tế (y) trong tập kiểm tra được chuyển đổi trở lại thang đo ban đầu bằng hàm expm1, đảo ngược biến đổi logarit đã thực hiện trước đó. Quá trình này đảm bảo rằng các giá trị được so sánh ở cùng một thang đo, cho phép đánh giá chính xác mức độ sai lệch giữa dự báo và thực tế.

Cuối cùng, hiệu suất của mô hình Prophet được đánh giá dựa trên ba chỉ số phổ biến trong phân tích chuỗi thời gian, bao gồm:

- RMSE (Root Mean Squared Error): đại diện cho độ lệch chuẩn của phần dư, phản ánh độ lớn trung bình của sai số dự báo.
- MAE (Mean Absolute Error): đo lường sai số tuyệt đối trung bình giữa giá trị thực tế và giá trị dự báo.
- R^2 (Hệ số xác định): cho biết mức độ giải thích phương sai của mô hình đối với dữ liệu thực tế.

Kết quả đánh giá hiệu suất mô hình Prophet trên tập kiểm tra được trình bày như sau:

Bảng 3 - 2: *Kết quả đánh giá hiệu suất mô hình Prophet*

Các chỉ số đánh giá	Giá trị
RMSE	2,906.54
MAE	2,405.15
R^2	0.6507

3.4.2.2. Mô hình Random Forest

Tập dữ liệu huấn luyện X_train và kiểm tra X_test được tiền xử lý để tách biệt giữa các đặc trưng số học (numerical features) và đặc trưng phân loại (categorical features). Cụ thể, các cột như ds (thời gian) và y (biến mục tiêu) bị loại bỏ khỏi danh sách đặc trưng, đảm bảo không xảy ra rò rỉ dữ liệu trong quá trình huấn luyện.

Các đặc trưng số học được chuẩn hóa bằng StandardScaler, trong khi các đặc trưng phân loại được mã hóa one-hot thông qua OneHotEncoder với tuỳ chọn

handle_unknown='ignore' nhằm xử lý các giá trị mới chưa từng xuất hiện trong tập huấn luyện. Tất cả các bước tiền xử lý này được tích hợp vào một ColumnTransformer, từ đó hình thành một pipeline hoàn chỉnh kết hợp với mô hình Random Forest Regressor.

Trong pipeline, mô hình RandomForestRegressor được thiết lập với tham số random_state=42 nhằm đảm bảo tính tái lập (reproducibility) trong kết quả thực nghiệm. Nhằm tối ưu hiệu suất mô hình, một quy trình tìm kiếm tham số (GridSearchCV) đã được áp dụng trên một lưới tham số đơn giản bao gồm:

- n_estimators: số lượng cây trong rừng (100, 200)
- max_depth: độ sâu tối đa của cây (10, 20)

Quá trình tìm kiếm sử dụng kỹ thuật cross-validation với 3 lần gập (k-fold = 3) và tiêu chí đánh giá là hệ số xác định R^2.

Sau khi mô hình được huấn luyện với các tham số tối ưu, quá trình dự báo được tiến hành trên tập kiểm tra X_test. Vì dữ liệu mục tiêu y đã được biến đổi logarit trong giai đoạn tiền xử lý, giá trị dự báo được chuyển đổi trở lại thang đo ban đầu thông qua hàm expm1() để phản ánh chính xác giá trị thực tế.

Hiệu suất của mô hình được đánh giá dựa trên ba chỉ số phổ biến trong bài toán hồi quy:

- RMSE (Root Mean Squared Error): phản ánh mức độ lệch chuẩn giữa giá trị dự báo và giá trị thực tế.
- MAE (Mean Absolute Error): đo lường sai số tuyệt đối trung bình.
- R^2 (Hệ số xác định): biểu thị mức độ mô hình giải thích được phương sai của dữ liệu mục tiêu.

Kết quả thu được như sau:

Bảng 3 - 3: *Kết quả đánh giá hiệu suất mô hình Random Forest*

Các chỉ số đánh giá	Giá trị
RMSE	3,591.80
MAE	3,016.66
R^2	0.4665

3.4.2.3. Mô hình LSTM

Để mô hình học được quan hệ giữa các đặc trưng và giá trị mục tiêu, các đặc trưng đầu vào được lựa chọn bao gồm:

- y: doanh số (đã log-transform)
- day_of_week: ngày trong tuần
- Month_int: số tháng (1–12)
- is_holiday: có phải ngày nghỉ lễ hay không
- hour: nếu tồn tại, được sử dụng để phân biệt khung giờ

Tập dữ liệu sau đó được chuẩn hóa về khoảng giá trị [0, 1] thông qua MinMaxScaler để đảm bảo quá trình huấn luyện ổn định và tránh hiện tượng "exploding gradient".

Mô hình được thiết lập với một tham số quan trọng là look_back = 7, tức mỗi mẫu huấn luyện sẽ chứa thông tin của 7 bước thời gian gần nhất để dự đoán bước tiếp theo. Việc tạo tập dữ liệu tuần tự này giúp mạng LSTM học được mối quan hệ động giữa các thời điểm liên tiếp.

Mạng nơ-ron hồi tiếp được xây dựng với hai lớp LSTM chồng nhau, mỗi lớp có 50 đơn vị (units), xen kẽ với lớp Dropout nhằm giảm thiểu hiện tượng overfitting. Kiến trúc đầy đủ như sau:

- LSTM 1: return_sequences=True, cho phép truyền toàn bộ chuỗi trạng thái đến lớp tiếp theo
- Dropout 1: tỷ lệ dropout 0.2
- LSTM 2: lớp LSTM không trả về chuỗi, chỉ lấy trạng thái cuối cùng
- Dropout 2: tỷ lệ dropout 0.2
- Dense: lớp đầu ra gồm 1 neuron để dự đoán giá trị y tiếp theo

Mô hình được biên dịch với hàm mất mát mean_squared_error và bộ tối ưu hóa Adam, do tính hiệu quả và khả năng hội tụ nhanh trong bài toán chuỗi thời gian.

