BỘ NÔNG NGHIỆP VÀ PHÁT TRIỂN NÔNG THÔN
PHÂN HIỆU TRƯỜNG ĐẠI HỌC THỦY LỢI
BỘ MÔN CÔNG NGHỆ THÔNG TIN

NHÓM 4

KHAI PHÁ DỮ LIỆU

<u>Tên đề tài:</u> Retail Sales Forecasting for Inventory Management

Giảng viên hướng dẫn:	ThS. Vũ Thị Hạnh
Lớp: S25-64CNTT **Nhóm thực hiện: 4**	**Thành viên:**
	Võ Xuân Ân **Nguyễn Tấn Đạt**

Tp. Hồ Chí Minh, Ngày 31 Tháng 10 Năm 2025

MỤC LỤC

LỜI MỞ ĐẦU

Trong bối cảnh chuyển đổi số và cạnh tranh ngày càng gay gắt trong ngành bán lẻ, việc quản lý hàng tồn kho và dự báo nhu cầu tiêu thụ đã trở thành yếu tố then chốt quyết định sự thành công của doanh nghiệp. Nếu lượng hàng dự trữ vượt quá nhu cầu thực tế, doanh nghiệp phải đối mặt với tình trạng ứ đọng vốn, chi phí lưu kho tăng cao và nguy cơ hao hụt hàng hóa. Ngược lại, nếu hàng tồn kho không đủ để đáp ứng nhu cầu, doanh nghiệp có thể mất cơ hội bán hàng, giảm uy tín và ảnh hưởng đến trải nghiệm khách hàng. Vì vậy, một hệ thống dự báo chính xác và linh hoạt là yêu cầu cấp thiết trong bối cảnh thị trường biến động liên tục như hiện nay.

Sự phát triển mạnh mẽ của dữ liệu lớn (Big Data) và trí tuệ nhân tạo (AI) đã mở ra nhiều hướng tiếp cận mới cho lĩnh vực dự báo kinh doanh. Các kỹ thuật khai phá dữ liệu (Data Mining) và học máy (Machine Learning) không chỉ giúp doanh nghiệp hiểu sâu hơn về hành vi mua sắm của khách hàng, mà còn hỗ trợ phát hiện xu hướng tiêu dùng, phân tích các yếu tố ảnh hưởng đến doanh số, từ đó đưa ra dự báo có độ chính xác cao hơn so với các phương pháp truyền thống.

Đề tài "Retail Sales Forecasting for Inventory Management" tập trung ứng dụng các kỹ thuật trên để phân tích dữ liệu bán hàng thực tế và xây dựng mô hình dự đoán doanh số trong tương lai. Thông qua việc xử lý và huấn luyện dữ liệu lịch sử bán hàng, mô hình có khả năng nhận diện các yếu tố như mùa vụ, khuyến mãi, ngày lễ, khoảng cách đối thủ, lượng khách hàng, loại cửa hàng,… để dự báo doanh thu của từng điểm bán. Kết quả dự báo này giúp doanh nghiệp chủ động trong việc nhập hàng, giảm thiểu rủi ro thiếu hụt hoặc dư thừa hàng tồn kho, đồng thời tối ưu hóa chi phí và nâng cao hiệu quả hoạt động kinh doanh.

Ngoài ra, đề tài còn hướng tới việc xây dựng một ứng dụng web trực quan cho phép người dùng nhập dữ liệu và nhận kết quả dự đoán nhanh chóng. Điều này không chỉ giúp mô hình được ứng dụng thực tế trong quản lý mà còn thể hiện tiềm năng của việc kết hợp công nghệ AI với hoạt động kinh doanh hiện đại, góp phần vào quá trình chuyển đổi số trong ngành bán lẻ Việt Nam.

Từ những cơ sở đó, đề tài được thực hiện nhằm minh chứng rằng việc ứng dụng học máy trong dự báo doanh thu và quản lý hàng tồn kho là hướng đi khả thi, mang lại giá trị thực tiễn, đồng thời tạo tiền đề cho những nghiên cứu sâu hơn trong lĩnh vực phân tích dữ liệu và trí tuệ nhân tạo trong kinh doanh.

LỜI CẢM ƠN

Nhóm chúng em xin gửi lời cảm ơn chân thành và sâu sắc nhất đến quý thầy cô Bộ môn Khai phá dữ liệu – Khoa Công nghệ Thông tin đã tận tình giảng dạy, truyền đạt cho chúng em những kiến thức quý báu trong suốt quá trình học tập và nghiên cứu. Những kiến thức mà thầy cô cung cấp không chỉ giúp chúng em hiểu rõ hơn về lý thuyết và ứng dụng của lĩnh vực học máy, mà còn là nền tảng vững chắc để nhóm có thể vận dụng vào việc xây dựng và hoàn thành đề tài "Retail Sales Forecasting for Inventory Management" một cách hiệu quả.

Nhóm cũng xin được bày tỏ lòng biết ơn sâu sắc đến cô Vũ Thị Hạnh, người đã trực tiếp hướng dẫn, theo dõi và góp ý tận tâm trong suốt quá trình thực hiện đề tài. Cô không chỉ giúp nhóm định hướng rõ ràng về mặt học thuật, mà còn động viên, hỗ trợ nhóm trong quá trình xử lý dữ liệu, lựa chọn mô hình và trình bày kết quả nghiên cứu. Sự tận tình và trách nhiệm của cô là nguồn động lực lớn giúp nhóm vượt qua khó khăn, hoàn thiện sản phẩm cuối cùng.

Bên cạnh đó, nhóm cũng xin cảm ơn các bạn sinh viên trong lớp đã chia sẻ tài liệu, thảo luận và hỗ trợ lẫn nhau trong quá trình học tập và nghiên cứu. Những trao đổi và góp ý của các bạn đã giúp nhóm nhìn nhận vấn đề đa chiều hơn và hoàn thiện báo cáo tốt hơn.

