College of Engineering and Computer Science

Business Performance Forecasting

A report from VinDataHub

Author:

Pham Dinh Hieu (BSc in Data Science)
Cao Pham Minh Dang (BSc in Data Science)
Nguyen Thi Bao Tien (BSc in Data Science)

Mục lục

1 Tóm tắt **3**

2 Giới thiệu bộ dữ liệu **3**

3 Phương pháp tiếp cận **3**

4 Kết quả quan trọng & Đề xuất **3**

4.1 Sự tương quan giữa giá bán và doanh thu . . . 3
4.2 Tỉ lệ doanh thu của các tổ hợp Category-Segment . . . 4
4.3 Xu hướng doanh thu theo thời gian . . . 4
4.4 Các cửa hàng trọng điểm . . . 5
4.5 Đề xuất kinh doanh . . . 5
4.6 Dự đoán doanh thu 2 năm tiếp theo . . . 6
4.6.1 SARIMAX . . . 6
4.6.2 LSTM . . . 6
4.6.3 Transformer . . . 6

1 Tóm tắt

Dựa vào mô hình SARIMAX và LSTM, chúng tôi dự đoán chính xác tính mùa vụ (seasonality) của bộ dữ liệu. Từ đó, bản báo cáo này sẽ giúp doanh nghiệp dự đoán những biến động kinh doanh và có chiến lược phù hợp để tối ưu hàng tồn kho và nâng cao lợi nhuận trong những năm kế tiếp.

2 Giới thiệu bộ dữ liệu

Bộ dữ liệu chứa 6233 giao dịch của 24103 cửa hàng. Kết quả kinh doanh được ghi lại từ năm 04/07/2010 đến 01/07/2022. Không tồn tại việc thiếu dữ liệu hoặc dữ liệu sai định dạng ở các hàng.

Chúng tôi đã ghép 3 bộ dữ liệu (**SalesFact**, **Geography**, và **Product**) và chọn các cột ['Date', 'Units', 'Revenue', 'COGS', 'Year', và 'Month']. Sau đó, ['Category', 'Segment'] được kết hợp với nhau để xác định nhóm sản phẩm có giá trị nhất và tối ưu hoá chiến lược cửa hàng.

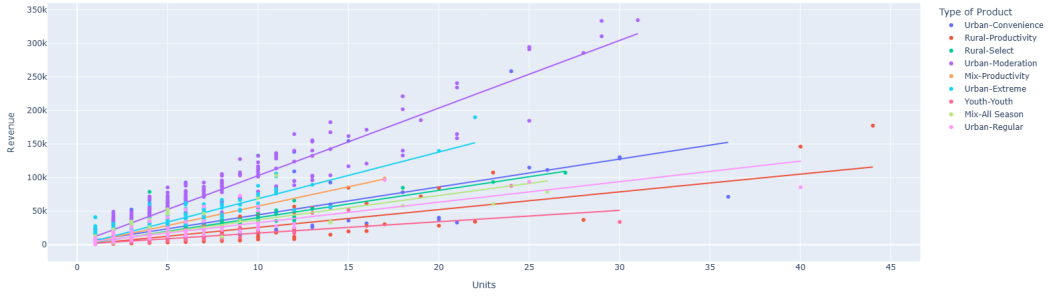
3 Phương pháp tiếp cận

Để thực hiện phân tích dữ liệu bán hàng, một quy trình có hệ thống đã được áp dụng nhằm đảm bảo độ chính xác và tin cậy trong kết quả. Phương pháp phân tích bao gồm các bước sau:

- Thu thập dữ liệu: Dữ liệu bán hàng từ năm 2012 đến 2020, bao gồm: **SalesFact** (Mã sản phẩm, Ngày, Mã Zip, Đơn vị, Lợi nhuận, Giá gốc); **Geography** (Mã Zip, Thành phố, Bang, Khu vực, Quận); **Product** (Danh mục, Phân khúc, Sản phẩm, Mã sản phẩm).
- Tiền xử lý dữ liệu: Trước khi phân tích, dữ liệu được xử lý để đảm bảo chất lượng và tính nhất quán. Các bước xử lý bao gồm loại bỏ dữ liệu trùng lặp, xử lý giá trị bị thiếu và khắc phục các lỗi hoặc sự không nhất quán trong dữ liệu.
- Phân tích dữ liệu khám phá (EDA): Phân tích dữ liệu khám phá được thực hiện để có cái nhìn sơ bộ về tập dữ liệu. Các thống kê mô tả, kỹ thuật trực quan hóa và biểu đồ được sử dụng để trả lời các câu hỏi kinh doanh.
- Báo cáo: Các phát hiện, thông tin chi tiết và đề xuất được tổng hợp thành một báo cáo phân tích dữ liệu toàn diện. Báo cáo trình bày kết quả phân tích một cách rõ ràng và có tổ chức, sử dụng hình ảnh trực quan, bảng biểu và giải thích súc tích.