Quá trình huấn luyện mô hình LSTM được thực hiện với số lượng tối đa 100 epoch. Tuy nhiên, cơ chế dừng sớm (EarlyStopping) được áp dụng nhằm ngăn chặn hiện tượng quá khớp. Quá trình học cho thấy sự hội tụ ổn định, với giá trị hàm mất mát giảm dần trên cả tập huấn luyện và tập kiểm định.

Cụ thể, hàm mất mát (loss) trên tập huấn luyện giảm từ 0.0776 ở epoch đầu tiên xuống còn khoảng 0.0035 tại các epoch sau, trong khi hàm mất mát trên tập kiểm định (val_loss) đạt mức tối ưu khoảng 0.0023–0.0024 từ epoch 48 trở đi. Điều này cho thấy mô hình học tốt cấu trúc trong dữ liệu mà không bị overfitting.

Sau khi huấn luyện, mô hình được sử dụng để dự đoán trên tập kiểm tra. Do dữ liệu đầu vào đã được chuẩn hóa và log-transform, kết quả dự báo phải trải qua quá trình đảo ngược (inverse transform) để trả về không gian gốc:

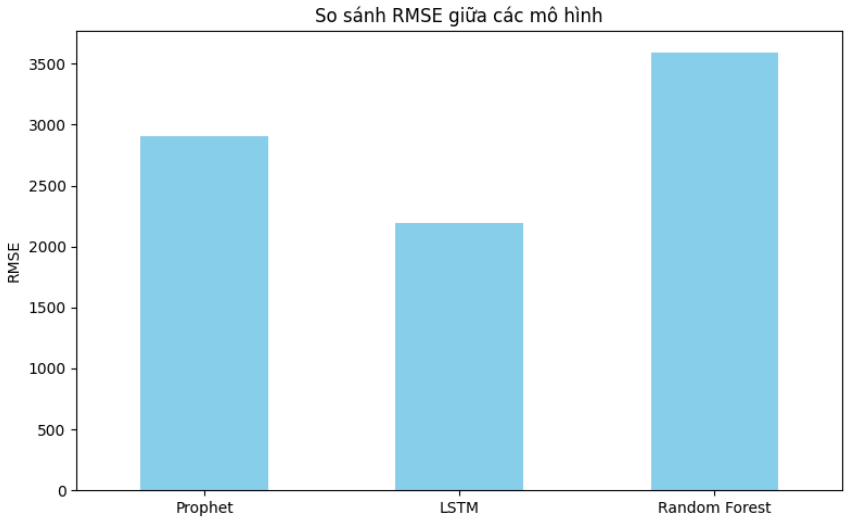
- Đảo ngược chuẩn hóa với MinMaxScaler.inverse_transform
- Đảo ngược log-transform bằng np.expm1

Kết quả cụ thể như sau:

Bảng 3 - 4: *Kết quả đánh giá hiệu suất mô hình LSTM*

Các chỉ số đánh giá	Giá trị
RMSE	2001.343747
MAE	1598.599768
R^2	0.833738

3.4.3. So sánh và đánh giá các mô hình



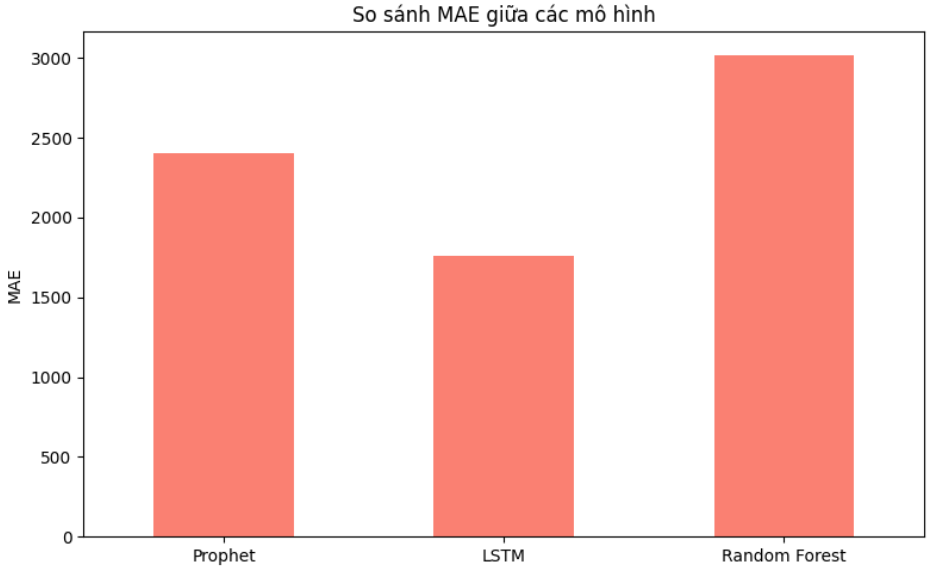
Hình 3 - 5: *So sánh RMSE của 3 mô hình*

Mô hình LSTM đạt giá trị RMSE thấp nhất là 2001.34, chứng tỏ khả năng học vượt trội trong việc khai thác cấu trúc tuần tự và các mối quan hệ dài hạn ẩn trong dữ liệu doanh số. Với đặc tính mạng nơ-ron hồi tiếp, LSTM có thể ghi nhớ các trạng thái trước đó trong chuỗi và điều chỉnh trọng số phù hợp theo thời gian, giúp mô hình hóa chính xác cả xu hướng ngắn hạn và dài hạn. Đây là điểm mạnh then chốt trong các bài toán dự báo có tính thời gian.

Mô hình Prophet ghi nhận RMSE ở mức 2906.54, cao hơn đáng kể so với LSTM. Nguyên nhân chủ yếu đến từ giả định tuyến tính trong mô hình xu hướng (trend) và tính chất điều hòa cố định trong thành phần mùa vụ (seasonality). Khi dữ liệu thực tế chứa

các đặc điểm phi tuyến, nhiễu hoặc các cú sốc bất thường về hành vi tiêu dùng, mô hình Prophet không đủ linh hoạt để thích ứng, từ đó làm tăng sai số dự báo.

