BÀI DỰ THI CUỘC THI PHÂN TÍCH DỮ LIỆU "DATAFLOW 2025: MASTERING THE DATA WAVES"

BÀI TOÁN: FORECASTING BUSINESS PERFORMANCE - DỰ BÁO HIỆU SUẤT KINH DOANH

Nhóm thực hiện: Monarchie

Hà Nội, tháng 02/2025

Chapter 1: MÔ TẢ BÀI TOÁN

1. Tóm tắt đề bài:

- Từ dữ liệu năm 2010 - 2020, phân tích và dự báo hiệu suất kinh doanh của một công ty thời trang tại Mỹ trong hai năm (2021–2022).
- Đánh giá xu hướng doanh thu, khối lượng bán hàng và tác động của các yếu tố kinh tế, thị trường đến hiệu suất kinh doanh.
- Dự đoán tình hình kinh doanh năm tiếp theo để hỗ trợ:
    - Lập ngân sách hàng tồn kho
    - Tối ưu chuỗi cung ứng
    - Quản lý hàng tồn kho hiệu quả
    - Nâng cao chiến lược kinh doanh trong ngành thời trang cạnh tranh

- **Mục tiêu phân tích:**
    - Phân tích dữ liệu lịch sử và trực quan hóa xu hướng doanh thu, khối lượng bán hàng theo thời gian.
    - Tạo và diễn giải biểu đồ xu hướng doanh thu, doanh số theo năm, tháng và quý.
    - Xác định các yếu tố ảnh hưởng đến doanh thu (tính mùa vụ, khu vực, dòng sản phẩm).

2. Mô tả các trường dữ liệu

- **Nguồn:** Tập dữ liệu nội bộ – Forecasting Business Performance
- **Giai đoạn huấn luyện:** 10 năm (2011 – 2020)
- **Giai đoạn kiểm tra:** 2 năm (2021 – 2022)
- Bảng chính **Salesfact** gồm các đặc trưng:
    - **ProductID:** Mã sản phẩm
    - **Date:** Ngày giao dịch
    - **Zip:** Mã khu vực bán hàng
    - **Units:** Số lượng sản phẩm bán được
    - **Revenue:** Doanh thu từ bán hàng
    - **COGS:** Giá vốn
- Tệp dữ liệu còn bao gồm 2 bảng phụ để bổ sung thông tin:
    - **Geography:** Gồm 39948 dòng và 5 cột
        - **Zip:** 39948 - non-null - int64
        - **City**: 39948 - non-null - object
        - **State**: 39948 - non-null - object
        - **Region**: 39948 - non-null - object
        - **District**: 39948 - non-null - object
    - **Product:** Gồm 2412 dòng và 4 cột
        - **Category:** 2412 - non-null - object
        - **Segment:** 2412 - non-null - object
        - **Product:** 2412 - non-null - object
        - **ProductID:** 2412 - non-null - int64

Chapter 2: PHÂN TÍCH KHÁM PHÁ DỮ LIỆU

1. Tiền xử lý dữ liệu

Để tổng hợp được hết tất cả dữ liệu đề bài và phục vụ mục đích phân tích cũng như thực thi các mô hình, Monarchie sẽ tiến hành ghép 3 bảng dữ liệu *train*, *geography* và *product* của đề bài thành một tập dữ liệu tổng hợp rồi mới thực hiện tiền xử lý dữ liệu thông qua các bước dưới đây:

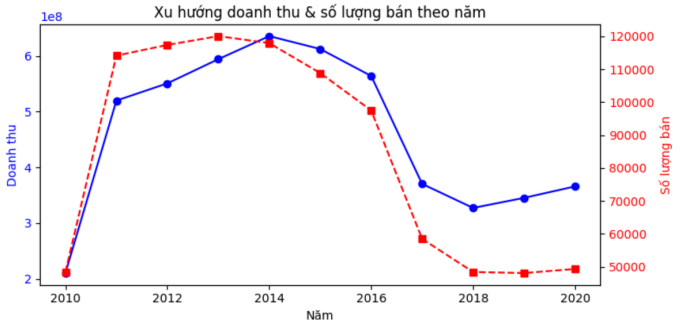
- Xử lý dữ liệu trùng lặp (Duplicates)
- Xử lý dữ liệu bị thiếu (Missing values)

2. Thống kê mô tả

Để phù hợp với mục đích thống kê và phân tích, ta sẽ thực hiện bước nhóm các cột Doanh thu (Revenue), Số lượng sản phẩm (Units) và Giá vốn (COGS) theo 3 cách: theo cột Mã sản phẩm (ProductID), cột Ngày giao dịch (Date) và cột Vùng miền (Region).

