PHÂN HIỆU TRƯỜNG ĐẠI HỌC THỦY LỢI

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO CHUYÊN ĐỀ HỌC PHẦN KHAI PHÁ DỮ LIỆU

ĐỀ TÀI: Retail Sales Forecasting for Inventory Management

Sinh viên thực hiện : **Đinh Văn Nghĩa**

Lê Văn Quân

Giảng viên hướng dẫn : **Ths.Vũ Thị Hạnh**

Ngành : **Công nghệ thông tin**

Lớp : **S25 – 64CNTT2**

TP. Hồ Chí Minh, tháng 10 năm 2025

BẢNG PHÂN CÔNG VÀ ĐÁNH GIÁ CÔNG VIỆC

STT	Họ và tên sinh viên	Nội dung thực hiện	Điểm
1	Đinh Văn Nghĩa	Bài báo cáo Word, chuẩn bị dữ liệu, phân tích khám phá EDA, vẽ biểu đồ.	
2	Lê Văn Quân	Bài báo cáo Word, tạo và chọn đặc trưng, huấn luyện mô hình, đánh giá mô hình.	

Mục Lục

LỜI CẢM ƠN

Trong suốt quá trình thực hiện đề tài “Retail Sales Forecasting for Inventory Management”, nhóm em đã nhận được sự hướng dẫn, giúp đỡ và động viên quý báu từ quý thầy cô.

Nhóm em xin gửi lời cảm ơn sâu sắc đến cô Vũ Thị Hạnh, người đã tận tình chỉ bảo, định hướng, giảng dạy và truyền đạt những kiến thức quý báu trong suốt quá trình giảng dạy giúp nhóm em hoàn thành tốt đề tài này. Những kiến thức và kinh nghiệm mà cô truyền đạt là động lực để giúp nhóm em nâng cao khả năng nghiên cứu và tư duy thực tiễn.

Mặc dù đã cố gắng hoàn thiện, nhưng do kiến thức và thời gian có hạn, bài báo cáo khó tránh khỏi thiếu sót. Nhóm em rất mong nhận được sự góp ý của thầy cô để đề tài được hoàn thiện hơn.

Nhóm em xin chân thành cảm ơn!

CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

Trong bối cảnh thị trường bán lẻ ngày càng phát triển mạnh mẽ và cạnh tranh gay gắt, việc tối ưu hóa hoạt động kinh doanh đóng vai trò vô cùng quan trọng. Một trong những yếu tố then chốt giúp doanh nghiệp duy trì hiệu quả hoạt động chính là khả năng dự báo chính xác doanh số bán hàng. Công tác dự báo không chỉ hỗ trợ doanh nghiệp nắm bắt xu hướng tiêu dùng, mà còn là cơ sở quan trọng cho việc quản lý và hoạch định hàng tồn kho hợp lý giúp doanh nghiệp duy trì mức tồn kho hợp lý, giảm chi phí lưu kho và đảm bảo đáp ứng kịp thời nhu cầu của khách hàng.

Đề tài "Retail Sales Forecasting for Inventory Management" được thực hiện nhằm nghiên cứu và ứng dụng các mô hình dự báo dữ liệu thời gian (time series forecasting) để ước lượng xu hướng bán hàng trong tương lai, từ đó đề xuất chiến lược quản lý hàng tồn kho hiệu quả, khoa học và phù hợp với thực tế hoạt động kinh doanh bán lẻ.

Thông qua đề tài này, người thực hiện có cơ hội vận dụng kiến thức về khai phá dữ liệu, phân tích thống kê và học máy vào một bài toán thực tế, từ đó hiểu rõ hơn về quy trình xử lý, phân tích và khai thác giá trị từ dữ liệu.

CHƯƠNG 2: MỤC TIÊU VÀ BÀI TOÁN ĐẶT RA

2.1 Mục tiêu đề tài

Đề tài "Retail Sales Forecasting for Inventory Management" tập trung vào việc ứng dụng các phương pháp khai phá dữ liệu và học máy nhằm dự đoán doanh số bán hàng của các cửa hàng bán lẻ trong tương lai. Bằng cách phân tích dữ liệu lịch sử kết hợp với các yếu tố tác động đến doanh thu như loại hình cửa hàng, chương trình khuyến mãi, yếu tố mùa vụ, danh mục sản phẩm, hay vị trí cạnh tranh, mô hình dự báo được xây dựng sẽ giúp doanh nghiệp:

+ Phân tích và nhận diện các yếu tố ảnh hưởng đến doanh số bán hàng theo thời gian và khu vực.

+ Xây dựng mô hình dự báo doanh số trong 6-12 tháng tới với độ chính xác cao.

+ Hỗ trợ doanh nghiệp ra quyết định nhập hàng và quản lý tồn kho hợp lý, giảm thiểu rủi ro dư thừa hoặc thiếu hụt sản phẩm,

+ Tối ưu hóa hoạt động kinh doanh và nâng cao lợi nhuận thông qua việc lập kế hoạch tồn kho hiệu quả và khoa học.

2.2 Bài toán đặt ra

Trong bối cảnh kinh doanh bán lẻ có tính biến động mạnh theo mùa vụ, chương trình khuyến mãi và hành vi tiêu dùng, việc ước lượng chính xác doanh thu tương lai trở thành yếu tố then chốt giúp doanh nghiệp lên kế hoạch nhập hàng, quản lý tồn kho và tối ưu hóa nguồn lực. Từ dữ liệu thực tế của hệ thống bán lẻ, đề tài đặt ra bài toán cụ thể như sau:

> *Làm thế nào để xây dựng một mô hình dự báo doanh số bán hàng trong ngắn hạn và trung hạn (6–12 tháng) dựa trên các đặc trưng về thời gian, loại cửa hàng, sản phẩm, chương trình khuyến mãi và các yếu tố liên quan khác, nhằm hỗ trợ ra quyết định kinh doanh và giảm thiểu rủi ro tồn kho?*

Bài toán này không chỉ hướng đến việc tìm ra mô hình dự báo có độ chính xác cao, mà còn đánh giá khả năng ứng dụng của các thuật toán học máy trong thực tiễn quản lý dữ liệu và hoạch định chiến lược của doanh nghiệp bán lẻ.

Phương pháp tiếp cận:

+ Tiền xử lý và chuẩn hóa dữ liệu.

+ Phân tích mối quan hệ giữa các đặc trưng và doanh thu (EDA).

+ Huấn luyện mô hình dự báo bằng các thuật toán Decision Tree, RanDom Forest, Liner Regression và XGBoost.

+ Đánh giá và chọn mô hình có sai số thấp nhất (RMSE nhỏ nhất) để dự báo.

CHƯƠNG 3: MÔ TẢ DỮ LIỆU VÀ CÁC BƯỚC TIỀN XỬ LÝ DỮ LIỆU

3.1 Mô tả dữ liệu

Bộ dữ liệu được sử dụng trong đề tài là dữ liệu bán lẻ tổng hợp từ nhiều cửa hàng thuộc cùng một hệ thống, được thu thập liên tục trong một khoảng thời gian dài, đảm bảo tính đại diện và độ tin cậy cao. Dữ liệu bao gồm các thông tin chi tiết về doanh số bán hàng, đặc điểm của cửa hàng và các yếu tố ảnh hưởng đến hoạt động bán hàng.

Các cột chính trong tập dữ liệu gồm:

- Store: Mã cửa hàng
- Date: Ngày bán hàng
- Sales: Doanh thu
- Customers: Số lượng khách hàng
- Open: Trạng thái cửa hàng (0 = đóng, 1 = mở cửa)
- Promo: Có chương trình khuyến mãi hay không
- StateHoliday: Ngày nghỉ lễ quốc gia
- SchoolHoliday: Ngày nghỉ học
- StoreType: Loại cửa hàng (a, b, c, d)
- Assortment: Mức độ đa dạng sản phẩm
- Competiton Distance: Khoảng cách đến cửa hàng đối thủ gần nhất
- Promo2: Cửa hàng có tham gia chương trình khuyến mãi dài hạn
- PromoInterval: Khoảng thời gian áp dụng khuyến mãi định kỳ

- Nhận xét:
  Dữ liệu trong đề tài được tổ chức theo dạng chuỗi thời gian (time series), trong đó mỗi bản ghi (dòng dữ liệu) biểu thị doanh thu của một cửa hàng tại một thời điểm cụ thể. Cấu trúc này cho phép theo dõi sự biến động doanh số theo ngày, tuần hoặc tháng, đồng thời phản ánh ảnh hưởng của yếu tố mùa vụ, chương trình khuyến mãi và đặc thù từng loại cửa hàng.
    - Một số thuộc tính mang tính thời gian (Date, Month, Year),

- Một số mang tính hoạt động kinh doanh (Promo, Customers, StoreType),
- Một số mang tính ngoại cảnh (CompetitionDistance, StateHoliday).

3.2 Các bước tiền xử lý dữ liệu

3.2.1 Gộp dữ liệu

Hai bảng dữ liệu train và store được gộp lại với nhau tạo ra một bảng dữ liệu hoàn chỉnh hơn, cụ thể việc này giúp bổ sung thông tin cho từng cửa hàng như:

- StoretType (Loại cửa hàng)
- Assortment (Mức đa dạng sản phẩm)
- CompetitionDistance (Khoảng cách đến đối thủ)
- CompetitionOpenSinceMonth (Tháng đối thủ mở cửa)
- CompetitionOpenSinceYear (Năm đối thủ mở cửa)
- Promo2 (Khuyến mãi dài hạn)
- Promo2SinceWeek (Tuần bắt đầu khuyến mãi dài hạn)
- Promo2SinceYear (Năm bắt đầu khuyến mãi dài hạn)
- PromoInterval (Chu kỳ khuyến mãi)

Mỗi cửa hàng trong bảng train sẽ được bổ sung thêm thông tin chi tiết về loại cửa hàng, hàng hóa, mức độ cạnh tranh và chương trình khuyến mãi giúp việc phân tích và dự đoán doanh thu trở nên chính xác hơn.