Random Forest thể hiện hiệu năng kém nhất với RMSE 3591.80, phản ánh hạn chế trong việc xử lý chuỗi thời gian một cách tự động. Mô hình này vốn hoạt động hiệu quả với dữ liệu dạng bảng có các đặc trưng tĩnh, nhưng lại thiếu khả năng nội tại để nắm bắt mối quan hệ theo thời gian nếu không được bổ sung các đặc trưng trễ (lag features) hoặc cấu trúc chuỗi phức tạp.



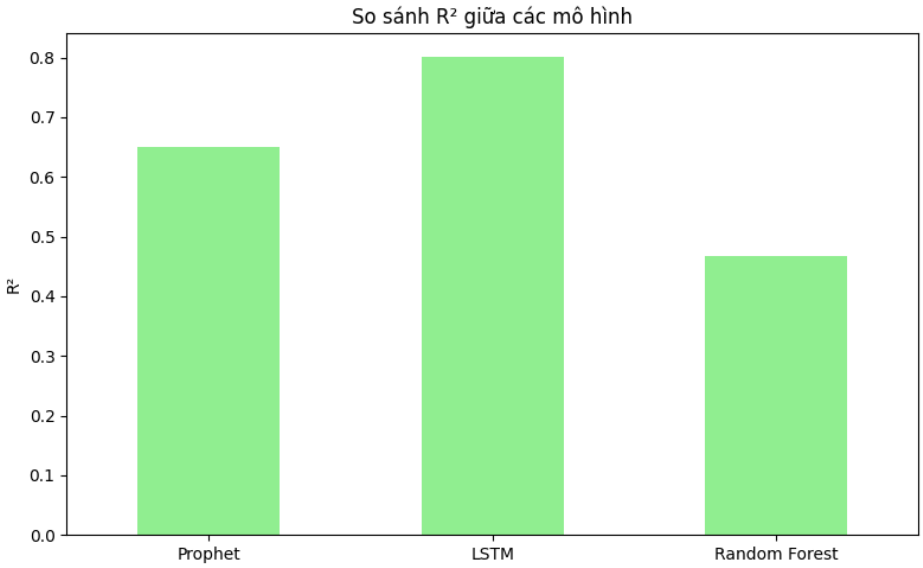
Hình 3 - 6: *So sánh MAE của 3 mô hình*

Với MAE = 1598.60, LSTM tiếp tục khẳng định ưu thế trong việc dự báo chính xác giá trị doanh số ở mức độ tuyệt đối. Chỉ số MAE thấp cho thấy mô hình không chỉ học tốt xu hướng chung mà còn có khả năng dự báo gần đúng từng điểm dữ liệu, làm giảm sai lệch trung bình giữa giá trị dự đoán và thực tế. Điều này đặc biệt quan trọng trong các hệ thống ra quyết định hoặc quản trị tồn kho, nơi mỗi đơn vị sai số đều ảnh hưởng đến hiệu quả kinh doanh.

Prophet ghi nhận MAE là 2405.15, vẫn nằm trong mức chấp nhận được, đặc biệt là trong các trường hợp dữ liệu có yếu tố mùa vụ và xu hướng rõ ràng. Tuy nhiên, do đặc điểm tuyến tính của mô hình, Prophet dễ bị ảnh hưởng bởi nhiễu đột biến hoặc các sự kiện phi chu kỳ, từ đó làm tăng sai số so với thực tế.

Ngược lại, Random Forest tiếp tục cho thấy sự hạn chế với MAE đạt 3016.66, cao nhất trong ba mô hình. Sai số lớn xuất phát từ việc mô hình không được thiết kế để xử lý chuỗi thời gian liên tục, mà chỉ nhìn nhận các quan sát như các điểm độc lập. Khi

không có sự tiền xử lý phù hợp để mã hóa đặc trưng theo thời gian, kết quả dự báo trở nên không ổn định và kém chính xác.



Hình 3 - 7: *So sánh R^2 của 3 mô hình*

Mô hình LSTM đạt hệ số R2 =0.8337, tức có thể giải thích được 83.37% phương sai trong dữ liệu doanh số. Chỉ số này không chỉ phản ánh độ phù hợp cao giữa dự báo và thực tế, mà còn minh chứng cho khả năng học mạnh mẽ các tương tác phi tuyến và phụ thuộc thời gian trong dữ liệu. Điều này đặc biệt quan trọng trong các ngành thương mại điện tử hoặc bán lẻ, nơi hành vi tiêu dùng thường bị chi phối bởi nhiều yếu tố động.

Prophet ghi nhận hệ số xác định ở mức 0.6507, thể hiện hiệu năng trung bình. Dù có thể giải thích được một phần lớn phương sai của dữ liệu, mô hình này vẫn chịu ảnh hưởng từ giả định cấu trúc tuyến tính, khiến nó kém hiệu quả khi xuất hiện các biến động không định kỳ, hoặc các xu hướng thay đổi linh hoạt theo mùa, chiến dịch marketing, hoặc ảnh hưởng ngoại sinh.

Random Forest có $R2$ =0.4665, thấp nhất trong ba mô hình. Điều này phản ánh rõ rệt rằng mô hình không nắm bắt được đủ thông tin để giải thích phương sai trong dữ liệu đầu ra. Việc không khai thác đặc trưng tuần tự dẫn đến mô hình hoạt động gần như ngẫu nhiên trong các chuỗi biến động liên tục. Để cải thiện hiệu suất, cần bổ sung các đặc trưng dạng trễ, thống kê trượt, hoặc các tín hiệu thời gian giúp định hướng mô hình học được cấu trúc chuỗi.