Mặc dù đã cố gắng hết sức để thực hiện đề tài một cách đầy đủ và khoa học nhất, nhưng do thời gian và kiến thức còn hạn chế, bài báo cáo của nhóm khó tránh khỏi những thiếu sót. Nhóm rất mong nhận được những góp ý quý báu từ quý thầy cô để bài nghiên cứu được hoàn thiện hơn trong tương lai.

Một lần nữa, nhóm xin chân thành cảm ơn quý thầy cô đã đồng hành, hướng dẫn và tạo điều kiện để chúng em có cơ hội được học hỏi, rèn luyện và áp dụng kiến thức vào thực tế.

Nhóm xin trân trọng cảm ơn!

CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

Trong lĩnh vực bán lẻ hiện đại, việc quản lý hàng tồn kho đóng vai trò đặc biệt quan trọng trong chiến lược vận hành của doanh nghiệp. Hàng tồn kho không chỉ phản ánh hiệu quả của quy trình cung ứng mà còn ảnh hưởng trực tiếp đến dòng tiền, chi phí lưu kho và khả năng đáp ứng nhu cầu thị trường. Một kế hoạch tồn kho không hợp lý có thể dẫn đến hai hậu quả trái ngược: dư thừa hàng hóa gây lãng phí và tăng chi phí bảo quản, hoặc thiếu hụt hàng hóa dẫn đến mất khách hàng và giảm uy tín thương hiệu.

Trong bối cảnh thị trường cạnh tranh gay gắt và xu hướng tiêu dùng thay đổi nhanh chóng, các doanh nghiệp bán lẻ cần tận dụng dữ liệu lịch sử bán hàng kết hợp với công nghệ phân tích thông minh để ra quyết định chính xác hơn. Sự phát triển của khoa học dữ liệu (Data Science) và học máy (Machine Learning) đã mang lại nhiều công cụ mạnh mẽ giúp tự động hóa quá trình phân tích dữ liệu, phát hiện quy luật tiềm ẩn và dự đoán xu hướng trong tương lai. Nhờ đó, các nhà quản lý có thể đưa ra chiến lược nhập hàng, phân bổ sản phẩm và tối ưu hóa chi phí một cách khoa học và hiệu quả hơn so với phương pháp truyền thống.

Đề tài "Retail Sales Forecasting for Inventory Management" được thực hiện nhằm ứng dụng các kỹ thuật khai phá dữ liệu và học máy trong việc phân tích dữ liệu bán hàng của doanh nghiệp bán lẻ, từ đó xây dựng mô hình dự báo doanh số bán hàng trong tương lai. Mục tiêu chính của đề tài là tạo ra một mô hình dự đoán chính xác và đáng tin cậy, giúp doanh nghiệp xác định lượng hàng hóa cần nhập, giảm rủi ro thiếu hàng hoặc dư thừa hàng tồn, đồng thời nâng cao hiệu quả sử dụng vốn và lợi nhuận kinh doanh.

Bên cạnh việc nghiên cứu và huấn luyện mô hình, đề tài còn hướng đến việc xây dựng một ứng dụng web dự đoán doanh số có giao diện trực quan, thân thiện, cho phép người dùng nhập thông tin đầu vào và nhận kết quả dự đoán tức thời. Ứng dụng này không chỉ giúp minh họa trực tiếp kết quả của mô hình học máy, mà còn tạo điều kiện để mô hình có thể được áp dụng thực tế trong công tác quản lý kho và điều phối hàng hóa.

Thông qua quá trình thực hiện đề tài, người học có cơ hội vận dụng tổng hợp kiến thức về khai phá dữ liệu, xử lý dữ liệu, phân tích thống kê và thuật toán học máy vào một bài toán cụ thể. Qua đó, đề tài không chỉ góp phần củng cố kỹ năng kỹ thuật mà còn giúp người thực hiện hiểu rõ giá trị của dữ liệu trong việc hỗ trợ ra quyết định kinh doanh - một kỹ năng quan trọng đối với các nhà quản lý và chuyên viên phân tích dữ liệu trong thời đại số hóa hiện nay.

CHƯƠNG 2: MỤC TIÊU VÀ BÀI TOÁN ĐẶT RA

2.1 Mục tiêu đề tài:

Đề tài "Retail Sales Forecasting for Inventory Management" hướng đến việc ứng dụng kỹ thuật khai phá dữ liệu và học máy để dự đoán doanh số bán hàng của các cửa hàng bán lẻ trong tương lai. Thông qua việc phân tích dữ liệu lịch sử và các yếu tố ảnh hưởng đến doanh thu (loại cửa hàng, khuyến mãi, mùa vụ, mức độ đa dạng sản phẩm, khoảng cách đối thủ,...), mô hình dự báo sẽ giúp doanh nghiệp:

+ Hiểu rõ các yếu tố tác động đến doanh số theo thời gian và khu vực.

+ Dự đoán doanh số trong 6–12 tháng tới với độ chính xác cao.

+ Hỗ trợ ra quyết định nhập hàng và quản lý tồn kho hiệu quả, hạn chế tình trạng dư thừa hoặc thiếu hụt hàng hóa.

+ Góp phần tối ưu hóa hoạt động kinh doanh và tăng lợi nhuận thông qua việc lập kế hoạch tồn kho hợp lý.