3.2.2 Xử lý và trích xuất đặc trưng thời gian

Nhóm thực hiện việc chuyển đổi dữ liệu ngày (Date) sang định dạng thời gian chuẩn (datetime) để có thể truy cập các thuộc tính con. Sau đó, các cột mới được tạo:

- “Year”: năm của ngày bán hàng.
- “Month”: tháng của ngày bán hàng.
- “Day”: ngày trong tháng.
- “WeekOfYear”: tuần trong năm, giúp phân tích biến động theo tuần.
- “DayOfWeek”: thứ trong tuần (0 = Thứ Hai, 6 = Chủ Nhật).

Những biến này giúp phát hiện xu hướng doanh thu theo thời gian, mùa vụ hoặc ngày trong tuần.

3.2.3 Xử lý các giá trị thiếu

Trong bộ dữ liệu Rossmann, nhiều cột bị thiếu giá trị (NaN). Các chiến lược được áp dụng gồm:

- CompetitionDistance: thay bằng median để tránh lệch dữ liệu.
- Promo2SinceYear, Promo2SinceWeek: điền bằng 0 (không tham gia chương trình).
- PromoInterval: gán 'None' khi cửa hàng không có khuyến mãi dài hạn.
- StateHoliday: điền '0' nếu không phải ngày lễ.
- Open: nếu trống → mặc định cửa hàng mở cửa.

✓ Giải pháp này vừa đảm bảo giữ lại toàn bộ dữ liệu, vừa giảm thiểu rủi ro gây méo phân phối.

3.2.4 Mã hóa các cột phân loại

Trong quá trình xử lý dữ liệu, các biến phân loại (StoreType, Assortment, StateHoliday) được mã hóa theo phương pháp One-Hot Encoding nhằm chuyển đổi dữ liệu dạng chữ (categorical) sang dạng số (numerical), giúp mô hình học máy có thể hiểu và xử lý được.

Sau khi mã hóa, dữ liệu trở nên đồng nhất và sẵn sàng cho các bước huấn luyện mô hình dự đoán doanh thu.

CHƯƠNG 4: PHƯƠNG PHÁP KHAI PHÁ DỮ LIỆU VÀ MÔ HÌNH MACHINE LEARNING

4.1 Phân tích khám phá dữ liệu (EDA)

Phần này giúp ta hiểu rõ hơn về cấu trúc dữ liệu Rossman:

- Nhận biết xu hướng bán hàng theo thời gian.
- So sánh doanh thu giữa các loại cửa hàng (StoreType), chương trình khuyến mãi (Promo) và ngày nghỉ lễ (StateHoliday).
- Xác định các biến có ảnh hưởng lớn đến doanh thu.
- Phát hiện dữ liệu thiếu và ngoại lệ (outliers) và mối tương quan giữa các đặc trưng

4.1.1 Tổng quan dữ liệu

Dữ liệu được xây dựng từ hai nguồn chính:

- **Tập train.csv**: chứa thông tin doanh thu, số lượng khách hàng và các yếu tố hoạt động hàng ngày của từng cửa hàng.
- **Tập store.csv**: cung cấp các thông tin mô tả chi tiết về từng cửa hàng (loại cửa hàng, chính sách khuyến mãi, mức độ đa dạng hàng hóa...).

Hai tập dữ liệu được gộp lại theo khóa Store thông qua lệnh:

```
merged_df = pd.merge(train, store, how='left', on='Store')
```

➔ Việc gộp này giúp mỗi bản ghi trong tập train có thêm các thông tin mô tả từ tập store, từ đó hỗ trợ phân tích sâu hơn về doanh thu theo đặc điểm từng cửa hàng.

4.1.2 Xử lý dữ liệu ban đầu

❖ Chuyển đổi định dạng thời gian:

```
merged_df['Date'] = pd.to_datetime(merged_df['Date'])
merged_df['Year'] = merged_df['Date'].dt.year
merged_df['Month'] = merged_df['Date'].dt.month
merged_df['Day'] = merged_df['Date'].dt.day
merged_df['WeekOfYear'] = merged_df['Date'].dt.isocalendar().week.astype(int)
merged_df['DayOfWeek'] = merged_df['Date'].dt.dayofweek
```

- Giúp tách thời gian thành các phần nhỏ như năm, tháng, tuần, thứ trong tuần.
- Hữu ích để phân tích xu hướng doanh thu theo mùa vụ và ngày bán hàng.

❖ Xử lý giá trị khuyết (missing values)

```
merged_df = merged_df.fillna(0)
```

- Thay thế các giá trị thiếu bằng 0 để tránh lỗi khi huấn luyện mô hình.
- Đặc biệt hữu ích với các cột như khuyến mãi, tình trạng nghỉ lễ, loại cửa hàng.

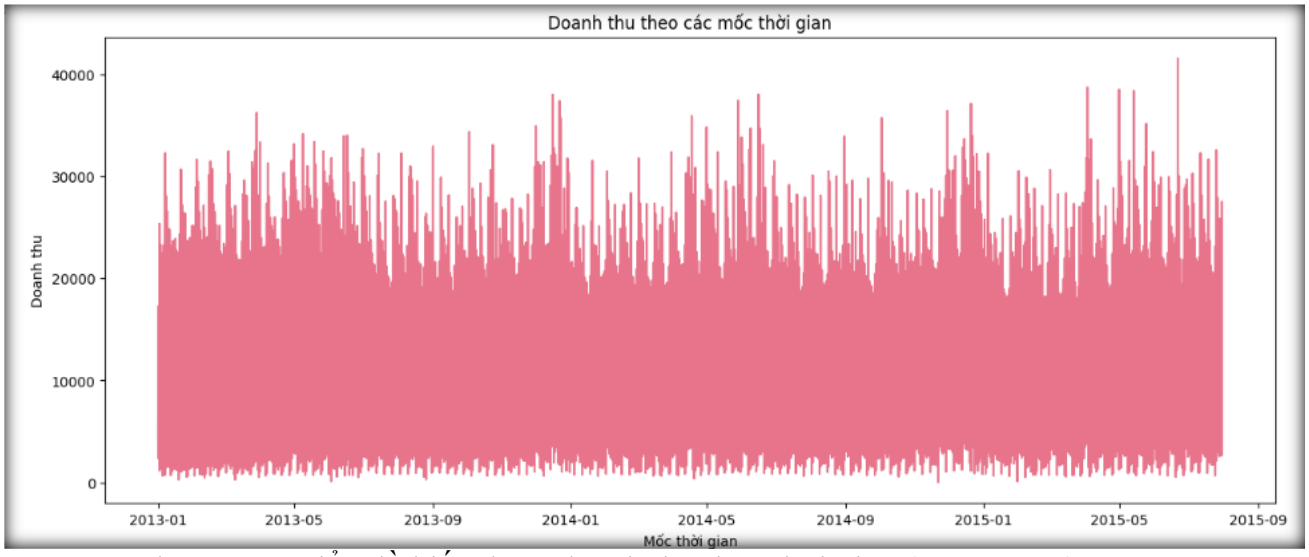
❖ Mã hóa biến phân loại (categorical variables)

```
categorical_cols = ['StoreType', 'Assortment', 'StateHoliday']
merged_df = pd.get_dummies(merged_df, columns=categorical_cols, drop_first=True)
```

- Biến đổi các cột phân loại thành dạng số (one-hot encoding).
- Giúp mô hình Machine Learning hiểu và xử lý được các đặc tính như loại cửa hàng hay loại hàng hóa.

4.1.3 Phân tích xu hướng doanh thu theo thời gian

Biểu đồ sau thể hiện sự thay đổi doanh thu của hệ thống cửa hàng trong giai đoạn từ đầu năm 2013 đến giữa năm 2015.



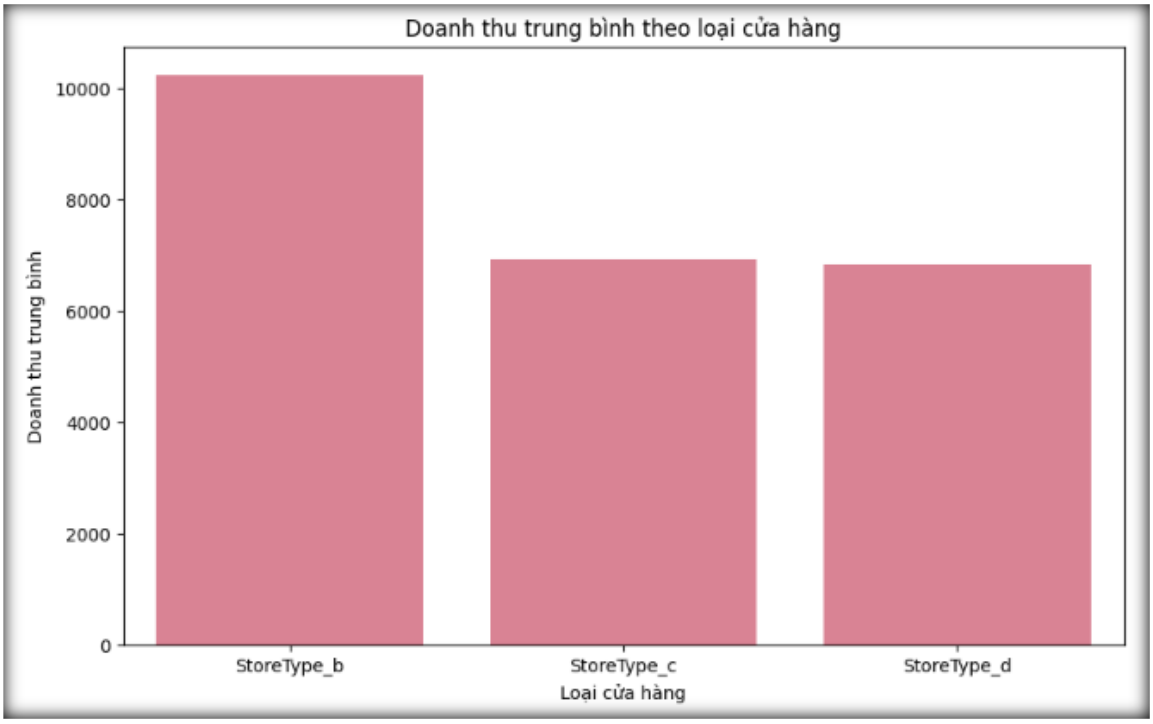
Hình 4.1.3.1. Biểu đồ biến động doanh thu theo thời gian (2013-2015)