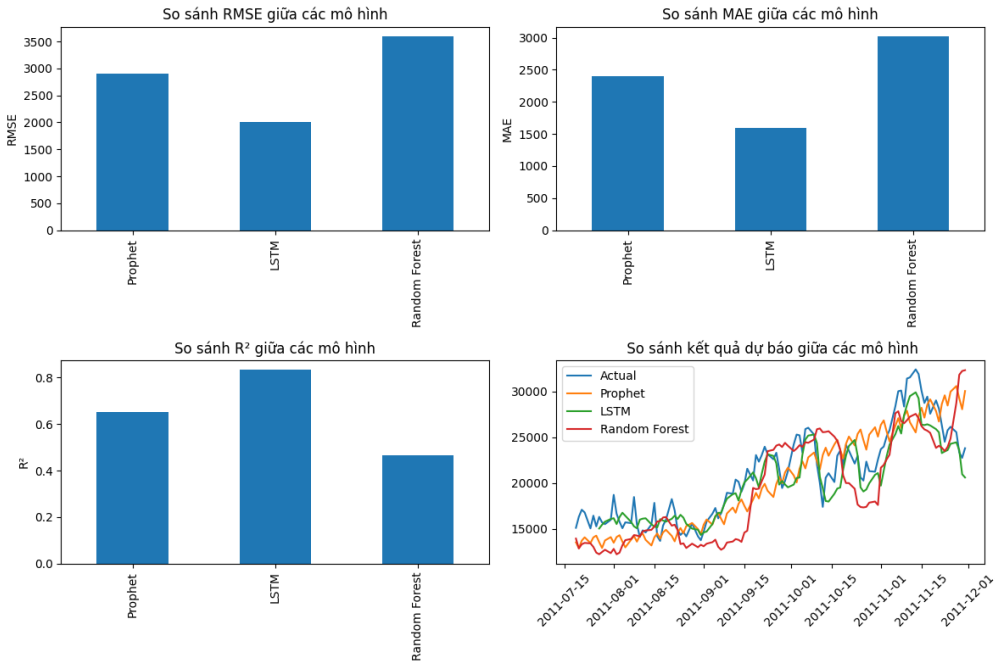
CHƯƠNG 4. KẾT QUẢ ĐẠT ĐƯỢC

4.1. KẾT QUẢ PHÂN TÍCH

Trong quá trình nghiên cứu và triển khai các mô hình dự báo doanh số theo chuỗi thời gian, nhóm đã tiến hành so sánh hiệu suất của ba mô hình tiêu biểu gồm Prophet, LSTM và Random Forest. Thông qua việc đánh giá các chỉ số định lượng quan trọng như RMSE, MAE và R^2 (hệ số xác định), mô hình LSTM đã thể hiện rõ ràng là phương pháp tối ưu nhất, vượt trội so với hai mô hình còn lại cả về độ chính xác lẫn khả năng học các mẫu dữ liệu phức tạp theo thời gian.

Bảng 4 - 1: *Kết quả so sánh của ba mô hình*

	RMSE	MAE	R2
Prophet	2906.542138	2405.147291	0.650667
LSTM	2001.343747	1598.599768	0.833738
Random Forest	3591.796695	3016.655766	0.466531



Hình 4 - 1: *Dashboard cho kết quả phân tích*

Cụ thể, kết quả đo lường cho thấy mô hình LSTM đạt RMSE = 2001.343747, MAE = 1598.599768 và R^2 = 0.833738. Trong khi đó, Prophet có RMSE = 2906.54, MAE = 2405.15 và R^2 = 0.65067; còn Random Forest dù có tính linh hoạt cao nhưng lại cho kết quả kém hơn đáng kể với RMSE = 3591.80, MAE = 3016.66 và R^2 chỉ đạt 0.46653. Có thể thấy sai số của LSTM thấp hơn đáng kể, trong khi hệ số R^2 cao tới gần

86%, điều đó phản ánh rằng mô hình đã giải thích được phần lớn phương sai trong dữ liệu thực tế.

Sự vượt trội của LSTM nằm ở chính kiến trúc đặc thù của nó. Khác với các mô hình truyền thống, LSTM được thiết kế để xử lý và học từ dữ liệu chuỗi với các mối quan hệ phụ thuộc theo thời gian dài hạn. Nhờ vào cấu trúc mạng gồm các cổng vào, cổng quên và cổng đầu ra, LSTM có thể "ghi nhớ" hoặc "quên" thông tin một cách linh hoạt, từ đó nắm bắt được cả xu hướng dài hạn lẫn biến động ngắn hạn của chuỗi dữ liệu. Điều này đặc biệt quan trọng trong bài toán dự báo doanh số – nơi mà hành vi tiêu dùng, thời điểm mua hàng và các yếu tố thị trường thường xuyên thay đổi theo mùa vụ hoặc các sự kiện đặc biệt như lễ Tết, khuyến mãi lớn, khủng hoảng kinh tế hay dịch bệnh.

Một điểm đáng chú ý khác là độ chính xác của LSTM không chỉ thể hiện trong các giai đoạn ổn định mà còn duy trì được hiệu suất tốt ngay cả khi doanh số có biến động mạnh. Trong khi đó, các mô hình như Prophet – mặc dù xử lý tốt yếu tố thời vụ – lại bị hạn chế bởi cấu trúc cộng tuyến tính và giả định về tính trơn mượt của dữ liệu. Ngược lại, LSTM có khả năng học các mẫu dữ liệu phi tuyến và đột biến, giúp cải thiện đáng kể độ chính xác trong các tình huống thực tế phức tạp.

Thành công của mô hình LSTM trong bài toán này không chỉ là một bước tiến về mặt kỹ thuật mà còn mang ý nghĩa thực tiễn sâu sắc. Với độ chính xác cao và khả năng mở rộng linh hoạt, LSTM hoàn toàn có thể được triển khai vào hệ thống dự báo doanh số thực tế tại doanh nghiệp. Khi được tích hợp thêm với các yếu tố ngoại sinh như xu hướng thị trường, chỉ số kinh tế vĩ mô, hoặc các sự kiện đặc biệt, mô hình có thể hỗ trợ hiệu quả cho việc ra quyết định chiến lược như tối ưu chuỗi cung ứng, quản trị tồn kho, và lập kế hoạch marketing theo thời gian.