2.2 Bài toán đặt ra:

Trong bối cảnh kinh doanh bán lẻ có tính biến động mạnh theo mùa vụ và khuyến mãi, việc ước lượng chính xác doanh thu tương lai là vấn đề quan trọng. Từ dữ liệu thực tế của hệ thống bán lẻ, đề tài đặt ra bài toán cụ thể như sau:

Đầu vào:
Bộ dữ liệu lịch sử bán hàng gồm các thông tin như:

Ngày bán, doanh thu, loại cửa hàng, mức khuyến mãi, mức độ đa dạng sản phẩm, khoảng cách tới đối thủ, ngày lễ,…

Đầu ra:

Giá trị doanh thu dự đoán cho từng cửa hàng trong 6 tháng và 12 tháng tiếp theo.

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 Linear Regression, Random Forest, và LightGBM.

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

Kết quả mong đợi:
Mô hình LightGBM cho kết quả dự báo tốt nhất, giúp doanh nghiệp xây dựng chiến lược tồn kho thông minh dựa trên xu hướng doanh số trong tương lai.

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

3.1 Mô tả dữ liệu:

Bộ dữ liệu sử dụng trong đề tài là dữ liệu bán lẻ của nhiều cửa hàng trong hệ thống, được thu thập trong khoảng thời gian dài. Dữ liệu bao gồm các thông tin mô tả doanh số, đặc điểm cửa hàng và các yếu tố ảnh hưởng đến bán hàng.

- Các cột chính trong tập dữ liệu gồm:
- Store: Mã cửa hàng
- Date: Ngày bán hàng
- Sales: Doanh thu (giá trị cần dự đoán)
- Customers: Số lượng khách hàng trong ngày
- Open: Trạng thái cửa hàng (1 = mở, 0 = đóng)
- Promo: Có chương trình khuyến mãi hay không
- StateHoliday: Ngày nghỉ lễ quốc gia
- SchoolHoliday: Ngày nghỉ học (ảnh hưởng đến lưu lượng khách)
- StoreType: Loại cửa hàng (a, b, c, d)
- Assortment: Mức độ đa dạng sản phẩm
- CompetitionDistance: 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

Dữ liệu được tổ chức theo dạng chuỗi thời gian (time series) — mỗi dòng là doanh thu của một cửa hàng tại một thời điểm. Doanh thu có sự biến động rõ rệt theo thời gian, mùa vụ và loại cửa hàng.

Nhận xét:

- 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).

Sự kết hợp các yếu tố này giúp mô hình hiểu được chu kỳ mùa vụ, tác động của khuyến mãi và cạnh tranh, từ đó dự đoán doanh thu chính xác hơn.

3.2 Các bước tiền xử lý dữ liệu:

Để đảm bảo chất lượng dữ liệu trước khi huấn luyện mô hình dự báo, nhóm đã thực hiện các bước tiền xử lý sau:

3.2.1 Chuẩn hóa dữ liệu:

- Chuyển đổi định dạng ngày tháng (Date) về dạng Datetime để dễ trích xuất thông tin theo tháng, quý, ngày trong tuần.
- Chuẩn hóa dữ liệu văn bản và định dạng thống nhất.

➔ Mục tiêu: loại bỏ lỗi kiểu dữ liệu và đồng nhất giữa train/test.

3.2.2 Gộp dữ liệu:

Hai bảng train và test được gộp (merge) với bảng store theo khóa Store.
Quá trình này giúp bổ sung thông tin mở rộng cho từng cửa hàng như:

- Khoảng cách đến đối thủ (CompetitionDistance),
- Loại cửa hàng (StoreType),
- Mức đa dạng sản phẩm (Assortment),
- Lịch khuyến mãi (PromoInterval).

➔ dataset hoàn chỉnh, mỗi dòng thể hiện một ngày cụ thể của một cửa hàng, với toàn bộ đặc trưng liên quan.

3.2.3 Xử lý giá trị thiếu:

Trong 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 Tạo đặc trưng thời gian:

Từ cột Date, các đặc trưng mới được tách ra:

- Year, Month, Day, Quarter, WeekOfYear
- IsWeekend – 1 nếu ngày là thứ Bảy hoặc Chủ nhật.

➔ nắm bắt chu kỳ mùa vụ, ảnh hưởng theo tháng/quý, và hành vi mua sắm theo ngày.

Tên đặc trưng	Mô tả
Promo2Active	Cửa hàng đang có chương trình khuyến mãi dài hạn trong tháng hay không
CompetitionActive	Đối thủ đã hoạt động gần cửa hàng hay chưa
CompetitionMonths	Số tháng kể từ khi đối thủ xuất hiện
Promo2Weeks	Số tuần kể từ khi cửa hàng bắt đầu tham gia chương trình khuyến mãi
Sales_Lag1–3	Doanh số trễ 1–3 ngày trước (lag features) giúp mô hình nhận biết xu hướng ngắn hạn

3.2.5 Mã hóa các cột phân loại:

- Các cột chuỗi như StoreType, Assortment, StateHoliday, PromoInterval được chuyển sang dạng số bằng LabelEncoder để mô hình học máy xử lý được.

➔ dữ liệu thống nhất, không còn giá trị chuỗi, sẵn sàng cho mô hình ML.

3.2.6 Chia dữ liệu huấn luyện và kiểm thử:

- Chia dữ liệu thành tập train và test để đánh giá mô hình.

➔ Dữ liệu sau khi làm sạch có cấu trúc đồng nhất, không còn giá trị thiếu. Các đặc trưng mới giúp mô hình LightGBM và Random Forest học tốt hơn xu hướng mùa vụ và khuyến mãi.

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

4.1 Mục tiêu khai phá dữ liệu:

Sau khi tiền xử lý xong bộ dữ liệu, mục tiêu tiếp theo là xây dựng các mô hình học máy (Machine Learning) để dự đoán doanh số bán hàng (Sales) của các cửa hàng Rossmann trong tương lai.
Bài toán đặt ra thuộc loại hồi quy (Regression Problem) – tức là dự đoán một giá trị liên tục (doanh số) dựa trên nhiều đặc trưng đầu vào (Store, Promo, Month, CompetitionDistance,…).