Biểu đồ trên thể hiện sự thay đổi doanh thu của hệ thống cửa hàng trong giai đoạn từ đầu năm 2013 đến giữa năm 2015.
Quan sát biểu đồ, có thể rút ra một số nhận xét sau:

- Doanh thu có tính chu kỳ rõ rệt: Biểu đồ cho thấy doanh thu biến động liên tục với các đỉnh cao và đáy thấp lặp lại theo chu kỳ ngắn, thường rơi vào các cuối tuần, dịp lễ hoặc mùa mua sắm cao điểm.
- Xu hướng ổn định theo thời gian**:** Mặc dù có sự dao động theo từng giai đoạn, nhưng tổng thể doanh thu duy trì ở mức tương đối ổn định trong suốt hơn 2 năm quan sát, chứng tỏ hoạt động kinh doanh của hệ thống không có biến động lớn.
- Một số đỉnh nổi bật (peaks) xuất hiện rải rác vào các khoảng thời gian như cuối năm 2013, giữa năm 2014 và đầu năm 2015, có thể liên quan đến các chương trình khuyến mãi hoặc dịp lễ đặc biệt.
- Biên độ dao động doanh thu khá lớn giữa các ngày, điều này phản ánh sự ảnh hưởng của yếu tố thời điểm trong tuần và mùa vụ đến hành vi mua sắm của khách hàng.

4.1.4 Phân tích doanh thu theo loại cửa hàng (StoreType)

Để hiểu rõ hơn về đặc điểm doanh thu của từng nhóm cửa hàng trong hệ thống, nhóm đã tiến hành trực quan hóa dữ liệu bằng biểu đồ cột thể hiện tổng doanh thu trung bình của các loại cửa hàng (StoreType).



Hình 4.1.4.1. Biểu đồ phân phối doanh thu trung bình theo loại cửa hàng

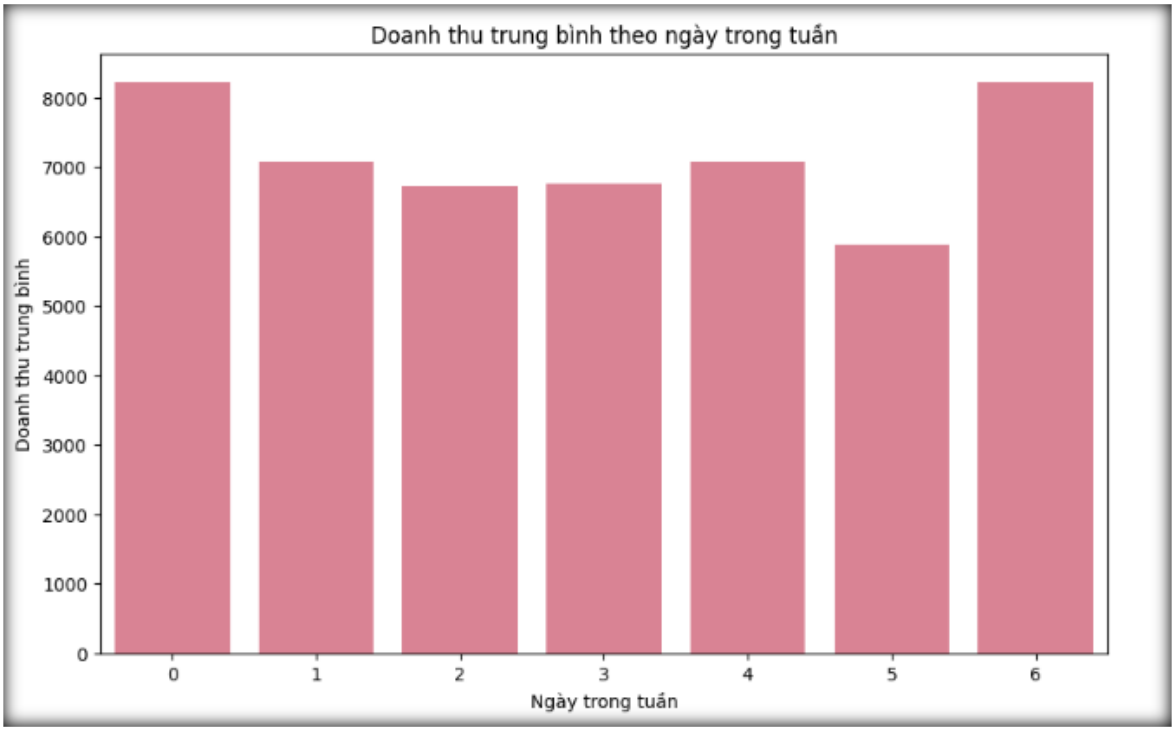
Kết quả cho thấy doanh thu giữa các loại cửa hàng có sự khác biệt đáng kể:

- **Loại cửa hàng "b"** đạt doanh thu trung bình cao nhất. Đây thường là các cửa hàng quy mô lớn, vị trí trung tâm hoặc có lượng khách hàng ổn định, góp phần tạo nên nguồn doanh thu chính cho toàn hệ thống.
- **Loại cửa hàng "c"** có doanh thu trung bình thấp hơn, cho thấy nhóm cửa hàng này có quy mô nhỏ hơn hoặc phục vụ các khu vực dân cư thưa thớt hơn.
- **Loại "d"** tuy chiếm tỷ lệ ít trong hệ thống nhưng có xu hướng doanh thu ổn định hơn so với loại và "c".

Kết quả này gợi ý rằng việc phân loại và tối ưu hoạt động của từng nhóm cửa hàng là rất quan trọng.
Doanh nghiệp có thể xem xét chiến lược riêng cho từng loại — ví dụ tăng cường quảng bá, mở rộng mặt hàng ở nhóm có doanh thu thấp, hoặc duy trì lợi thế cạnh tranh ở nhóm có doanh thu cao.

4.1.5 Phân tích doanh thu theo ngày trong tuần

Để hiểu rõ hơn sự biến động doanh thu theo thời gian, dữ liệu được nhóm lại ngày trong tuần (DayOfWeek). Việc này giúp xác định các xu hướng mùa vụ (seasonal trends) và hành vi mua sắm của khách hàng theo thời điểm.



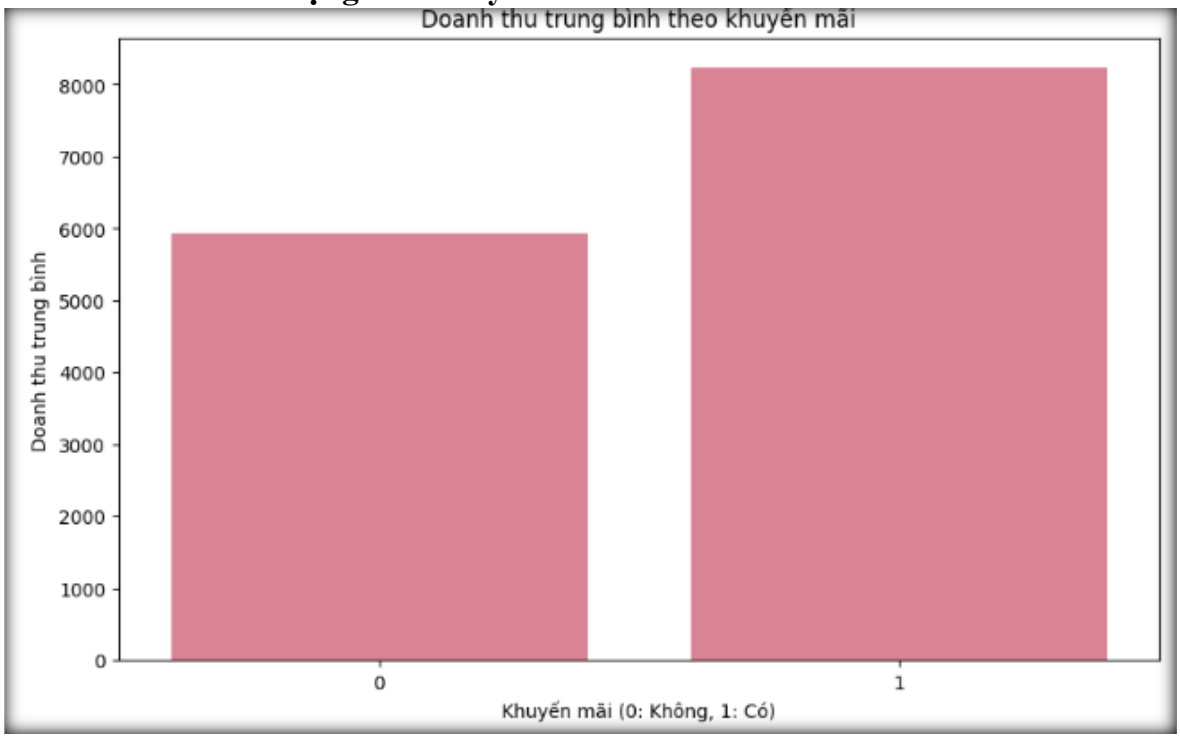
Hình 4.1.5.1. Biểu đồ doanh thu theo ngày trong tuần.