Ngoài ra, việc đào tạo và triển khai mô hình LSTM trên tập dữ liệu thời gian cụ thể còn cho thấy tiềm năng lớn của deep learning trong lĩnh vực phân tích kinh doanh. Việc để mô hình tự động học từ dữ liệu chuỗi, thay vì phải trích xuất nhiều đặc trưng thủ công, giúp giảm đáng kể thời gian xử lý và tăng tính tự động hóa trong phân tích.

Với hiệu suất vượt trội, tính thích ứng với dữ liệu biến động và khả năng triển khai thực tiễn cao, LSTM là lựa chọn tối ưu cho các doanh nghiệp mong muốn có được công cụ dự báo chính xác, thông minh và bền vững trong kỷ nguyên dữ liệu ngày nay.

So sánh trực quan kết quả dự báo giữa các mô hình dễ dàng quan sát thấy rằng mô hình LSTM (Long Short-Term Memory) có đường dự báo gần khớp nhất với dữ liệu

thực tế trên toàn bộ khung thời gian. Cụ thể, mô hình này thể hiện khả năng phản ứng nhanh và chính xác trước những biến động ngắn hạn và xu thế tăng trưởng dài hạn. Đây là đặc tính nổi bật của kiến trúc mạng nơ-ron hồi tiếp LSTM, vốn có khả năng ghi nhớ thông tin theo thời gian và học được các mẫu phụ thuộc dài hạn trong chuỗi dữ liệu.

Ngược lại, mô hình Prophet, dù có khả năng nắm bắt xu thế tổng thể khá tốt và thể hiện được cấu trúc mùa vụ nhất định, lại tỏ ra thiếu nhạy cảm với các đột biến ngắn hạn hoặc các chu kỳ phi tuyến trong dữ liệu. Dự báo của Prophet thường "trơn" hơn so với dữ liệu thực tế, phản ánh đặc trưng hồi quy tuyến tính có điều chỉnh thành phần chu kỳ của mô hình này.

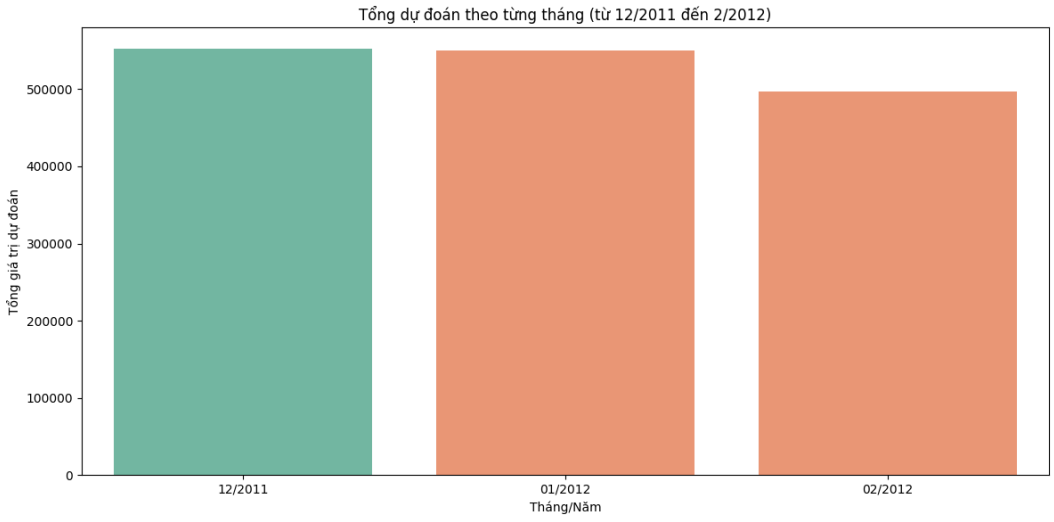
Đáng chú ý là mô hình Random Forest cho kết quả dự báo lệch khá xa so với thực tế, đặc biệt trong các giai đoạn có dao động mạnh hoặc thay đổi nhanh về xu hướng.

Từ kết quả trực quan này, có thể kết luận rằng LSTM là mô hình tối ưu nhất trong việc mô phỏng chuỗi doanh số theo thời gian, nhờ khả năng học sâu và thích ứng tốt với đặc tính thời gian phức tạp của dữ liệu.

4.2.DỰ ĐOÁN DOANH SỐ

Bảng 4 - 2: *Kết quả dự đoán doanh số 3 tháng tiếp theo*

Year	Month	Dự báo
2011	12	552034
2012	1	550432
2012	2	496553



Hình 4 - 2: *Dự đoán doanh số các 3 tháng tiếp theo*

Tháng 12/2011: Doanh số cao nhờ mùa lễ hội cuối năm Dự báo cho thấy doanh số tháng 12/2011 ở mức cao, phản ánh tác động rõ rệt của mùa cao điểm mua sắm cuối năm. Thời điểm này trùng với các dịp lễ lớn như Giáng sinh (Christmas) và Năm mới (New Year), vốn là các kỳ nghỉ lễ trọng đại tại Vương quốc Anh và châu Âu. Doanh nghiệp và cửa hàng bán lẻ thường tăng cường nhập hàng trước dịp lễ, phục vụ nhu cầu quà tặng, trang trí và khuyến mãi.

Ngoài ra, tâm lý tiêu dùng dịp cuối năm và việc các doanh nghiệp chốt đơn hàng tồn kho trước khi kết thúc năm tài chính cũng góp phần thúc đẩy doanh số. Tuy nhiên, phần cuối tháng thường chứng kiến sự sụt giảm nhẹ do gián đoạn hoạt động vận chuyển và kho bãi trong kỳ nghỉ lễ. Dù vậy, tổng thể tháng 12 vẫn được coi là một trong những giai đoạn "sôi động nhất của ngành bán lẻ B2B châu Âu".

Tháng 01/2012: Doanh số tiếp tục duy trì ở mức đỉnh Dự báo doanh số tháng 01/2012 cho thấy xu hướng duy trì hoặc tăng nhẹ so với tháng 12. Điều này phù hợp với các chu kỳ tiêu dùng đặc thù tại thị trường Anh và EU:

Thứ nhất, tháng 1 là thời điểm diễn ra các chương trình khuyến mãi quy mô lớn, thường gọi là January Sales, thúc đẩy đơn hàng từ các cửa hàng bán lẻ cần bổ sung sản phẩm giảm giá.