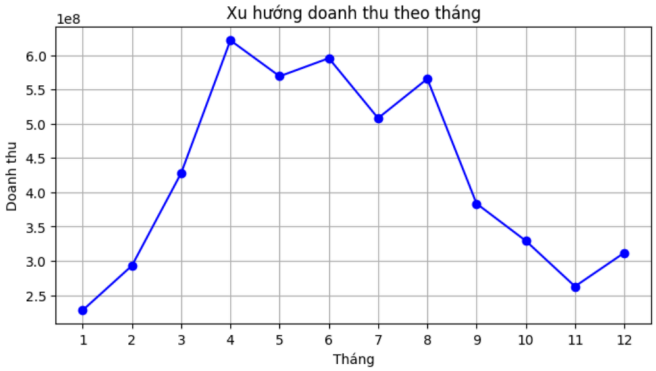
2.1. Phân tích xu hướng kinh doanh chung theo thời gian

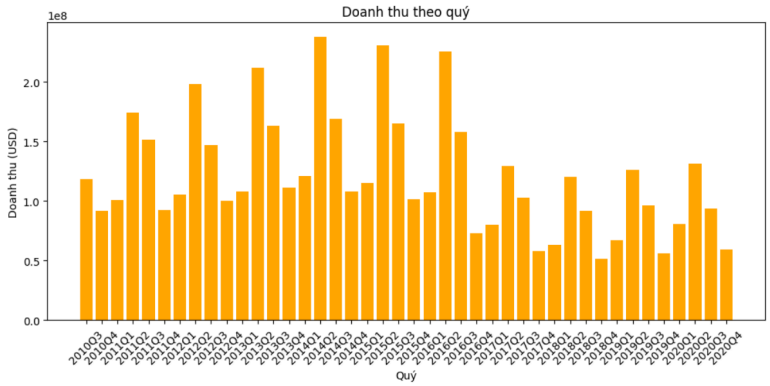
- Đầu tiên, ta sẽ trực quan hóa tổng doanh thu và tổng số lượng bán theo từng năm để thấy được xu hướng tổng quan.



Nhận xét: Ở giai đoạn từ năm 2010 đến khoảng 2013, cả doanh thu và số lượng sản phẩm có sự tăng trưởng đáng kể. Tuy nhiên từ khoảng năm 2014 trở đi, hai đại lượng đều có sự sụt giảm đáng kể, cho đến năm 2018 mới có khởi sắc.

- Để thấy rõ hơn xu hướng thời gian, ta tiếp tục trực quan hóa dữ liệu theo tháng và theo quý.

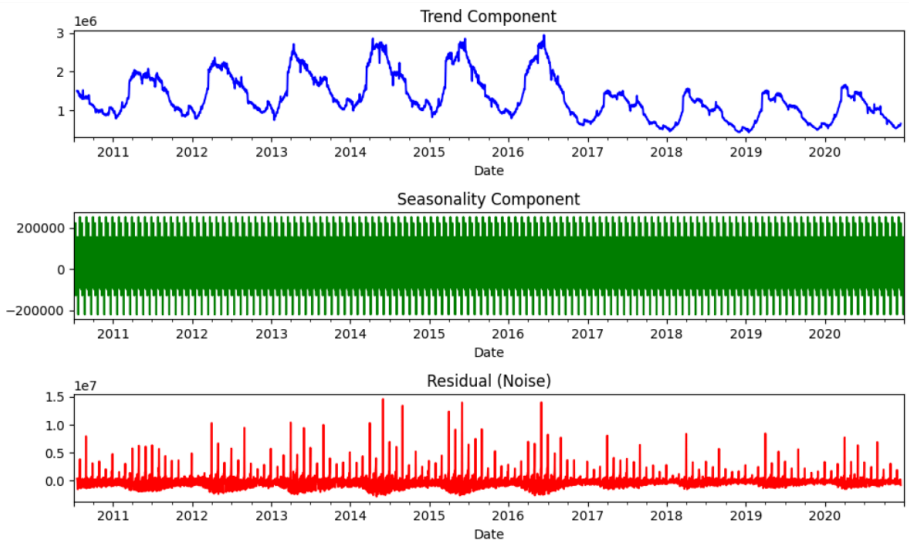




Quan sát cho thấy, thời gian từ tháng 4 đến tháng 8 là khoảng thời gian có doanh thu cao nhất. Tuy nhiên nếu xét theo quý, quý 1 và quý 2 mới là thời gian có tổng doanh thu cao nhất của tất cả các năm.