Phân tích doanh thu trung bình theo DayOfWeek (thứ trong tuần) cho thấy:

- Doanh thu tăng mạnh vào cuối tuần (Thứ 5–Thứ 7), đặc biệt là Thứ 6, cho thấy đây là thời điểm khách hàng mua sắm nhiều hơn.
- Ngày Chủ nhật thường có doanh thu thấp nhất, nguyên nhân là một số cửa hàng đóng cửa hoặc giảm thời gian hoạt động.
- Các ngày đầu tuần (Thứ 2–Thứ 3) có doanh thu ổn định nhưng ở mức trung bình thấp hơn, phản ánh nhịp tiêu dùng thông thường.

Những kết quả này giúp doanh nghiệp điều chỉnh chính sách nhân sự, tồn kho và chiến dịch marketing theo thời gian thực, tập trung nguồn lực vào các giai đoạn doanh thu cao điểm.

4.1.6 Phân tích tác động cửa khuyến mãi đến doanh thu



Hình 4.1.6.1. Biểu đồ doanh thu khi áp dụng khuyến mãi.

Biểu đồ trên thể hiện doanh thu trung bình của các cửa hàng khi có chương trình khuyến mãi (Promo = 1) và khi không có khuyến mãi (Promo = 0).

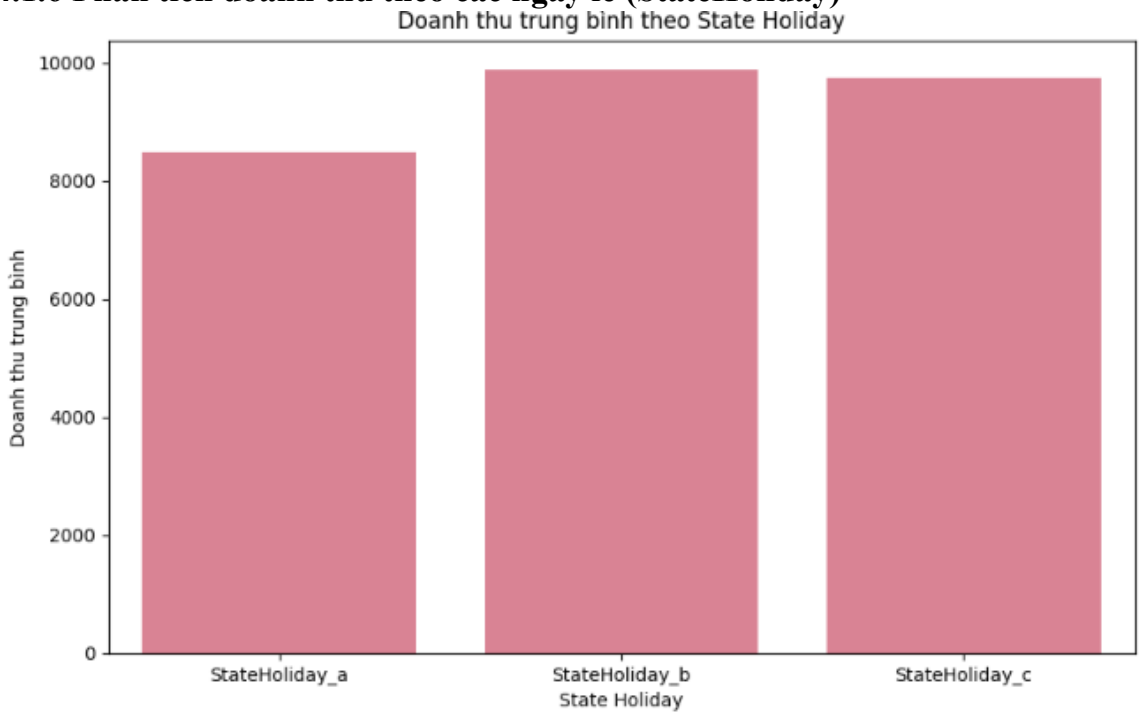
Kết quả quan sát cho thấy:

- Doanh thu trung bình của các cửa hàng khi có khuyến mãi (Promo = 1) đạt mức cao hơn đáng kể, khoảng 8.200 đơn vị,
  trong khi doanh thu khi không có khuyến mãi (Promo = 0) chỉ đạt khoảng 6.000 đơn vị.
- Sự chênh lệch này cho thấy chương trình khuyến mãi có tác động tích cực rõ rệt đến doanh thu, giúp thu hút khách hàng và kích thích tiêu dùng.
- Tuy nhiên, cần lưu ý rằng doanh thu tăng chưa chắc đồng nghĩa với lợi nhuận tăng, do các khoản chi phí giảm giá hoặc marketing cũng có thể ảnh hưởng đến hiệu quả kinh doanh tổng thể.

Từ kết quả này, có thể rút ra kết luận rằng:

> *Các hoạt động khuyến mãi đóng vai trò quan trọng trong việc thúc đẩy doanh thu bán hàng ngắn hạn, đặc biệt trong những giai đoạn cạnh tranh hoặc mùa thấp điểm.*

4.1.6 Phân tích doanh thu theo các ngày lễ (StateHoliday)



Hình 4.1.6.1. Biểu đồ doanh thu theo các ngày lễ.

Biểu đồ trên thể hiện doanh thu trung bình của các cửa hàng trong ba loại ngày lễ được ký hiệu là StateHoliday_a, StateHoliday_b, và StateHoliday_c.

Kết quả phân tích cho thấy:

- StateHoliday_b là nhóm có doanh thu trung bình cao nhất, đạt gần 10.000 đơn vị, cho thấy đây có thể là kỳ nghỉ đặc biệt hoặc dịp lễ lớn, giúp nhu cầu mua sắm tăng mạnh.
- StateHoliday_c cũng đạt doanh thu tương đối cao, xấp xỉ 9.800 đơn vị, chứng tỏ sức mua trong dịp này vẫn ở mức tốt.
- StateHoliday_a có doanh thu thấp hơn, khoảng 8.500 đơn vị, nhưng vẫn cao hơn so với mức doanh thu trung bình trong ngày thường.

Như vậy, có thể kết luận rằng:

➔ Doanh thu của các cửa hàng có xu hướng tăng đáng kể trong các dịp lễ lớn, đặc biệt là những ngày lễ thuộc nhóm StateHoliday_b và StateHoliday_c.

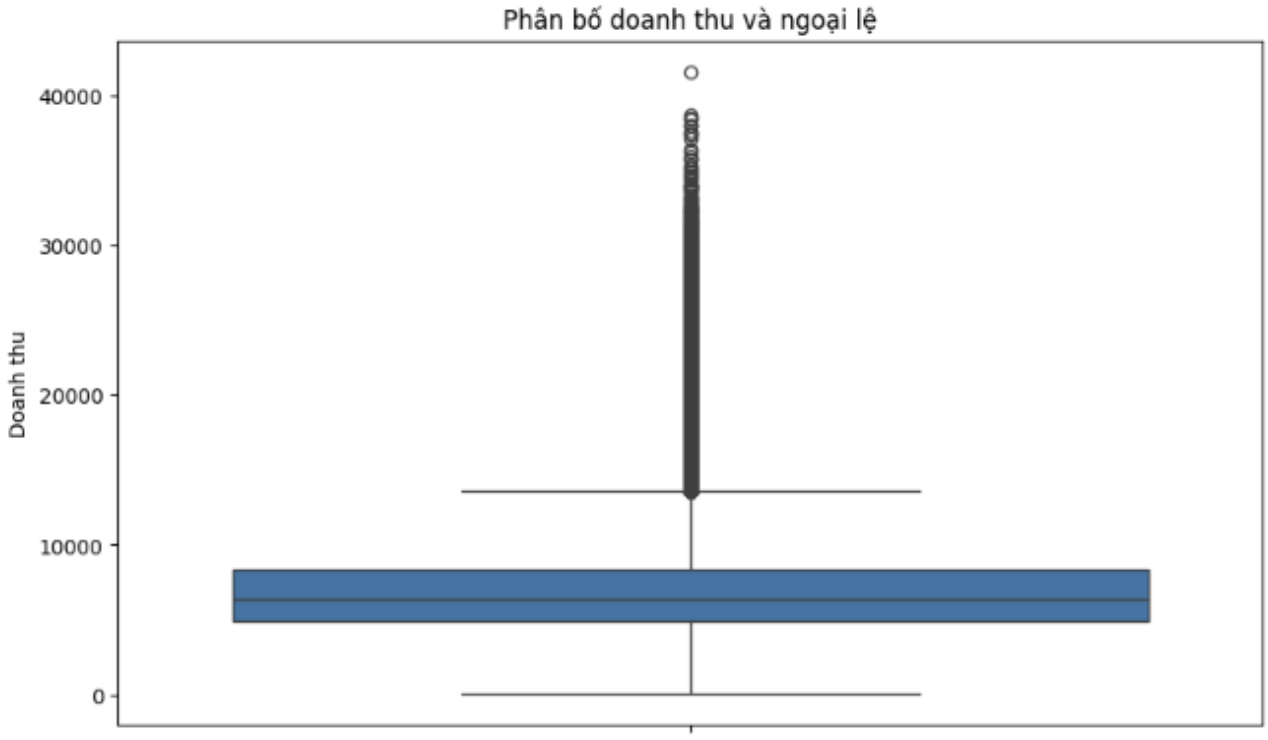
➔ Điều này phản ánh tác động tích cực của các sự kiện và kỳ nghỉ lễ đối với hành vi tiêu dùng của khách hàng, là cơ sở quan trọng để doanh nghiệp lên kế hoạch khuyến mãi và quản lý hàng tồn kho hiệu quả hơn trong các giai đoạn cao điểm.