4 Kết quả quan trọng & Đề xuất

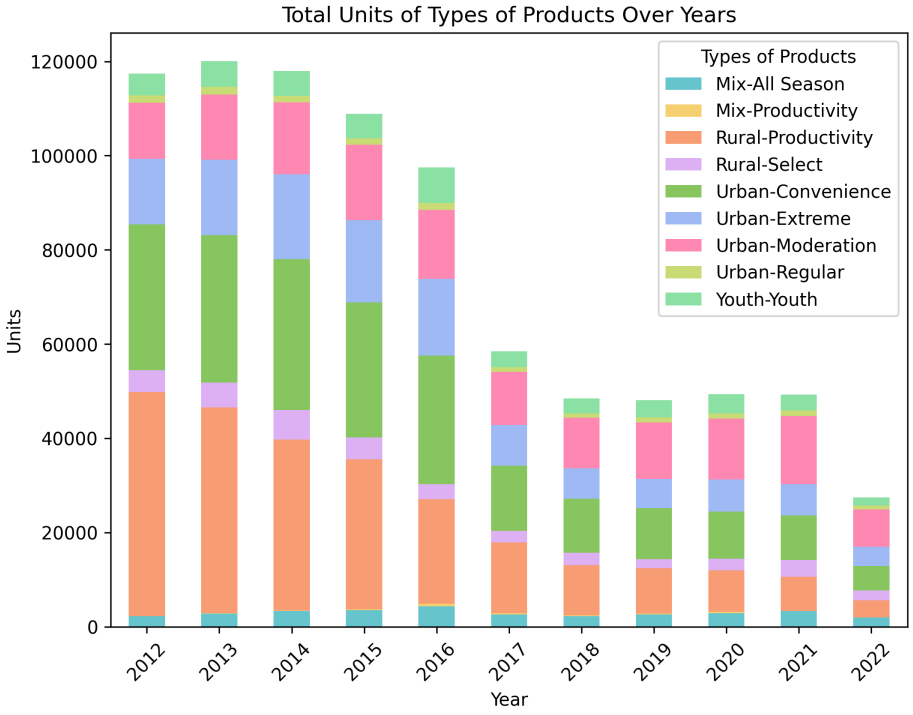
4.1 Sự tương quan giữa giá bán và doanh thu



Hình 1: Giá bán so với doanh thu

Urban-Moderation là loại sản phẩm có doanh thu cao nhất, có nghĩa là nhu cầu thị trường đối với sản phẩm này đang mạnh. Rural-Productivity và Rural-Select có đường xu hướng bằng phẳng hơn, có nghĩa là nhu cầu thị trường yếu hơn so với các danh mục đô thị.

4.2 Tỉ lệ doanh thu của các tổ hợp Category-Segment



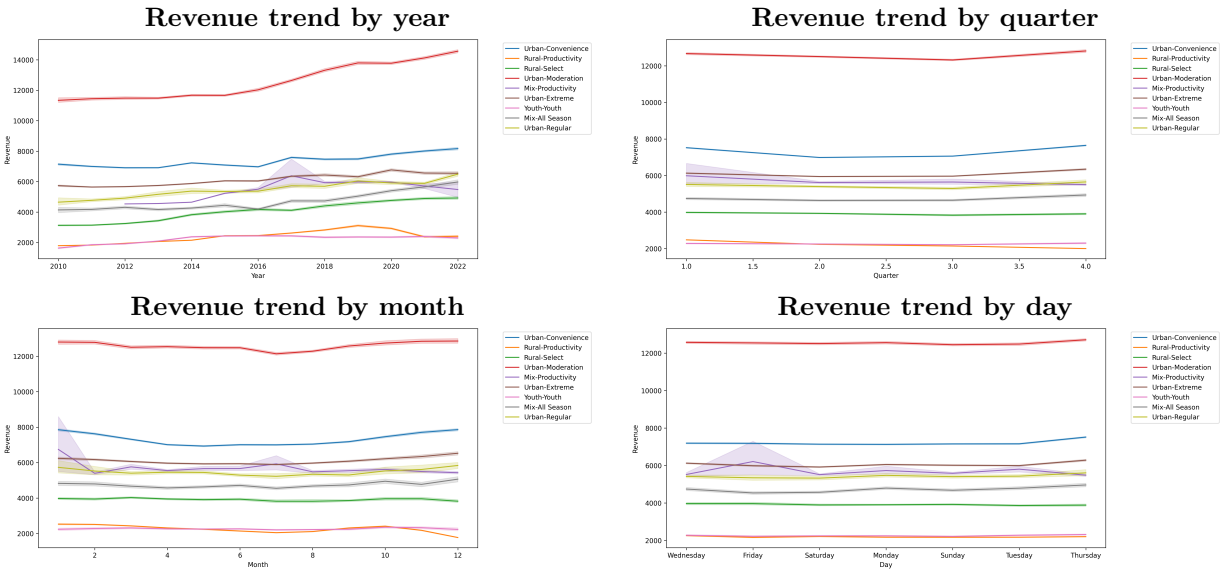
Urban-Moderation, Urban-Convenience: Thời trang phù hợp với cuộc sống hàng ngày luôn giữ được thị phần qua mỗi năm và còn giúp doanh nghiệp đứng vững trong nền kinh tế khó khăn (2018-2022).