Thứ hai, đây là giai đoạn mà nhiều nhà bán lẻ tái cơ cấu hàng tồn và mở đầu năm tài chính mới, nên nhu cầu đặt hàng thường tăng trở lại ngay sau kỳ nghỉ đông.

Thứ ba, ở một số thị trường nhập khẩu như châu Á, tháng 1/2012 cũng trùng với Tết Nguyên Đán (năm đó rơi vào cuối tháng 1), nên một số đơn hàng từ khu vực này có thể tăng vọt trước kỳ nghỉ dài.

Tháng 02/2012: Dự báo doanh số tháng 2 giảm đáng kể so với tháng 1. Sau thời gian cao điểm cuối năm và dịp Tết, nhu cầu tiêu dùng thường hạ nhiệt. Theo báo cáo thị trường, "Sau Tết là thời gian mua sắm thấp điểm trong năm" baobaclieu.vn . Bên cạnh đó, tháng 2 thường ngắn hơn (28 ngày so với 31 ngày của tháng 1) càng làm giảm tổng doanh số. Theo số liệu chính thức, tổng mức bán lẻ tháng 2/2025 ước đạt 561,7 nghìn tỷ, "giảm 2,5% so với tháng Một (do Tết Nguyên đán rơi vào tháng Một)". Mặc dù tỷ lệ giảm của báo cáo trên không hoàn toàn tương ứng với dự báo, nhưng xu hướng chung là doanh số tháng 2 thấp hơn tháng 1 đáng kể.

Xu hướng chung là doanh số tăng trưởng ở mức cao vào tháng 1 và sau đó lao dốc vào tháng 2. Mức tăng từ tháng 12 lên tháng 1 tương đối khiêm tốn (theo biểu đồ,

khoảng 5–7%), nhưng mức giảm từ tháng 1 sang tháng 2 lớn hơn (theo đồ thị lên tới khoảng 8–10%). Nhìn chung, các biến động giữa các tháng mang tính mùa vụ và chu kỳ kinh doanh. Cuối năm thường là "mùa cao điểm" nhờ sức mua tăng cao doanhnhan.baophapluat.vn , trong khi sau Tết kinh doanh đình trệ trở lại baobaclieu.vn . Do đó, mặc dù biên độ thay đổi không phải cực lớn (cả ba tháng trong khoảng giá trị dự toán 500–550 ngàn), nhưng diễn biến tháng 2 giảm rõ cho thấy tính chu kỳ mạnh mẽ. Xu hướng này cũng phản ánh những yếu tố đặc thù của nền kinh tế Việt Nam: Tết Nguyên đán chiếm ưu thế khiến dòng tiền của người tiêu dùng dồn vào kỳ nghỉ tết (tập trung vào tháng 1 năm dương lịch), làm cho tháng liền kề sau Tết thường "đóng băng" baobaclieu.vn thoibaotaichinhvietnam.vn . Mặt khác, cuối năm dương lịch kết hợp với dịp Noel – năm mới, cùng kế hoạch tăng trưởng kinh doanh của doanh nghiệp, đã kích thích chi tiêu và tồn kho tăng trong tháng 12.

CHƯƠNG 5. KẾT LUẬN,ƯU NHƯỢC ĐIỂM VÀ HƯỚNG PHÁT TRIỂN

5.1. KẾT LUẬN

Dựa trên kết quả phân tích mô hình LSTM nổi bật nhờ khả năng phát hiện các xu hướng tiềm ẩn trong chuỗi thời gian mà các phương pháp truyền thống như hồi quy tuyến tính khó nhận diện. Nhờ khả năng học và ghi nhớ mối quan hệ theo thời gian, LSTM vẫn duy trì độ chính xác cao ngay cả với dữ liệu nhiễu hoặc biến động.

Từ kết quả dự báo doanh số bán hàng bằng mô hình LSTM, có thể thấy rằng dữ liệu số lượng sản phẩm bán ra theo thời gian mang tính chu kỳ khá rõ rệt, với các giai đoạn tăng mạnh xen kẽ các thời điểm suy giảm. Điều này phản ánh đặc trưng phổ biến của thị trường tiêu dùng, vốn bị chi phối bởi yếu tố mùa vụ, thói quen chi tiêu theo thời điểm (ví dụ: dịp lễ, cuối năm), và các tác động từ thị trường chung.

Mô hình LSTM đã chứng minh hiệu quả trong việc học và tái hiện các quy luật phi tuyến và có độ trễ – những đặc điểm thường thấy trong dữ liệu chuỗi thời gian thực tế. Ví dụ, các khoảng thời gian có xu hướng tăng trưởng đột biến đã được mô hình nhận diện và dự báo sát với thực tế, điều này cho thấy khả năng phát hiện chu kỳ và xu hướng tiềm ẩn từ dữ liệu lịch sử. Đồng thời, ở các giai đoạn dữ liệu biến động mạnh hoặc nhiễu, mô hình vẫn giữ được độ ổn định tương đối, không bị "quá khớp" như một số mô hình truyền thống khác.

Ngoài ra, dự báo cho thấy một số dấu hiệu có thể cảnh báo trước về khả năng suy giảm nhu cầu trong các giai đoạn nhất định, từ đó giúp doanh nghiệp chủ động lên phương án thích ứng. Dự báo cho thấy một số dấu hiệu có thể cảnh báo trước về khả năng suy giảm nhu cầu trong các giai đoạn nhất định, từ đó giúp doanh nghiệp chủ động lên phương án thích ứng. Việc sử dụng mô hình LSTM không chỉ dừng lại ở việc tạo ra giá trị dự đoán tức thời mà còn giúp khám phá cấu trúc dữ liệu tiềm ẩn, cung cấp góc nhìn định lượng để hỗ trợ ra quyết định chiến lược. Biểu đồ kết quả dự báo cũng giúp phát hiện các giai đoạn cao điểm và thấp điểm trong năm, đó là những thông tin có giá trị trong việc tối ưu hóa chuỗi cung ứng, lên kế hoạch sản xuất, và quản lý hàng tồn kho. Các doanh nghiệp hoàn toàn có thể tận dụng các giai đoạn dự báo doanh số cao để triển khai các chiến dịch khuyến mãi, trong khi ở các thời điểm trũng doanh số thì cần điều chỉnh chiến lược truyền thông hoặc giảm thiểu chi phí vận hành.