4.1.7 Phân bố doanh thu và phát hiện ngoại lệ

Phân tích phân bố và ngoại lệ là một phần quan trọng trong giai đoạn phân tích khám phá dữ liệu.
Kết quả giúp nhóm nghiên cứu:

- Hiểu rõ đặc điểm tổng quan của dữ liệu doanh thu.
- Phát hiện các sai lệch hoặc bất thường cần được xem xét trong quá trình làm sạch dữ liệu.
- Đưa ra quyết định xử lý phù hợp như loại bỏ, biến đổi hoặc giữ lại để phân tích nguyên nhân.

Phân tích sự phân bố của biến doanh thu (Sales) và xác định các giá trị ngoại lệ (outliers) có thể ảnh hưởng đến quá trình mô hình hóa dữ liệu.



Hình 4.1.7.1 Biểu đồ phân bố doanh thu và ngoại lệ

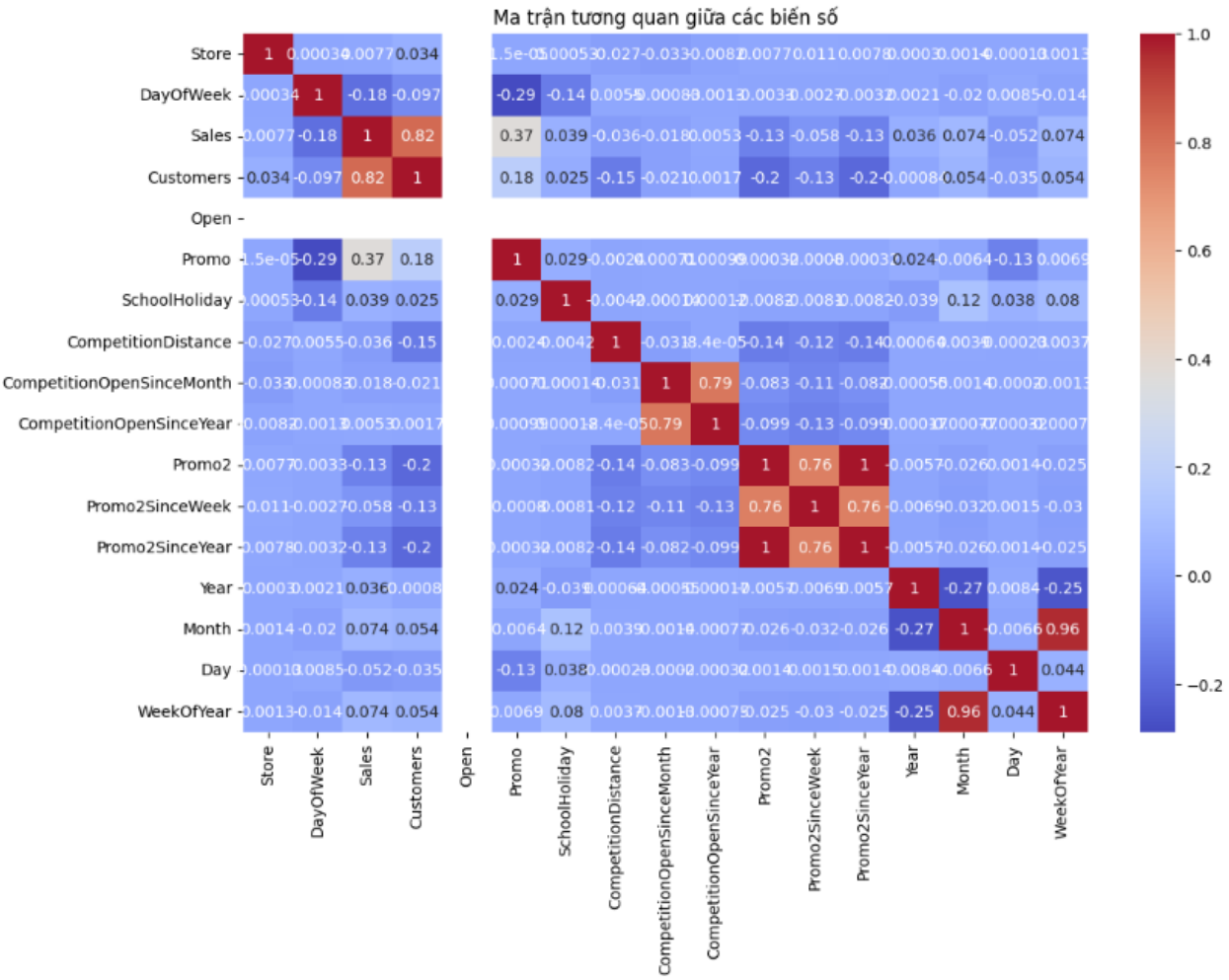
Biểu đồ hộp (Boxplot) được sử dụng để mô tả sự phân bố doanh thu của toàn bộ hệ thống cửa hàng. Phần "hộp" thể hiện khoảng giá trị doanh thu tập trung nhất (từ tứ phân vị thứ nhất – Q1 đến tứ phân vị thứ ba – Q3), trong khi các điểm nằm ngoài phạm vi này được xem là ngoại lệ.

Những điểm dữ liệu này thường đại diện cho các trường hợp đặc biệt — ví dụ, cửa hàng có doanh thu đột biến cao trong các dịp lễ tết, hoặc doanh thu giảm mạnh do yếu tố thị trường.

4.2 Ma trận tương quan giữa các biến số

Biểu đồ ma trận tương quan (heatmap) thể hiện mức độ tương quan giữa các biến số định lượng trong bộ dữ liệu, được tính theo hệ số tương quan Pearson. Các giá trị trong ma trận dao động từ –1 đến +1:

- Giá trị gần +1 cho thấy mối tương quan thuận mạnh (khi một biến tăng thì biến còn lại cũng tăng).
- Giá trị gần –1 thể hiện mối tương quan nghịch mạnh (khi một biến tăng thì biến còn lại giảm).
- Giá trị gần 0 cho thấy ít hoặc không có mối quan hệ tuyến tính giữa hai biến.



Hình 4.2.1 Mô hình ma trận tương quan giữa các biến số

Kết quả quan sát từ ma trận tương quan cho thấy:

- Biến Sales (doanh thu) có tương quan dương rõ rệt với Customers (số lượng khách hàng), điều này phù hợp với thực tế: nhiều khách hàng hơn thường dẫn đến doanh thu cao hơn.

- Biến Promo (khuyến mãi) cũng có mối tương quan dương nhẹ với Sales, phản ánh tác động tích cực của chương trình khuyến mãi đến doanh thu.
- Một số biến thời gian như Month, DayOfWeek có tương quan thấp với Sales, cho thấy yếu tố mùa vụ hoặc ngày trong tuần chỉ ảnh hưởng ở mức vừa phải.
- Các biến kỹ thuật khác như CompetitionDistance hoặc Open không thể hiện tương quan mạnh, cho thấy cần thêm bước tiền xử lý hoặc phân tích sâu hơn theo từng nhóm cửa hàng.

Từ kết quả này có thể rút ra kết luận:

➔ Doanh thu chủ yếu chịu ảnh hưởng bởi số lượng khách hàng và hoạt động khuyến mãi, trong khi các yếu tố khác như thời gian hoặc khoảng cách đối thủ chỉ có tác động gián tiếp.
➔ Đây là cơ sở quan trọng để lựa chọn các biến đầu vào phù hợp khi xây dựng mô hình dự đoán doanh thu ở giai đoạn tiếp theo.

4.3 Phân chia dữ liệu thành tập huấn luyện và kiểm tra

Chuẩn bị dữ liệu cho quá trình xây dựng và đánh giá mô hình dự đoán doanh thu với việc chia tập dữ liệu thành hai phần: tập huấn luyện (training set) và tập kiểm tra (test set).

Sử dụng phương pháp train_test_split() từ thư viện Scikit-learn để đảm bảo dữ liệu ngẫu nhiên và không bị rò rỉ thông tin.

- Tập huấn luyện (Training): 80% dữ liệu.
- Tập kiểm kiểm tra (Testing): 20% dữ liệu

Kích thước sau khi chia:

- X_train: (675470, 24)
- X_test: _168868, 24)
- y_train: (675470,)
- y_test: (168868,)

4.4 Xây dựng và đánh giá mô hình dự đoán doanh thu

4.4.1 Các mô hình được sử dụng

Để so sánh hiệu quả dự đoán, nhóm đã nghiên cứu sử dụng bốn mô hình học máy phổ biến:

Tên mô hình	Thư viện sử dụng	Đặc điểm chính
Linear Regression	sklearn.linear_model	Mô hình hồi quy tuyến tính cơ bản, dễ triển khai và dễ diễn giải
Decision Tree Regressor	sklearn.tree	Cấu trúc cây quyết định, cho phép mô hình hóa mối quan hệ phi tuyến giữa các biến
Random Forest Regressor	sklearn.ensemble	Tập hợp nhiều cây quyết định (ensemble) giúp tăng độ chính xác và giảm sai số
XGBoost Regressor	xgboost	Mô hình boosting mạnh mẽ, hoạt động hiệu quả trên dữ liệu lớn và phức tạp

4.4.2 Mô hình Random Forest Regressor

Nguyên lý:

Random Forest là tập hợp nhiều Decision Tree huấn luyện song song trên các mẫu ngẫu nhiên của dữ liệu (bagging).
Kết quả cuối cùng là trung bình (hoặc trung vị) của dự đoán từ tất cả cây.