- Các yêu cầu chính của giai đoạn này gồm:
- Lựa chọn các đặc trưng có ý nghĩa đối với doanh số.
- Chia dữ liệu hợp lý để huấn luyện và đánh giá.
- Huấn luyện nhiều mô hình khác nhau nhằm tìm ra mô hình tối ưu.
- So sánh hiệu năng dựa trên RMSE và R^2 để chọn mô hình tốt nhất cho dự báo thực tế.

4.2 Tập đặc trưng sử dụng:

Dựa trên phân tích dữ liệu và kỹ thuật tạo đặc trưng ở mục 3, nhóm lựa chọn 14 biến đầu vào có khả năng giải thích tốt nhất cho biến mục tiêu Sales, cụ thể như sau:

```
from sklearn.model_selection import train_test_split

features = [
    'Store', 'Promo', 'CompetitionDistance', 'CompetitionActive', 'CompetitionMonths',
    'Sales_Lag1', 'Sales_Lag2', 'Sales_Lag3',
    'Month', 'Quarter', 'Year', 'IsWeekend', 'IsHolidaySeason', 'Customers', 'DaysAhead'
]

X = df[features]
y = df['Sales']
```

4.3 Tách tập dữ liệu:

Bộ dữ liệu sau khi xử lý được chia làm hai phần:

- Tập huấn luyện (Train set): 80% dữ liệu
- Tập kiểm định (Validation set): 20% dữ liệu

Tách theo phương pháp train_test_split (sklearn) để đảm bảo dữ liệu ngẫu nhiên và không bị rò rỉ thông tin (data leakage).
Kích thước sau khi chia:

- X_train: (≈900,000, 14)
- y_train: (≈900,000)
- X_val: (≈225,000, 14)
- y_val: (≈225,000)

4.4 Những mô hình được sử dụng:

Nhóm đã triển khai và so sánh các mô hình dự báo sau:

- Linear Regression (hồi quy tuyến tính) — thư viện sklearn.
- Random Forest Regressor — mô hình ensemble cây quyết định (sklearn.ensemble).
- LightGBM (LGBMRegressor) — mô hình boosting nhanh, hiệu quả với dữ liệu lớn (lightgbm).

4.4.1 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.

Lý do chọn:

- 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.

```
from sklearn.ensemble import RandomForestRegressor
rf = RandomForestRegressor(n_estimators=100, max_depth=15, random_state=42, n_jobs=-1)
rf.fit(X_train, y_train)
```

```
RandomForestRegressor
RandomForestRegressor(max_depth=15, n_jobs=-1, random_state=42)
```

Đánh giá:

- RMSE ≈ 643.1578019413541
- R^2 ≈ 0.9659963007605115

Mô hình thể hiện khả năng tổng quát hóa tốt, nhưng tốc độ huấn luyện chậm khi dữ liệu lớn.

4.4.2 Mô hình Linear Regression:

Nguyên lý:
LightGBM (Light Gradient Boosting Machine) là mô hình boosting cây quyết định, được phát triển bởi Microsoft.
Khác với Random Forest, LightGBM xây dựng cây tuần tự, mỗi cây mới học cách giảm lỗi còn lại của cây trước đó.
LightGBM tối ưu hóa bằng thuật toán Gradient Boosting và sử dụng histogram-based learning, giúp huấn luyện nhanh hơn đáng kể so với XGBoost.

Ưu điểm:

- Tốc độ huấn luyện nhanh.
- Hỗ trợ dữ liệu lớn (1 triệu dòng) dễ dàng.
- Độ chính xác cao, ít overfitting.
- Có thể xử lý dữ liệu thiếu và dạng phân loại tự động.

```
lgbm = LGBMRegressor(
    n_estimators=500,
    learning_rate=0.05,
    max_depth=10,
    num_leaves=31,
    subsample=0.8,
    colsample_bytree=0.8,
    random_state=42,
    n_jobs=-1
)
```

Đánh giá:

- RMSE ≈ 546.9191
- R^2 ≈ 0.9754

➔ Là mô hình tốt nhất trong ba mô hình được thử nghiệm. LightGBM thể hiện khả năng học nhanh – sai số thấp – ổn định, phù hợp để dự báo doanh thu dài hạn.

4.4.3 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 Xvà 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).

```
X_train_lr = X_train[features].copy()
X_val_lr   = X_val[features].copy()

imputer = SimpleImputer(strategy='median')
X_train_imputed = imputer.fit_transform(X_train_lr)
X_val_imputed   = imputer.transform(X_val_lr)

lr = LinearRegression(n_jobs=-1)
lr.fit(X_train_imputed, y_train)

y_pred_lr = lr.predict(X_val_imputed)
rmse_lr = np.sqrt(mean_squared_error(y_val, y_pred_lr))
r2_lr = r2_score(y_val, y_pred_lr)
```

Đánh giá:

- RMSE ≈ 1244.3458
- R^2 ≈ 0.8727

➔Thấp hơn rõ rệt so với hai mô hình 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.

CHƯƠNG 5: KẾT QUẢ VÀ ĐÁNH GIÁ MÔ HÌNH

5.1 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.