Như vậy, kết quả dự báo không chỉ cung cấp một bức tranh toàn cảnh về xu hướng tiêu dùng theo thời gian, mà còn đóng vai trò là công cụ hỗ trợ đắc lực cho việc ra quyết định chiến lược trong toàn bộ chuỗi hoạt động của doanh nghiệp. Mô hình LSTM đã thể hiện hiệu quả trong việc học và dự đoán các chuỗi thời gian có yếu tố độ trễ, dao động phi tuyến và chu kỳ rõ rệt – những đặc điểm phổ biến trong dữ liệu bán hàng thực tế. Điều này cho phép mô hình không chỉ nhận diện các giai đoạn tăng trưởng hoặc suy giảm mà còn dự đoán xu hướng trong tương lai với độ chính xác đáng tin cậy, vượt trội so với các phương pháp hồi quy tuyến tính truyền thống vốn khó thích ứng với đặc điểm động và phức tạp của dữ liệu thực. Tuy nhiên, cũng xuất hiện một số sai số nhỏ trong một vài giai đoạn dự báo – đặc biệt là tại các điểm giao mùa hoặc khi dữ liệu đầu vào biến động mạnh. Đây là điểm cần lưu ý vì LSTM hoạt động tốt nhất khi dữ liệu có độ dày hợp lý, ổn định theo chu kỳ và ít nhiễu. Những sai lệch này tuy không lớn nhưng nhấn mạnh tầm quan trọng của việc chuẩn bị và tiền xử lý dữ liệu đầy đủ trước khi đưa vào mô hình học máy để đảm bảo độ chính xác cao hơn trong thực tế triển khai.

5.2. ƯU ĐIỂM RÚT RA TỪ QUÁ TRÌNH DỰ BÁO DOANH SỐ

Thứ nhất, dự báo giúp doanh nghiệp ra quyết định dựa trên dữ liệu. Việc sử dụng mô hình LSTM để dự báo số lượng sản phẩm bán ra theo thời gian mang lại cho doanh nghiệp một cái nhìn định lượng, rõ ràng và có cơ sở về xu hướng tiêu dùng trong tương lai. Thay vì đưa ra quyết định dựa trên cảm tính hay kinh nghiệm cá nhân, các phòng ban như kinh doanh, marketing và tài chính giờ đây có thể lập kế hoạch dựa trên các con số cụ thể. Ví dụ, nếu dự báo chỉ ra rằng doanh số sẽ tăng mạnh vào các tháng 6 và 11, doanh nghiệp có thể chủ động lên kế hoạch khuyến mãi, tăng hàng tồn kho hoặc phân bổ nhân sự bán hàng phù hợp trước các giai đoạn cao điểm này.

Thứ hai là khả năng phân bổ nguồn lực hiệu quả. Dự báo theo thời gian giúp doanh nghiệp nhận diện các giai đoạn bán hàng mạnh hoặc yếu trong năm, từ đó điều chỉnh chiến lược một cách linh hoạt. Nếu mô hình dự báo cho thấy quý 3 thường có doanh số cao nhất trong năm, doanh nghiệp có thể tập trung ngân sách quảng bá, tăng cường đội ngũ bán hàng và đẩy mạnh sản xuất trong quý 2 để chuẩn bị cho nhu cầu tăng cao. Ngược lại, các tháng thấp điểm có thể là thời gian thích hợp để tối ưu vận hành, đào tạo nhân viên hoặc giảm chi phí tạm thời.

Thứ ba là khả năng tối ưu chuỗi cung ứng. Với dữ liệu dự báo chính xác theo thời gian, doanh nghiệp có thể lập kế hoạch sản xuất và phân phối hiệu quả hơn. Điều này

giúp hạn chế tối đa tình trạng hết hàng trong mùa cao điểm hoặc tồn kho quá mức trong mùa thấp điểm – hai yếu tố ảnh hưởng trực tiếp đến chi phí và mức độ hài lòng của khách hàng. Chẳng hạn, nếu dự báo cho thấy doanh số sẽ sụt giảm nhẹ trong tháng 2, doanh nghiệp có thể điều chỉnh mức đặt hàng nguyên liệu hoặc giảm công suất sản xuất để tiết kiệm chi phí vận hành.

Cuối cùng, quan trọng vẫn là cải thiện hiệu quả tài chính. Việc lập kế hoạch tài chính sẽ trở nên sát với thực tế hơn nhờ các dự báo định lượng. Doanh nghiệp có thể dự tính doanh thu, biên lợi nhuận kỳ vọng và cân nhắc các phương án đầu tư dựa trên thông tin có độ tin cậy cao. Ví dụ, nếu dự báo cho thấy xu hướng tăng trưởng ổn định trong nửa cuối năm, doanh nghiệp có thể mạnh dạn mở rộng quy mô hoặc đàm phán tốt hơn với đối tác cung ứng, thay vì e ngại do thiếu thông tin chắc chắn về nhu cầu.

5.3. NHƯỢC ĐIỂM RÚT RA TỪ QUÁ TRÌNH DỰ BÁO DOANH SỐ

Mặc dù mô hình LSTM mang lại hiệu quả cao trong việc dự báo chuỗi thời gian, vẫn tồn tại một số hạn chế cần được xem xét. Trước hết là độ nhạy của mô hình với chất lượng dữ liệu đầu vào. Nếu dữ liệu lịch sử có chứa nhiễu, thiếu giá trị hoặc phân bố không đồng đều theo thời gian, mô hình có thể học sai xu hướng, từ đó đưa ra dự báo chưa chính xác tại một số thời điểm cụ thể. Ví dụ, nếu trong giai đoạn dữ liệu có các sự kiện bất thường như chương trình khuyến mãi mạnh, đại dịch hay biến động thị trường nhưng không được chú thích rõ, mô hình sẽ coi đó là xu hướng bình thường, dẫn đến sai lệch trong tương lai.