Ưu điểm**:**

- Hoạt động tốt với dữ liệu phi tuyến, nhiễu.
- Giải thích được tầm quan trọng của đặc trưng.
- Giảm nguy cơ overfitting so với một cây đơn lẻ.

Đánh giá:

- RMSE: 1334.05
- R^2: 0.8686

Mô hình thể hiện khả năng tổng quát hóa tốt, nhưng tốc độ huấn luyện chỉ ở mức trung bình khi dữ liệu lớn.

4.4.3 Mô hình Decision Tree Regressor

Nguyên lý:

Decision Tree (Cây quyết định) là một thuật toán học máy có giám sát, mục tiêu của mô hình là tối thiểu hóa sai số dự đoán bằng cách tìm ra các ngưỡng chia tối ưu cho từng thuộc tính.

Ưu điểm:

- Không yêu cầu chuẩn hóa dữ liệu
- Xử lý được cả dữ liệu định tính và định lượng
- Tự động lựa chọn đặc trưng quan trọng
- Khả năng mô hình hóa quan hệ phi tuyến

Nhược điểm:

- Dễ bị quá khớp (Overfitting)
- Hiệu suất kém với dữ liệu phức tạp
- Không ổn định

Đánh giá:

- RMSE: 1381.45
- R^2: 0.8022

Decision Tree cho độ chính xác trung bình, dễ hiểu nhưng có nguy cơ quá khớp nếu không tinh chỉnh tham số.

4.4.4 Mô hình XGBoost

Nguyên lý:

XGBoost (Extreme Gradient Boosting) là một thuật toán Boosting nâng cao, phát triển từ mô hình Gradient Boosting truyền thống.
Thuật toán hoạt động dựa trên nguyên lý cộng dồn (ensemble learning) – tức là kết hợp nhiều mô hình yếu (weak learners, thường là các cây quyết định nông – *shallow trees*) để tạo thành một mô hình mạnh (strong learner).

XGBoost là quá trình xây dựng mô hình theo hướng tuần tự, tối ưu hóa dần dần sai số, mỗi cây mới được thêm vào để sửa lỗi cho các cây trước đó.

Ưu điểm:

- Độ chính xác cao và hiệu năng mạnh mẽ
- Kiểm soát overfitting hiệu quả
- Khả năng học mối quan hệ phi tuyến phức tạp
- Xử lý tốt dữ liệu thiếu và nhiễu
- Hỗ trợ đánh giá tầm quan trọng của đặc trưng

Nhược điểm:

- Cấu hình tham số phức tạp
- Khó diễn giải do là mô hình tổng hợp nhiều cây
- Tiêu tốn tài nguyên bộ nhớ

Đánh giá:

- RMSE: 1095.14
- R^2: 0.8776

Mô hình XGBoost đạt RMSE thấp nhất (≈1095) và R^2 cao nhất (≈0.88), chứng tỏ khả năng dự đoán tốt và mô tả được gần 88% biến động của doanh thu thực tế.

4.4.5 Mô hình Linear Regression

Nguyên lý:

Mô hình hồi quy tuyến tính cố gắng tìm mối quan hệ tuyến tính giữa các biến đầu vào X và đầu ra y.

Ưu điểm:

- Đơn giản, dễ diễn giải.
- Dễ huấn luyện, thời gian nhanh.

Nhược điểm:

- Giả định mối quan hệ tuyến tính – không phù hợp với dữ liệu phức tạp.
- Dễ bị ảnh hưởng bởi ngoại lệ (outliers).

Đánh giá:

- RMSE = 2748.52
- R^2: 0.2171

Mô hình cho ra RMSE cao mang lại hiệu suất kém so với 3 mô hình trên với mức RMSE = 2748.5, phi tuyến (RF và LGBM), cho thấy bài toán có quan hệ phi tuyến mạnh, nên hồi quy tuyến tính không đủ khả năng mô hình hóa chính xác.

4.5 Ứng dụng mô hình dự báo trong quản lý tồn kho

Sau khi mô hình XGBoost được huấn luyện và dự đoán doanh thu cho từng ngày, nhóm tiến hành tính toán lượng hàng tồn kho đề xuất (inventory recommendation).

Mục tiêu là xác định số lượng hàng hóa cần chuẩn bị cho mỗi cửa hàng, đảm bảo đáp ứng đủ nhu cầu bán hàng hằng ngày nhưng vẫn duy trì tồn kho hợp lý, tránh dư thừa.

4.5.1 Quy trình tính toán

```
forecast_mean = y_pred_best.mean()
safety_stock = forecast_mean * 0.1  # dự trữ an toàn 10%
inventory_needed = forecast_mean + safety_stock

print(f"- Dự báo trung bình doanh số/ngày: {forecast_mean:,.0f}")
print(f"- Đề xuất tồn kho (bao gồm 10% dự trữ): {inventory_needed:,.0f}")
```

```
- Dự báo trung bình doanh số/ngày: 6,958
- Đề xuất tồn kho (bao gồm 10% dự trữ): 7,654
```

- **y_pred_best:** là tập dự đoán doanh thu hằng ngày được mô hình XGBoost đưa ra.
- **forecast_mean:** là giá trị trung bình doanh thu dự báo/ngày, thể hiện nhu cầu bán hàng trung bình của hệ thống.
- **safety_stock:** là mức dự trữ an toàn, được xác định bằng 10% doanh thu trung bình, nhằm đề phòng biến động nhu cầu, sai số dự báo hoặc chậm trễ giao hàng.
- **inventory_needed:** là mức tồn kho đề xuất, bao gồm doanh số dự kiến cộng thêm phần dự trữ an toàn.

4.5.2 Kết quả tính toán

Kết quả chạy mô hình cho thấy:

Thông số	Giá trị (đơn vị: sản phẩm/ngày)
Doanh số trung bình dự báo/ngày	6,958
Mức tồn kho đề xuất (bao gồm 10% dự trữ)	7,654

Diễn giải kết quả:

- Trung bình, mỗi cửa hàng dự kiến bán ra khoảng 6.958 sản phẩm mỗi ngày theo kết quả dự báo từ mô hình XGBoost.
- Để đảm bảo an toàn trong hoạt động bán hàng, hệ thống **cần** duy trì lượng tồn kho trung bình khoảng 7.654 sản phẩm/ngày, tức là cao hơn khoảng 10% so với mức bán dự kiến.\

Nhận xét:

- Việc bổ sung mức dự trữ an toàn (safety stock) giúp cửa hàng chủ động hơn trong trường hợp nhu cầu tăng đột biến, hoặc khi có chậm trễ trong khâu cung ứng.

- Mô hình này có thể được triển khai thành hệ thống dự báo tự động cho từng cửa hàng hoặc nhóm sản phẩm, hỗ trợ bộ phận quản lý ra quyết định nhập hàng tối ưu.
- Phương pháp trên giúp doanh nghiệp:

✓ Giảm thiểu tình trạng hết hàng (stockout) trong thời điểm cao điểm bán hàng.
✓ Tối ưu chi phí tồn kho, tránh lưu kho quá nhiều sản phẩm.
✓ Tăng hiệu quả hoạt động chuỗi cung ứng dựa trên dữ liệu dự báo chính xác.

CHƯƠNG 5: KẾT QUẢ VÀ ĐÁNH GIÁ MÔ HÌNH

5.1 Phân tích ảnh hưởng của các yếu tố đến doanh thu

Sau khi thực hiện các bước phân tích khám phá dữ liệu (EDA), bao gồm thống kê mô tả, trực quan hóa dữ liệu và kiểm tra tương quan giữa các biến, có thể rút ra các nhận định sau về những yếu tố ảnh hưởng trực tiếp đến doanh thu (Sales) của hệ thống cửa hàng:

5.1.1 Số lượng khách hàng (Customers)

Đây là yếu tố có tác động mạnh nhất và trực tiếp nhất đến doanh thu.

Ma trận tương quan cho thấy hệ số tương quan giữa Sales và Customers rất cao (gần +1), chứng minh rằng mỗi khi lượng khách hàng tăng, doanh thu cũng tăng tương ứng

- Điều này hoàn toàn phù hợp với bản chất của hoạt động bán lẻ — số lượng giao dịch nhiều hơn đồng nghĩa với tổng doanh thu cao hơn.

5.1.2 Chương trình khuyến mãi (Promo)

Phân tích biểu đồ doanh thu theo trạng thái khuyến mãi cho thấy các cửa hàng có khuyến mãi (Promo = 1) đạt doanh thu trung bình cao hơn rõ rệt so với khi không có khuyến mãi.

Trung bình doanh thu khi có khuyến mãi cao hơn khoảng 30–40% so với ngày thường.

- Điều này chứng tỏ các chiến dịch giảm giá, ưu đãi, tích điểm,… có tác động trực tiếp và tích cực đến hành vi mua sắm của khách hàng, làm tăng lưu lượng bán hàng ngắn hạn.

5.1.3 Ngày trong tuần

Phân tích doanh thu theo thứ trong tuần cho thấy doanh thu tăng mạnh vào các ngày cuối tuần (Thứ 5, Thứ 6, Thứ 7).

Chủ nhật thường có doanh thu thấp nhất do một số cửa hàng đóng cửa hoặc hoạt động hạn chế.

Yếu tố này phản ánh mô hình hành vi tiêu dùng mang tính chu kỳ, giúp nhà bán lẻ lên kế hoạch nhân sự và khuyến mãi hiệu quả hơn vào những ngày cao điểm.

5.1.4 Ngày lễ và sự kiện đặc biệt (StateHoliday)

Doanh thu có xu hướng tăng rõ rệt trong các kỳ nghỉ lễ lớn (StateHoliday_b và StateHoliday_c).

Đây là giai đoạn nhu cầu mua sắm tăng cao, thường đi kèm với các hoạt động quảng bá và khuyến mãi đặc biệt.