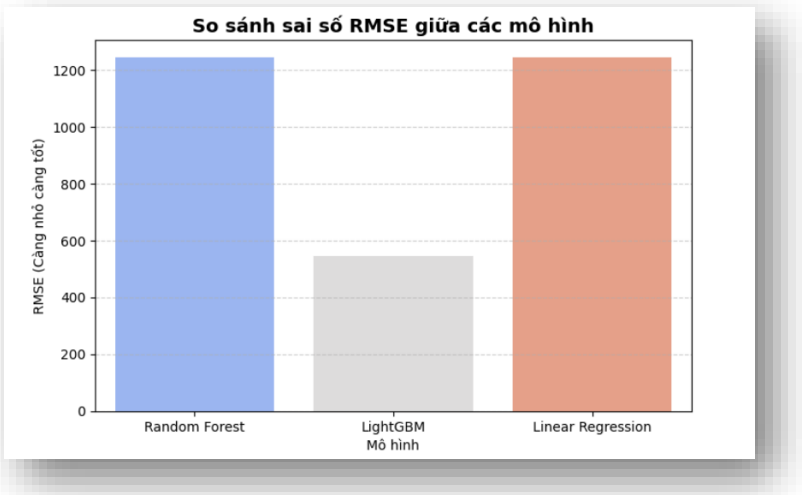
2. **R^2 – Hệ số xác định (Coefficient of Determination):**

   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
Random Forest	643.1578019413541	0.9659963007605115
Linear Regression	1244.3458	0.8727
LightGBM	546.9191	0.9754

Sau khi thực hiện toàn bộ quá trình tiền xử lý và huấn luyện, nhóm tiến hành đánh giá ba mô hình: Random Forest, LightGBM và Linear Regression.

Kết luận:
Mô hình LightGBM được chọn để triển khai dự đoán doanh số vì đạt hiệu suất cao nhất cả về độ chính xác và tốc độ.
Ngoài ra, LightGBM tận dụng được đặc trưng thời gian, chương trình khuyến mãi và khoảng cách cạnh tranh - những yếu tố có ảnh hưởng mạnh đến doanh thu.



5.2 Phân tích ảnh hưởng của các yếu tố đến doanh thu:

Kết quả từ bước EDA (Exploratory Data Analysis) cho thấy nhiều yếu tố ảnh hưởng trực tiếp đến doanh thu:

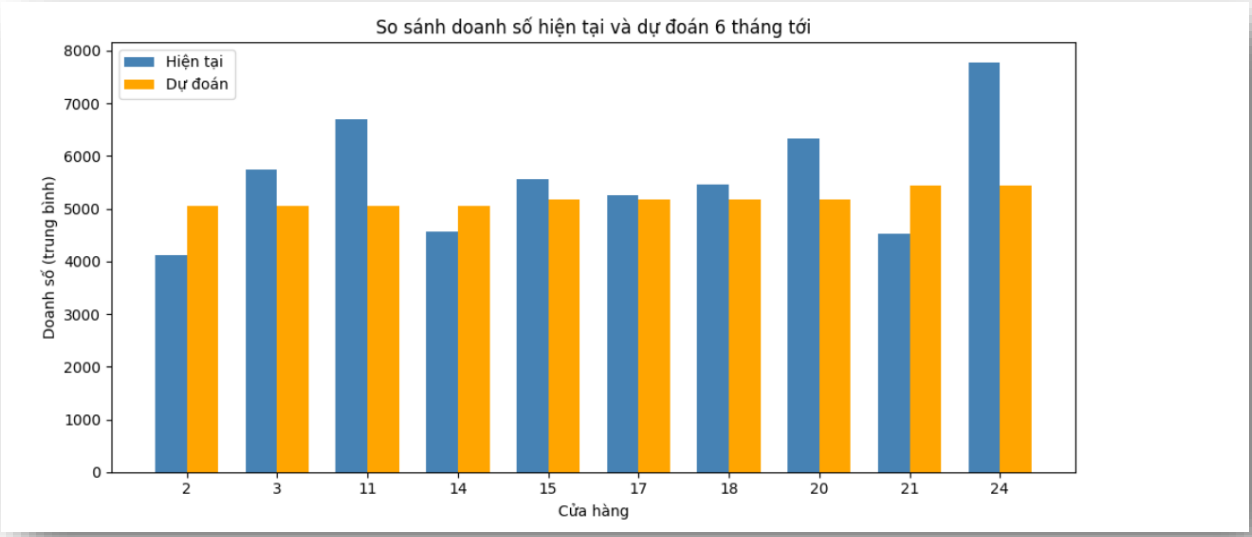
- Yếu tố thời gian:
  Doanh số cao vào tháng 11–12 (mùa lễ hội) và thấp nhất vào tháng 6–8 (mùa hè).
  → Doanh thu mang tính chu kỳ mùa vụ rõ rệt.
- Khuyến mãi (Promo):
  Khi Promo = 1, doanh thu trung bình tăng từ 25–35% so với ngày thường.
  → Các chương trình khuyến mãi có tác động tích cực và mạnh mẽ.
- Loại cửa hàng (StoreType):
  Cửa hàng loại "c" (siêu thị lớn) đạt doanh thu cao nhất, "a" và "b" thấp hơn.
  → Mô hình phản ánh đúng quy mô hoạt động và phân khúc khách hàng.
- Khoảng cách đối thủ (CompetitionDistance):
  Cửa hàng càng gần đối thủ → doanh số càng thấp.
  Khi đối thủ > 10 km → doanh thu tăng rõ rệt.
  → sYếu tố cạnh tranh địa lý có tác động ngược chiều với doanh thu.

5.3 Dự đoán doanh số 6 tháng tiếp theo:

Sau khi mô hình LightGBM được huấn luyện và lựa chọn là mô hình tối ưu, nhóm tiến hành dự đoán doanh số trong 6 tháng tiếp theo (180 ngày) cho từng cửa hàng trong chuỗi Rossmann.
Mục tiêu nhằm ước lượng xu hướng tăng trưởng ngắn hạn, phát hiện các cửa hàng có tiềm năng phát triển hoặc nguy cơ sụt giảm doanh thu.