Tiếp theo là tính khó giải thích của mô hình. Mặc dù LSTM cho kết quả dự báo tốt, nhưng việc lý giải tại sao mô hình đưa ra một giá trị cụ thể tại thời điểm nào đó lại không dễ dàng như với các mô hình tuyến tính. Điều này khiến cho việc trình bày kết quả dự báo với các phòng ban phi kỹ thuật hoặc nhà quản lý cấp cao trở nên khó khăn hơn, nếu không có công cụ trực quan hóa và phân tích bổ sung.

Ngoài ra, mô hình cũng yêu cầu dữ liệu đầu vào có độ dài đủ lớn và ổn định để học được các chu kỳ, xu hướng ngắn và dài hạn. Trong trường hợp dữ liệu quá ngắn hoặc thay đổi bất thường qua các giai đoạn, hiệu quả của mô hình sẽ bị ảnh hưởng. Đặc biệt là với những tập dữ liệu có tính mùa vụ rõ rệt, nhưng mẫu dữ liệu chưa bao phủ đủ chu kỳ, mô hình có thể bỏ qua hoặc đánh giá sai các điểm biến động quan trọng.

Việc triển khai các mô hình LSTM đòi hỏi kiến thức chuyên sâu về học sâu (deep learning), hiểu biết về cách xử lý chuỗi thời gian, cùng với phần mềm và phần cứng phù

hợp. Nếu doanh nghiệp chưa có đội ngũ kỹ thuật đủ năng lực, việc triển khai có thể gặp khó khăn hoặc cần thuê ngoài, dẫn đến chi phí cao. Ngoài ra, việc huấn luyện mô hình với dữ liệu lớn cần thời gian xử lý dài và tài nguyên máy tính mạnh (GPU, bộ nhớ lớn), đặc biệt nếu doanh nghiệp muốn cập nhật mô hình liên tục theo dữ liệu mới.

5.4. ĐỊNH HƯỚNG PHÁT TRIỂN VÀ ĐỀ XUẤT CHO DOANH NGHIỆP

Để khai thác tối đa giá trị từ các mô hình dự báo như LSTM, doanh nghiệp cần đầu tư vào hệ thống quản lý dữ liệu để tăng cường chiến lược dữ liệu cải thiện chất lượng thu thập và xây dựng một chiến lược phát triển đồng bộ từ dữ liệu, mô hình đến tổ chức vận hành.Việc chuẩn hóa quy trình thu thập, lưu trữ và làm sạch dữ liệu sẽ giúp đảm bảo chất lượng dữ liệu đầu vào, từ đó nâng cao độ chính xác của mô hình. Doanh nghiệp có thể thiết lập hệ thống tự động kiểm tra các điểm ngoại lệ trong doanh số theo từng tháng, hoặc áp dụng các chỉ số chất lượng dữ liệu như mức độ đầy đủ, tính nhất quán và tính hợp lệ.

Chất lượng dự báo phụ thuộc lớn vào chất lượng và độ phong phú của dữ liệu. Doanh nghiệp nên đầu tư xây dựng cơ sở dữ liệu tập trung, trong đó không chỉ lưu trữ doanh số theo thời gian mà còn tích hợp các yếu tố ảnh hưởng như: chiến dịch marketing, khuyến mãi, yếu tố mùa vụ, tình hình thị trường, phản hồi khách hàng, v.v. Việc làm giàu dữ liệu giúp mô hình học được mối liên hệ phức tạp và phản ánh thực tế kinh doanh tốt hơn.

Hơn nữa, thay vì chỉ dựa vào phân tích bên ngoài hoặc các mô hình dựng sẵn, chúng ta nên đầu tư phát triển đội ngũ phân tích dữ liệu nội bộ (data analyst, data scientist). Việc này không chỉ giúp nâng cao tính chủ động trong việc cập nhật mô hình mà còn hỗ trợ ra quyết định nhanh chóng trong các tình huống biến động. Ngoài ra, doanh nghiệp có thể huấn luyện đội ngũ kinh doanh hiểu cách đọc các biểu đồ dự báo, từ đó phối hợp nhịp nhàng hơn giữa kỹ thuật và vận hành.

Doanh nghiệp nên ứng dụng dự báo vào hoạch định chuỗi cung ứng và marketing, lồng ghép kết quả dự báo vào các hệ thống hoạch định doanh nghiệp để điều phối tồn kho, sản xuất và phân phối phù hợp theo từng khu vực và từng tháng. Thị trường luôn biến động, do đó việc duy trì hiệu quả mô hình dự báo là vô cùng cần thiết, các doanh nghiệp cần có quy trình định kỳ đánh giá lại độ chính xác của mô hình, cập nhật dữ liệu mới, và điều chỉnh kiến trúc hoặc thông số khi cần thiết. Việc này đảm bảo rằng hệ thống luôn phản ánh đúng thực tiễn và thích ứng với thay đổi nhanh chóng.

TÀI LIỆU THAM KHẢO

[1]. Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2015), Time Series Analysis: Forecasting and Control, Wiley, New Jersey.

[2]. Hyndman, R. J., & Athanasopoulos, G. (2018), Forecasting: Principles and Practice, OTexts, Australia.

[3]. Taylor, S. J., & Letham, B. (2018), "Forecasting at scale", The American Statistician, 72(1), pp. 37–45.

[4]. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013), An Introduction to Statistical Learning, Springer, New York.

[5]. Hochreiter, S., & Schmidhuber, J. (1997), "Long Short-Term Memory", Neural Computation, 9(8), pp. 1735–1780.

[6]. TS. Chu Bình Minh (Chủ biên), TS. Lê Xuân Huy, Tài liệu học tập Học máy cho ngành Khoa học dữ liệu. Trường Đại học Kinh tế - Kỹ thuật Công nghiệp, Năm 2025. Lưu hành nội bộ.