Ngược lại, một số kỳ nghỉ ngắn (StateHoliday_a) có thể khiến doanh thu giảm nhẹ do người tiêu dùng dành thời gian nghỉ ngơi hoặc di chuyển.

5.1.5 Loại cửa hàng (StoreType)

Cửa hàng loại A thường có doanh thu cao nhất, có thể do vị trí tốt, quy mô lớn và tập trung ở khu vực có sức mua cao.

Các loại cửa hàng nhỏ hơn (như B hoặc C) có doanh thu thấp hơn, phù hợp với đặc thù khu dân cư hoặc vùng ven.

- Điều này chứng minh rằng yếu tố đặc trưng của cửa hàng (StoreType) là một biến quan trọng trong mô hình dự đoán doanh thu.

5.1.6 Mùa vụ (Tháng trong năm – Month)

Phân tích doanh thu theo tháng cho thấy sự tăng mạnh vào các tháng cuối năm (tháng 11–12) và giảm vào đầu năm (tháng 1–2).

Xu hướng này thể hiện tác động của yếu tố mùa vụ trong ngành bán lẻ, đặc biệt liên quan đến kỳ nghỉ lễ, sự kiện cuối năm và thói quen tiêu dùng của khách hàng.

5.1.7 Khoảng cách cạnh tranh (CompititonDistance)

Biến này có ảnh hưởng nhưng không mạnh, vì tác động của khoảng cách đối thủ còn phụ thuộc vào vị trí cụ thể và chiến lược giá.

Tuy nhiên, các cửa hàng có khoảng cách xa đối thủ thường duy trì doanh thu ổn định hơn, trong khi cửa hàng gần đối thủ dễ biến động doanh thu mạnh hơn theo thời gian.

5.2 Tiêu chí đánh giá mô hình

Hai chỉ số được sử dụng:

1. **RMSE – Root Mean Squared Error:**

   Đo sai số trung bình của dự đoán so với thực tế.

   → RMSE càng nhỏ càng tốt.

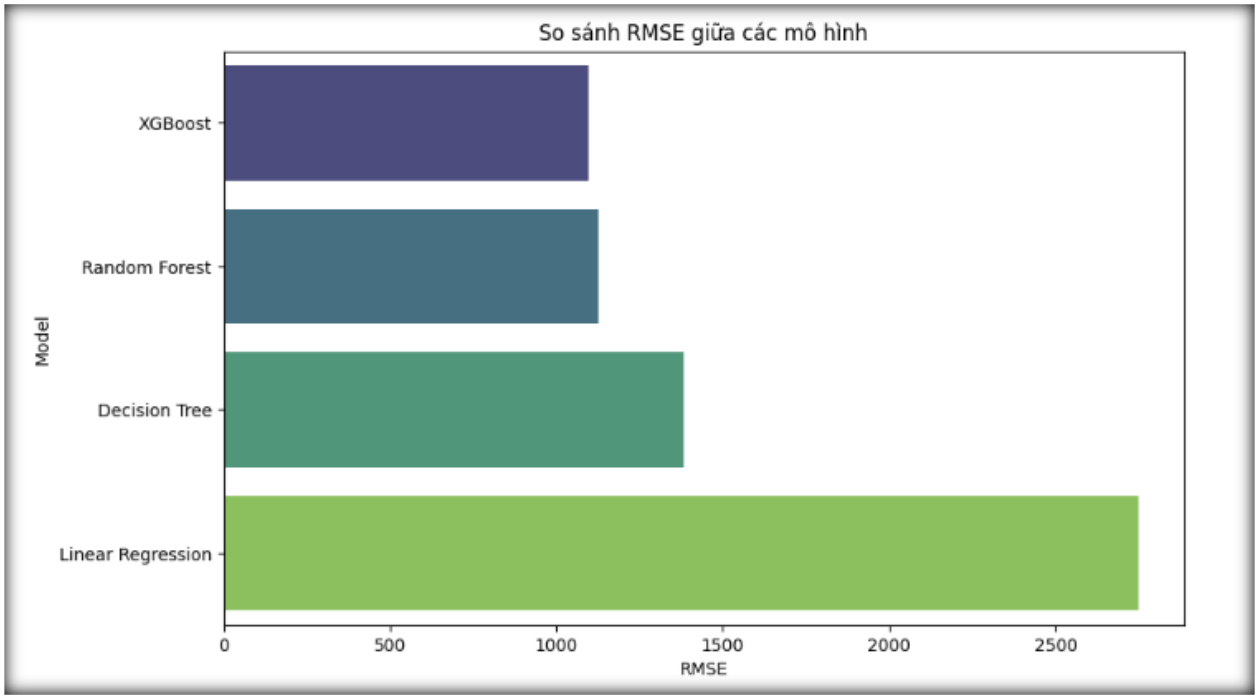
2. **R^2 – Hệ số xác định (Coefficient of Determination):**

   Phản ánh mức độ giải thích của mô hình đối với dữ liệu (từ 0 đến 1).
   → R^2 càng cao càng tốt.

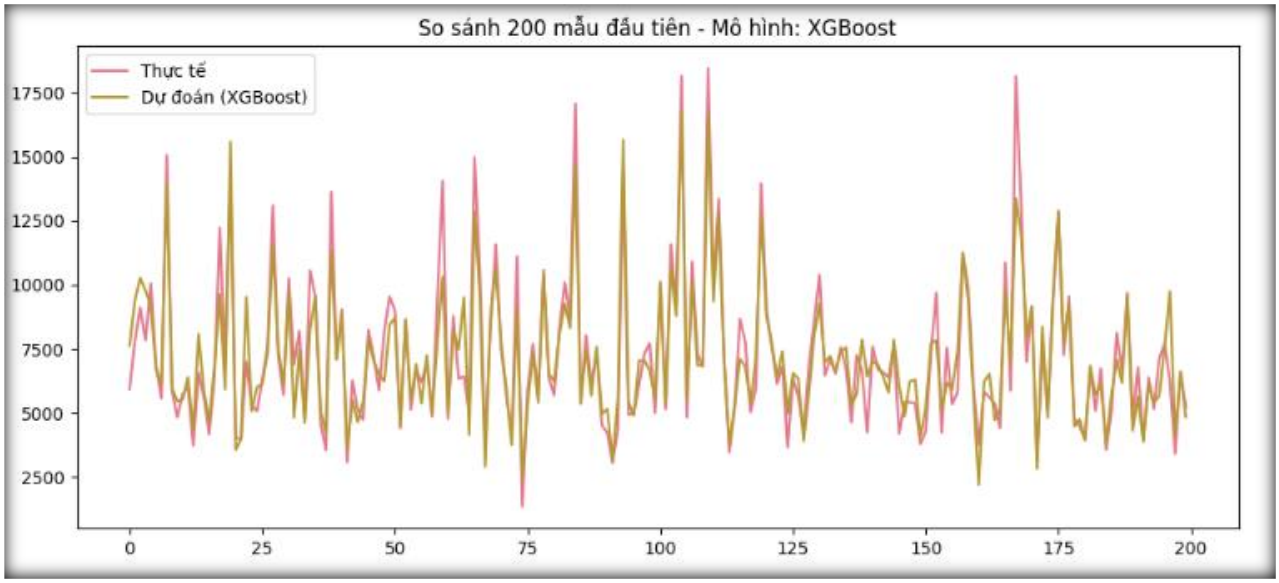
Mô hình	RMSE	R^2
XGBoost	1095.14	0.8776
Random Forest	1334.05	0.8686
Decision Tree	1381.45	0.8022
Linear Regression	2748.52	0.2171

Sau khi thực hiện toàn bộ quá trình tiền xử lý và huấn luyện, nhóm đưa ra kết luận sau:

- Trong các mô hình được thử nghiệm, XGBoost cho kết quả tốt nhất cả về độ chính xác và khả năng khái quát.
- Các đặc trưng ảnh hưởng mạnh nhất đến doanh thu bao gồm: Customers, Promo, DayOfWeek, Month, và StoreType.
- Kết quả mô hình có thể ứng dụng trong thực tế để dự báo doanh thu ngắn hạn và hỗ trợ quản lý tồn kho.



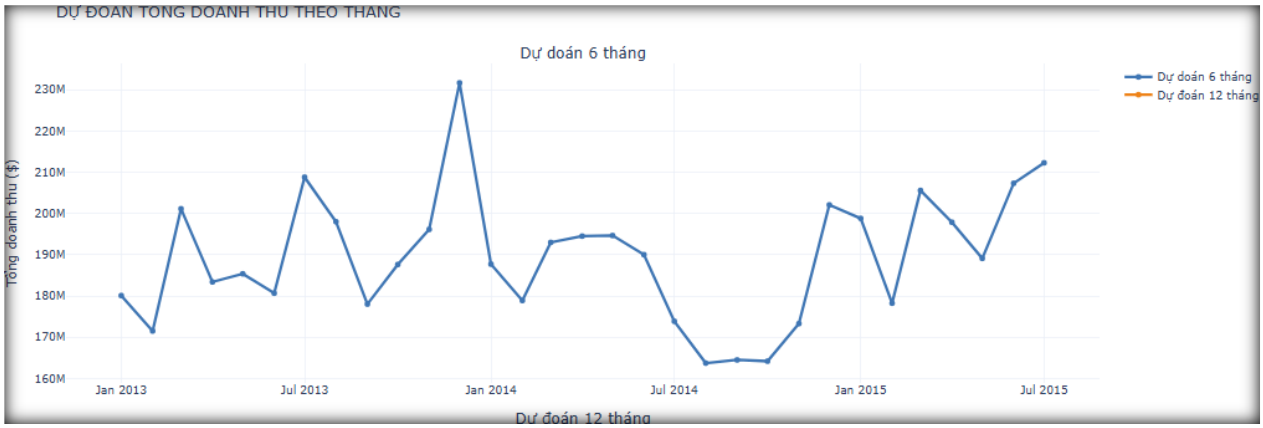
Hình 5.2.1 So sánh RMSE giữa các mô hình học máy



Hình 5.2.2 So sánh giá trị thực tế và dự đoán trên 200 mẫu đầu tiên

5.3 Dự đoán doanh số 6 tháng tiếp theo

Sau khi mô hình XGBoost được huấn luyện và lựa chọn là mô hình tối ưu, nhóm tiến hành dự đoán doanh số trong 6 tháng tiếp theo cho từng cửa hàng trong chuỗi Rossmann.