Rural-Productivity, Youth-Youth: Doanh số lớn trong nửa đầu thập kỷ nhưng giảm dần qua từng năm. Nếu mục tiêu là tối ưu hóa doanh thu trong thời gian tới, có thể cân nhắc chuyển hướng đầu tư sang các danh mục đang có xu hướng tăng trưởng ổn định hơn.

Youth-Youth: Tích hợp vào danh mục đô thị để tiếp cận khách hàng tốt hơn.

Mix-All Season, Mix-Productivity: Giảm tỷ trọng nếu không còn hiệu quả.

4.3 Xu hướng doanh thu theo thời gian



Theo năm

Tăng trưởng doanh thu chậm lại trong 2 năm gần đây, có thể do thị trường bão hòa hoặc chiến lược marketing chưa tối ưu.

Theo quý

Quý 4 luôn có doanh thu cao nhất, chủ yếu nhờ tháng 12. Quý 2 thường giảm nhẹ do không có nhiều dịp mua sắm đặc biệt.

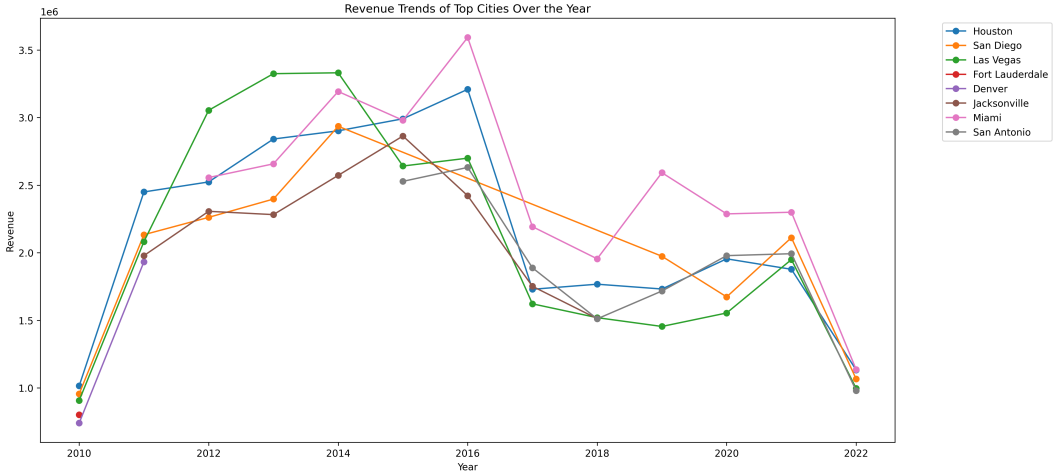
Theo tháng

Tháng 1 - 2 có mức tăng nhẹ, khả năng do nhu cầu mua sắm sau kỳ nghỉ lễ và Tết Nguyên Đán. Tháng 5 & 9 có doanh thu thấp nhất, có thể do không có sự kiện mua sắm lớn nào.

Theo ngày

Doanh số ổn định, nhưng cao hơn vào thứ Sáu, có thể do khách hàng có thói quen mua sắm để xả stress sau các ngày đi làm trong tuần.

4.4 Các cửa hàng trọng điểm



Hình 2: Top 5 cửa hàng qua từng năm

Houston, Las Vegas, Miami, San Diego, Jacksonville, San Antonio là nhóm các thành phố có doanh thu cao nhất qua từng năm nên cần cân nhắc mở rộng quy mô, cung cấp chương trình VIP và tối ưu trải nghiệm khách hàng.

4.5 Đề xuất kinh doanh

Chiến lược Tiếp thị Mục tiêu

- Mở rộng quy mô và tận dụng xu hướng tiêu dùng ở các thành phố dẫn đầu.
- Tăng cường đầu tư vào nhóm Rural-Productivity & Urban-Convenience, vì đây là hai phân khúc mang lại doanh thu cao nhất.
- Đẩy mạnh tiếp thị cho Youth-Youth và Mix-All Season nhằm gia tăng thị phần và tối ưu hóa doanh thu.