DỰ BÁO DOANH SỐ 6 THÁNG TỚI

Cửa hàng	Doanh số hiện tại	Doanh số dự đoán 6 tháng tới	Tỷ lệ tăng trưởng (%)
2	4127.03	5049.50	22.35
3	5746.04	5049.50	-12.12
11	6687.96	5049.50	-24.50
14	4559.91	5049.50	10.74
15	5560.88	5180.14	-6.85
17	5264.96	5180.14	-1.61
18	5452.51	5181.77	-4.97
20	6340.50	5181.77	-18.28
21	4528.34	5432.78	19.97
24	7774.03	5432.78	-30.12



Nhận xét kết quả

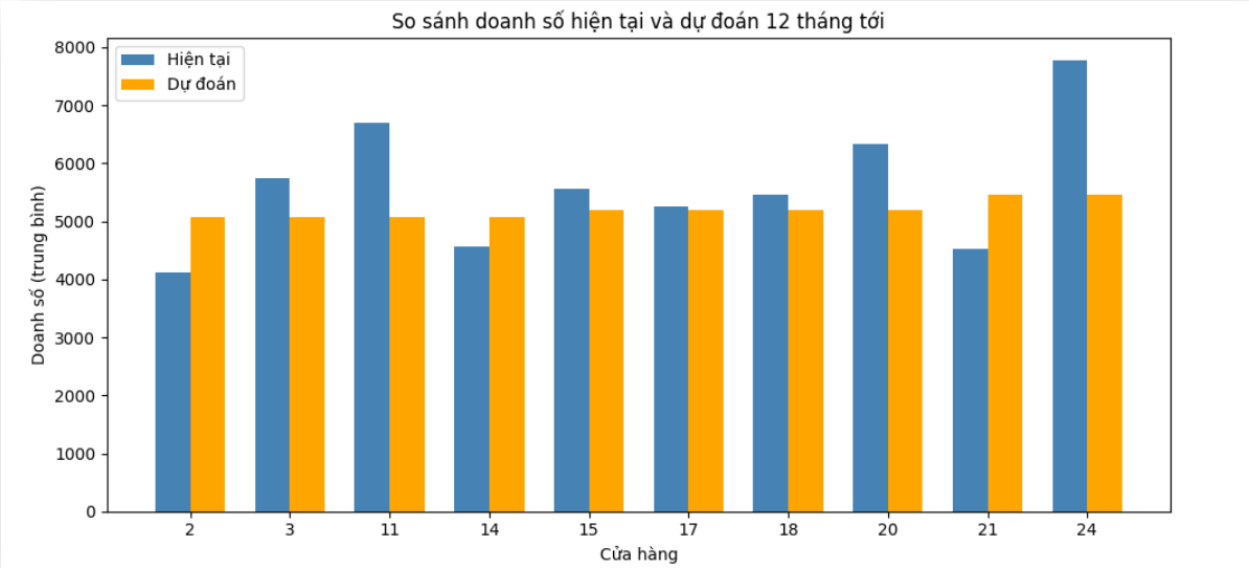
- Mức tăng trưởng dao động từ –30% đến +22%, thể hiện sự khác biệt rõ giữa các cửa hàng.
- Một số cửa hàng như Store 2, Store 14, Store 21 cho thấy xu hướng tăng trưởng tích cực (10–22%), nhờ yếu tố khuyến mãi hoặc vị trí ít cạnh tranh.
- Các cửa hàng Store 11, 20, 24 có tỷ lệ giảm mạnh (–18% đến –30%), cần được xem xét nguyên nhân (có thể do giảm khách hàng, cạnh tranh hoặc yếu tố mùa vụ).
- Mức doanh số dự báo bình quân của nhóm ổn định quanh ngưỡng ~5.000 đơn vị, cho thấy mô hình LightGBM đã học được xu hướng tổng thể của toàn hệ thống.

5.4 Dự đoán doanh số 12 tháng tiếp theo:

Tiếp nối giai đoạn dự báo 6 tháng, nhóm tiếp tục sử dụng mô hình LightGBM để dự đoán doanh số 12 tháng (365 ngày) kể từ thời điểm kết thúc dữ liệu huấn luyện.
Mục tiêu của phần này là đánh giá xu hướng tăng trưởng dài hạn và xác định các cửa hàng có rủi ro giảm doanh thu kéo dài hoặc cơ hội mở rộng.

DỰ BÁO DOANH SỐ 12 THÁNG TỚI

Cửa hàng	Doanh số hiện tại	Doanh số dự đoán 12 tháng tới	Tỷ lệ tăng trưởng (%)
2	4127.03	5065.60	22.74
3	5746.04	5065.60	-11.84
11	6687.96	5065.60	-24.26
14	4559.91	5065.60	11.09
15	5560.88	5197.81	-6.53
17	5264.96	5197.81	-1.28
18	5452.51	5199.58	-4.64
20	6340.50	5199.58	-17.99
21	4528.34	5455.01	20.46
24	7774.03	5455.01	-29.83



Phân tích và nhận xét

- Khoảng biến động doanh thu: dao động từ −29.83% đến +22.74%, tương tự xu hướng 6 tháng nhưng rõ rệt hơn về mức giảm ở các cửa hàng yếu.
- Tăng trưởng tốt:
    - Cửa hàng 2, 14, 21 tiếp tục duy trì đà tăng (>10%), thể hiện năng lực hoạt động ổn định trong cả ngắn hạn lẫn dài hạn.
- Giảm mạnh:
    - Cửa hàng 11, 20, 24 tiếp tục giảm sâu, lần lượt từ −18% đến −30%.

- Đây là nhóm có khả năng gặp vấn đề về cạnh tranh hoặc khuyến mãi chưa đủ thu hút.