2.2. Mô tả các yếu tố ảnh hưởng đến doanh thu

- Ở bước này, ta thực hiện trực quan hóa để kiểm tra tính mùa vụ và nhiễu của dữ liệu theo thời gian.



  - Xu hướng chung (Trend Component): Công ty có xu hướng đạt doanh thu cao vào khoảng thời gian giữa năm và giảm dần vào cuối năm.
  - Tính mùa vụ (Seasonality Component): Doanh thu của công ty có tính mùa vụ tương đối mạnh, có một mẫu hình lặp đi lặp lại nhiều lần.
  - Độ nhiễu (Noise): Giai đoạn năm 2010 - 2012 có biến động không quá nhiều, tuy nhiên từ năm 2013 - 2017 dữ liệu có nhiều biến động mạnh.

  Kết luận: Doanh thu có tính mùa vụ và bị ảnh hưởng bởi một số yếu tố vào những khoảng thời gian nhất định.
- Tiếp theo, ta thực hiện trực quan hóa tổng doanh thu theo Khu vực (Region), theo Danh mục sản phẩm (Category) và theo Phân khúc khách hàng (Segment). Quan sát cho thấy:
  - Khu vực phía Đông có doanh thu cao hơn nhiều so với hai khu vực còn lại. Như vậy đây là khu vực quan trọng nhất, cần tập trung phát triển các chiến lược kinh doanh. Ngoài ra, doanh nghiệp có thể tập trung phát triển thêm ở khu vực phía Tây và vùng Trung tâm để cải thiện tổng doanh thu.

- Danh mục sản phẩm có doanh thu cao nhất là Urban, sau đó đến Rural, cuối cùng Mix và Youth có tổng doanh thu thấp nhất.
- Phân khúc có doanh thu cao nhất là Convenience và Moderation, theo sau đó là Extreme, Productivity và các phân khúc còn lại.

Chapter 3: ỨNG DỤNG MÔ HÌNH HỌC MÁY

1. Mô hình XGBOOST

- Kết quả kiểm thử mô hình được đánh giá thông qua các hệ số sau:
  - Chỉ số R2 của mô hình là: 1.0
  - Chỉ số MAPE của mô hình là: 0.0
  - Chỉ số RMSE của mô hình là: 0.0
- Mức độ chính xác của mô hình: 100.0%. Như vậy, mô hình được xây dựng có khả năng dự đoán chính xác hoàn toàn doanh thu của doanh nghiệp trong thời gian chỉ định.
- Tính toán sai số giữa dữ liệu gốc và dữ liệu dự đoán theo hai chỉ số: Mean Squared Error (MSE), Mean Absolute Error (MAE):
  - Mean Squared Error (MSE): 2346491.92
  - Mean Absolute Error (MAE): 852.51

2. Mô hình SARIMA

- Dựa vào ACF và PACF plot, chúng tôi chọn bộ order (5,0,0,12) cho mô hình SARIMA
- Sau khi chạy mô hình:
  - **Các tham số AR**: Hầu hết các tham số AR (AutoRegressive) có giá trị p-value rất nhỏ, điều này cho thấy chúng có ảnh hưởng mạnh đến dự báo, ngoại trừ một số tham số như ar.L2, ar.L4 và ar.L5 có p-value lớn hơn 0.05.
  - **Mối quan hệ mùa vụ**: Các tham số mùa vụ (ar.S.L12, ar.S.L24, ...) có p-value rất thấp, cho thấy mối quan hệ mùa vụ rõ ràng trong dữ liệu, đặc biệt là tại các độ trễ dài hạn như L12, L24, L36.
- Kết quả đánh giá mô hình:
  - RMSE: 1328340.58
  - R-squared (R2): -0.7491372665755636

3. So sánh 2 mô hình

	XG Boost	SRIMA
Các chỉ số đánh giá hiệu suất	Chỉ số đánh giá hiệu suất khá cao, cho thấy khả năng dự đoán chính xác rất cao	Chỉ số đánh giá hiệu suất không cao
Thời gian thực hiện	Khá nhanh, khoảng 30 giây	21 giây
Độ chính xác	Độ chính xác tuyệt đối	Độ chính xác thấp