Hình 5.3.1 Biểu đồ dự đoán doanh thu 6 tháng

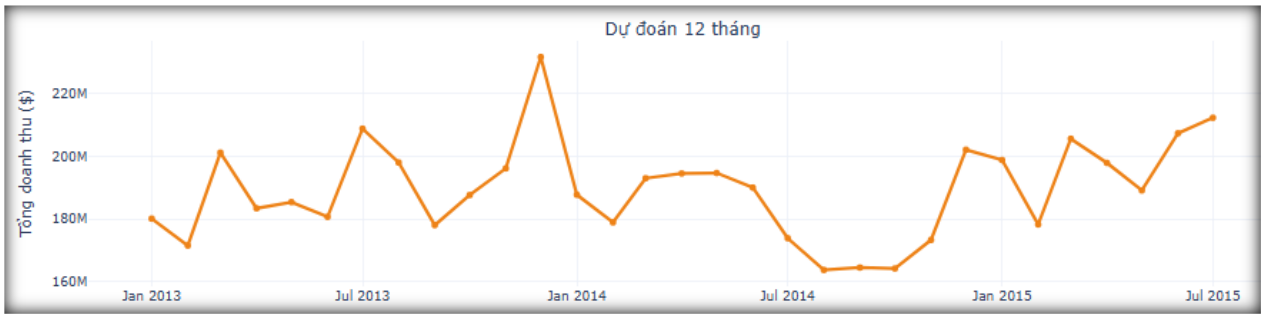
Đánh giá và nhận xét:

- Doanh thu của hệ thống có tính chu kỳ rõ rệt, tăng mạnh vào cuối năm và giảm nhẹ giữa năm.
- Các mô hình dự báo hoạt động ổn định, phản ánh được xu hướng thực tế của dữ liệu quá khứ.

- Kết quả dự báo cho thấy doanh nghiệp có triển vọng tăng trưởng tích cực trong nửa cuối năm 2015, đặc biệt nếu duy trì chiến dịch khuyến mãi và mở rộng danh mục sản phẩm phù hợp.
- Đây là cơ sở quan trọng để bộ phận quản lý lập kế hoạch nhập hàng, phân bổ nhân sự và ngân sách marketing theo mùa vụ.

5.4 Dự đoán doanh số 12 tháng tiếp theo

Sau khi dự báo 6 tháng, nhóm tiếp tục sử dụng mô hình XGBoost để dự đoán doanh số 12 tháng kể từ thời điểm kết thúc dữ liệu huấn luyện.



Hình 5.4.1 Biểu đồ dự đoán doanh thu 12 tháng

Đánh giá và nhận xét:

- Kết quả doanh thu tương đối so với dự đoán 6 tháng, cùng thể hiện xu hướng tăng trưởng tích cực về doanh thu
- Sự ổn định tương đối trong dự báo thể hiện rằng mô hình hoạt động đáng tin cậy trong dự báo dài hạn (12 tháng).

CHƯƠNG 6: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1 Kết luận

Qua quá trình nghiên cứu, phân tích và triển khai mô hình trên bộ dữ liệu bán lẻ của hệ thống cửa hàng Rossmann, nhóm đã hoàn thành các mục tiêu chính đặt ra trong quy trình khai phá dữ liệu (Data Mining), bao gồm:

1. **Phân tích khám phá dữ liệu (EDA):**

- Làm rõ đặc điểm của dữ liệu doanh thu theo thời gian, loại cửa hàng, ngày trong tuần, chương trình khuyến mãi và các ngày lễ đặc biệt.

- Xác định được các yếu tố ảnh hưởng trực tiếp đến doanh thu, bao gồm: Customers, Promo, DayOfWeek, StoreType, StateHoliday và Month.
- Phát hiện tính mùa vụ rõ rệt trong doanh thu — tăng mạnh vào cuối năm và giảm vào giai đoạn giữa năm.

2. Tiền xử lý và chuẩn bị dữ liệu:

- Thực hiện các bước làm sạch dữ liệu, xử lý giá trị thiếu (fillna), chuyển đổi định dạng thời gian, mã hóa biến phân loại (One-hot Encoding) và chuẩn hóa các biến số cần thiết.
- Gộp dữ liệu từ các bảng train và store giúp mô hình có thêm thông tin mô tả chi tiết từng cửa hàng, làm tăng chất lượng đầu vào cho mô hình học máy.

3. Xây dựng và đánh giá mô hình Machine Learning:

- Triển khai và so sánh bốn mô hình học máy phổ biến: Linear Regression, Decision Tree, Random Forest và XGBoost.
- Kết quả cho thấy XGBoost là mô hình cho hiệu năng tốt nhất, đạt RMSE ≈ 1095 và $R^2 \approx 0.88$, thể hiện khả năng dự đoán doanh thu với độ chính xác cao và ổn định.
- Mô hình đã thể hiện rõ khả năng nắm bắt xu hướng và quy luật mùa vụ của doanh thu trong hệ thống bán lẻ.

4. Ứng dụng mô hình trong thực tiễn:

- Dựa trên kết quả dự báo, đề tài đề xuất cơ chế tính toán tồn kho tối ưu, với mức dự trữ an toàn 10% so với doanh thu trung bình dự kiến.
- Kết quả giúp tối ưu quản lý hàng hóa, giảm chi phí tồn kho và tránh rủi ro thiếu hàng trong mùa cao điểm.
- Ngoài ra, mô hình dự báo 6 tháng và 12 tháng cho thấy xu hướng tăng trưởng ổn định và khả năng phục hồi doanh số theo chu kỳ, giúp doanh nghiệp lập kế hoạch bán hàng hiệu quả hơn.

➢ **Kết luận chung:**

Đề tài "Phân tích và dự báo doanh thu cửa hàng bằng phương pháp khai phá dữ liệu và mô hình học máy" đã hoàn thành tốt mục tiêu nghiên cứu đề ra, minh chứng cho tiềm năng ứng dụng mạnh mẽ của Data Mining trong lĩnh vực bán lẻ.

Kết quả không chỉ giúp doanh nghiệp ra quyết định dựa trên dữ liệu mà còn mở ra hướng phát triển bền vững trong quản trị thông minh.

6.2 Hướng phát triển

Trong tương lai, đề tài có thể được mở rộng và cải tiến theo các hướng sau:

1. Mở rộng nguồn dữ liệu:
    - Bổ sung thêm dữ liệu về giá bán, danh mục sản phẩm, chiến dịch marketing, yếu tố địa lý và thời tiết, nhằm nâng cao khả năng dự đoán và phân tích nguyên nhân.
    - Thu thập thêm dữ liệu từ nhiều năm và nhiều khu vực để mô hình hóa xu hướng tổng thể chính xác hơn.
2. Cải tiến mô hình dự báo:
    - Áp dụng các mô hình Deep Learning (LSTM, GRU) để nắm bắt tốt hơn các chuỗi thời gian phức tạp.
    - Tối ưu siêu tham số tự động bằng Grid Search hoặc Bayesian Optimization để tăng hiệu năng dự báo.
    - Kết hợp các mô hình (ensemble) như Stacking hoặc Hybrid Model nhằm tận dụng ưu điểm của từng thuật toán.
3. Phát triển hệ thống ứng dụng thực tế:
    - Xây dựng dashboard dự báo động (interactive dashboard) giúp nhà quản lý xem trực tiếp xu hướng doanh thu theo thời gian thực.
    - Tích hợp mô hình vào hệ thống quản trị bán hàng (POS) để tự động gợi ý mức tồn kho, lập kế hoạch nhập hàng hoặc đề xuất khuyến mãi.
4. Đánh giá rủi ro và mô phỏng kinh doanh:
    - Ứng dụng mô hình trong phân tích kịch bản (scenario analysis) để dự báo doanh thu trong các trường hợp khác nhau (thay đổi giá, khuyến mãi, số lượng khách hàng…).
    - Kết hợp với mô phỏng Monte Carlo để xác định phạm vi biến động doanh thu và tồn kho an toàn.

CHƯƠNG 7: TÀI LIỆU THAM THẢO

7.1 Tài liệu

1. Scikit-learn Documentation: https://scikit-learn.org/stable/documentation.html
2. Streamlit Documentation: https://docs.streamlit.io
3. Pandas Documentation: https://pandas.pydata.org/docs

4. Kaggle Retail Sales Dataset (tham khảo dữ liệu mô phỏng): https://www.kaggle.com/
5. Kaggle. (2015). *Rossmann Store Sales Dataset*. Retrieved from https://www.kaggle.com/competitions/rossmann-store-sales

7.2 Các thư viện sử dụng trong bài

1. Python 3 - ngôn ngữ lập trình chính.
2. Scikit-learn, XGBoost, RandomForest, DecisionTree - huấn luyện và đánh giá mô hình học máy.
3. Pandas, NumPy - xử lý và phân tích dữ liệu.
4. Math – tính toán dữ liệu.
5. Matplotlib, Plotly, Seaborn - trực quan hóa kết quả.
6. Microsoft Word (docx) - biên soạn và trình bày báo cáo.