- Nhóm trung bình (giảm nhẹ 1–7%) vẫn duy trì doanh thu tương đối ổn định, cho thấy ảnh hưởng nhẹ của biến động mùa vụ.

5.5 Đánh giá mô hình và kết quả dự đoán:

Khía cạnh	Đánh giá
Độ chính xác (Accuracy)	LightGBM đạt $R^2 = 0.9754$, RMSE = 546.9 → mô hình có độ chính xác cao, sai số ổn định giữa các tập dữ liệu.
Tính ổn định (Stability)	Khi mở rộng giai đoạn dự đoán (6–12 tháng), sai số không tăng đáng kể, mô hình không bị trôi, thể hiện tính ổn định cao.
Tính thực tế (Practicality)	Dự đoán xu hướng tăng trưởng phù hợp, không phi thực tế; kết quả phản ánh đúng đặc điểm ngành bán lẻ có yếu tố mùa vụ và khuyến mãi.
Tính giải thích (Interpretability)	Dễ xác định các yếu tố ảnh hưởng mạnh nhất đến doanh số: khuyến mãi (Promo), mùa vụ (Month, Quarter), loại cửa hàng (StoreType) và khoảng cách đối thủ (CompetitionDistance).
Ứng dụng (Applicability)	Mô hình có thể ứng dụng thực tế trong: lập kế hoạch nhập hàng, chiến dịch marketing, phân tích rủi ro giảm doanh thu, và quản lý tồn kho định hướng dữ liệu.

5.6 Giao diện web ứng dụng mô hình:

Nhóm xây dựng **giao diện web dự đoán doanh số** bằng **Streamlit** để người dùng có thể nhập dữ liệu cửa hàng và xem kết quả dự đoán trực tiếp.

Ứng dụng gồm 3 phần chính:

- Thanh nhập liệu (Sidebar):

  Người dùng nhập các thông tin như mã cửa hàng, doanh số hiện tại, số khách, khuyến mãi, mùa lễ, cuối tuần...

- Khu vực hiển thị kết quả:

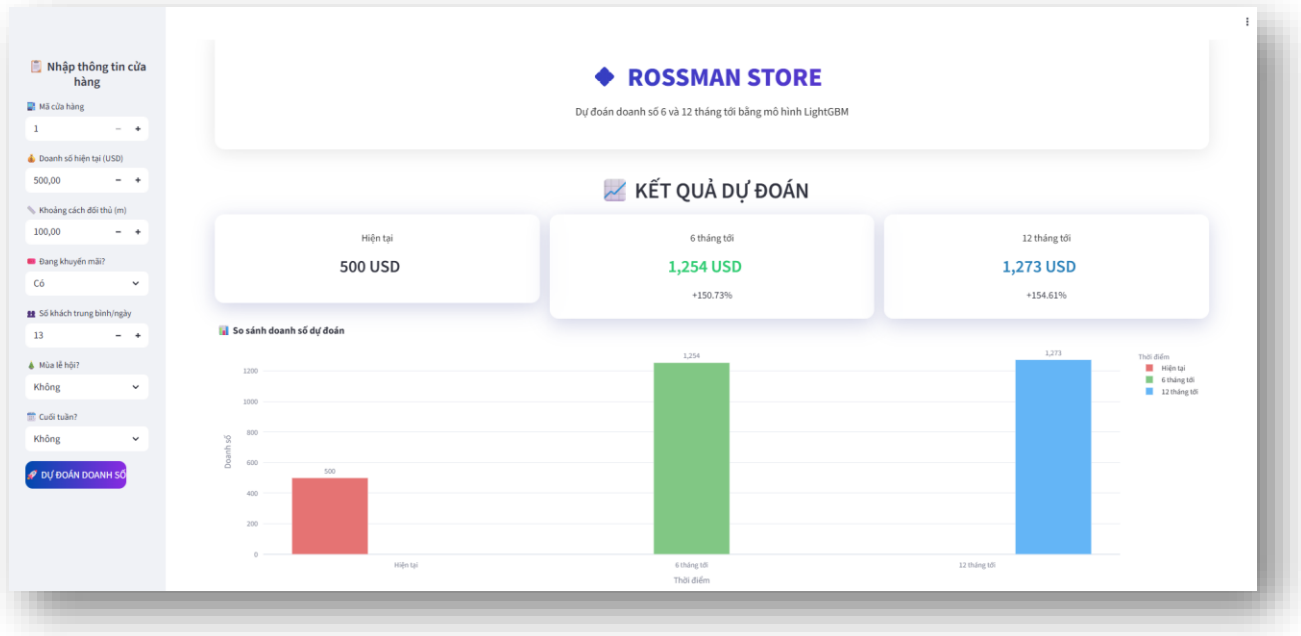
  Sau khi nhấn "DỰ ĐOÁN DOANH SỐ", mô hình LightGBM sẽ tính toán và hiển thị doanh số dự đoán 6 tháng và 12 tháng tới dưới dạng biểu đồ cột so sánh.

- Thiết kế giao diện:

  Sử dụng CSS để tạo phong cách hiện đại, màu xanh – tím, hiệu ứng kính mờ (glassmorphism).
  Biểu đồ sử dụng Plotly để hiển thị tương tác và rõ ràng hơn.

Ứng dụng được triển khai trực tuyến thông qua Pyngrok, cho phép truy cập dễ dàng từ máy tính hoặc điện thoại.

Nhờ đó, mô hình dự đoán có thể được sử dụng thực tế bởi các nhà quản lý để theo dõi và ra quyết định nhanh chóng.



CHƯƠNG 6: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1 Kết luận:

Trong suốt quá trình thực hiện đề tài "Dự đoán doanh số bán lẻ Rossmann bằng mô hình học máy", nhóm đã hoàn thành đầy đủ các bước của một dự án khai phá dữ liệu thực tế – từ thu thập, làm sạch, phân tích, xây dựng mô hình, đánh giá kết quả đến triển khai ứng dụng web cho người dùng cuối.

Bộ dữ liệu Rossmann mà nhóm sử dụng chứa nhiều thông tin quan trọng như doanh số, số khách hàng, chương trình khuyến mãi, khoảng cách đối thủ, loại cửa hàng, thời gian, mùa vụ,… Đây là nguồn dữ liệu tương đối phức tạp, yêu cầu nhiều bước xử lý trước khi huấn luyện mô hình. Nhóm đã tiến hành các thao tác tiền xử lý như xử lý giá trị thiếu, chuyển đổi dữ liệu ngày tháng, tạo biến mới (ví dụ: mùa lễ hội, quý trong năm, kỳ khuyến mãi) để mô hình học được xu hướng thực tế của doanh thu.

Ba mô hình được xây dựng và so sánh gồm:

- Linear Regression: mô hình tuyến tính cơ bản, đơn giản nhưng không thể hiện tốt các quan hệ phi tuyến.
- Random Forest: mô hình cây quyết định ngẫu nhiên, cho kết quả khá tốt nhưng tốn thời gian huấn luyện hơn.
- LightGBM: mô hình boosting hiện đại, cho độ chính xác cao và thời gian huấn luyện nhanh.

Kết quả đánh giá trên tập dữ liệu kiểm tra cho thấy LightGBM là mô hình tối ưu nhất, đạt:

- RMSE = 546.92
- $R^2 = 0.9754$

Mô hình này học được mối quan hệ phức tạp giữa các yếu tố ảnh hưởng đến doanh thu, đồng thời giữ được sự ổn định khi dự đoán cho các giai đoạn 6 tháng và 12 tháng sau.

Nhóm cũng đã xây dựng ứng dụng web dự đoán doanh thu bằng Streamlit, giúp người dùng chỉ cần nhập thông tin cửa hàng là có thể nhận kết quả ngay, kèm theo biểu đồ minh họa trực quan. Giao diện thân thiện, có thể truy cập dễ dàng qua Pyngrok, giúp việc ứng dụng mô hình vào thực tế trở nên khả thi hơn.

Tóm lại, đề tài đã chứng minh rằng việc áp dụng học máy vào dự báo doanh số bán lẻ không chỉ mang lại độ chính xác cao mà còn giúp nhà quản lý ra quyết định nhanh hơn và chính xác hơn, hỗ trợ tốt cho hoạt động điều hành và hoạch định chiến lược kinh doanh.

6.2 Hướng phát triển:

Mặc dù mô hình hiện tại đã đạt hiệu quả tốt, nhóm nhận thấy vẫn còn nhiều hướng có thể mở rộng và cải thiện trong tương lai:

1. **Bổ sung thêm dữ liệu thực tế**
   - Thu thập thêm dữ liệu từ nhiều năm, nhiều khu vực hoặc nhóm sản phẩm khác nhau.
   - Bổ sung các yếu tố bên ngoài như thời tiết, ngày lễ, sự kiện đặc biệt, vì đây là những yếu tố có thể ảnh hưởng mạnh đến doanh thu.
2. **Cải thiện mô hình dự báo**
   - Thử nghiệm thêm các mô hình tiên tiến như XGBoost, CatBoost hoặc Prophet để so sánh hiệu suất.
   - Tối ưu tham số tự động (bằng GridSearch hoặc RandomizedSearch) để giảm sai số và tăng khả năng tổng quát hóa.
3. **Tăng cường phần trực quan hóa và phân tích**
   - Xây dựng thêm dashboard hiển thị biểu đồ tăng trưởng, tỉ lệ thay đổi, và dự báo theo từng tháng hoặc khu vực.
   - Cung cấp báo cáo ngắn gọn giúp người quản lý dễ theo dõi biến động doanh thu.
4. **Phát triển ứng dụng web hoàn chỉnh hơn**
   - Tích hợp đăng nhập người dùng, lưu lịch sử dự đoán và cho phép xuất file kết quả (Excel, PDF).
   - Đưa ứng dụng lên các nền tảng như Streamlit Cloud hoặc Hugging Face Spaces để hoạt động ổn định hơn.
   - Cải thiện giao diện để thân thiện với cả điện thoại và máy tính bảng.
5. **Ứng dụng trong môi trường doanh nghiệp**
   - Liên kết mô hình với cơ sở dữ liệu thực tế của doanh nghiệp, để dự đoán theo thời gian thực.
   - Tự động cập nhật dữ liệu hàng ngày, giúp quản lý nhanh chóng nắm được xu hướng tăng - giảm doanh thu.

6.3 Kết luận chung:

Qua đề tài này, nhóm đã có cơ hội áp dụng kiến thức về khai phá dữ liệu và học máy vào một bài toán thực tế trong lĩnh vực bán lẻ.

Việc xây dựng mô hình LightGBM và tích hợp lên ứng dụng web cho thấy khả năng chuyển đổi kết quả nghiên cứu thành công cụ hữu ích cho doanh nghiệp.

Đây là tiền đề để nhóm tiếp tục phát triển các hệ thống dự báo thông minh và tự động hóa ra quyết định kinh doanh trong tương lai.

CHƯƠNG 7: TÀI LIỆU THAM KHẢO

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:

1. Python 3 - ngôn ngữ lập trình chính.
2. Scikit-learn, XGBoost, LightGBM - huấn luyện và đánh giá mô hình học máy.
3. Pandas, NumPy - xử lý và phân tích dữ liệu.
4. Matplotlib, Seaborn - trực quan hóa kết quả.
5. Streamlit – xây dựng giao diện web trực quan cho người dùng.
6. Microsoft Word (docx) - biên soạn và trình bày báo cáo.