Chiến lược Dự trữ & Mở rộng Sản phẩm

- Nếu tập trung vào tăng trưởng, mở rộng dòng sản phẩm thời trang Urban-Productivity.
- Đẩy mạnh Urban-Convenient, tận dụng xu hướng mua sắm đô thị.

Chiến lược Bán hàng & Định giá

- Urban-Moderation" có mức tăng trưởng doanh thu mạnh trên mỗi đơn vị, do đó cân nhắc mở rộng dòng sản phẩm này hoặc tăng cường các chiến dịch tiếp thị để tận dụng tiềm năng.
- Một số dòng sản phẩm, đặc biệt là Rural-Productivity, có thể cần tăng sản lượng bán ra hoặc áp dụng chiến lược giảm giá theo gói để cải thiện doanh thu.

Chiến lược thời gian

- Tận dụng mùa cao điểm (Tháng 12 & Tháng 1-2) để đẩy mạnh khuyến mãi, chiến dịch tiếp thị sớm nhằm tối đa hoá doanh số.
- Cải thiện doanh thu trong giai đoạn thấp điểm (Tháng 5 & 9, Quý 2) bằng cách triển khai chương trình mua sắm ưu đãi và xây dựng chiến lược content marketing để giữ chân khách hàng.

4.6 Dự đoán doanh thu 2 năm tiếp theo

4.6.1 SARIMAX

Các bước thực hiện (**File code**):

- Phân tích dữ liệu cho thấy max là 17,472,120.75, min là 1,101.87, cần chuẩn hóa do phân tán lớn.
- Quan sát thấy mùa vụ rõ và quan hệ chặt chẽ giữa doanh thu và số lượng đơn vị.
- Kiểm định Dickey-Fuller (p = 0.18) cho thấy dữ liệu chưa dừng ⇒ áp dụng sai phân bậc 1.
- Dùng SARIMAX với tham số (1, 0, 0) và (1, 0, 2, 12), tích hợp Units và COGS.

Kết quả: RMSE: 981,791.33, R^2: 0.044, MAPE: 57.3%. (**Link ảnh plot**)

Đánh giá: SARIMAX thể hiện khả năng nổi bật trong việc **dự đoán chính xác mùa vụ và xu hướng**, bắt được cái yếu tố bất thường trong bộ test nhưng thường lệch 1 ngày nên cho error cao.

4.6.2 LSTM

Các bước thực hiện (**File code**):

- Do dữ liệu phân hóa lớn, sử dụng MinMaxScaler cho các features
- Quan sát thấy mối quan hệ chặt chẽ giữa ngày trong tuần, số lượng đơn vị, và COGS với doanh thu, sử dụng những yếu tố đó để dự đoán doanh thu
- Dữ liệu được chia thành chuỗi thời gian với n_steps_in = 30 và n_steps_out = 1
- Sử dụng mô hình stacked LSTM với 3 tầng (hidden_size = 200), tối ưu bằng ADOPT, loss RMSE

Kết quả: RMSE: 590015.9317, R2: 0.6549, MAPE: 58.57% (**Link ảnh plot**)

Đánh giá: Mô hình LSTM thể hiện khả năng **dự đoán tốt xu hướng và quan hệ phi tuyến giữa các biến**, với R^2 khá cao (0.6549) cho thấy khả năng giải thích biến động doanh thu. Tuy nhiên, MAPE cao (58.57%) chỉ ra sai số phần trăm lớn, có thể do mô hình chưa bắt kịp các biến động bất thường hoặc lệch pha nhẹ trong dữ liệu test.

4.6.3 Transformer

Các bước thực hiện (**File code**):

- Do dữ liệu phân hóa lớn, sử dụng MinMaxScaler cho các features
- Quan sát thấy mối quan hệ chặt chẽ giữa ngày trong tuần, số lượng đơn vị, và COGS với doanh thu, sử dụng những yếu tố đó để dự đoán doanh thu
- Dữ liệu được chia thành chuỗi thời gian với n_steps_in = 30 và n_steps_out = 1 để dự đoán doanh thu.
- Sử dụng mô hình Transformer với 2 tầng encoder, kích thước vector nhúng d_model = 64, và 4 đầu attention. Mô hình được tối ưu bằng ADOPT, với hàm mất mát RMSELoss.

Kết quả: R^2: 0.4584, MAPE: 49.17%, RMSE: 739130.4162 (**Link ảnh plot**)

Đánh giá: Cách xử lí và các chỉ số được set khá tương tự LSTM. Để cải thiện Transformer, có thể tăng số **encoder layers** hoặc số **attention heads** để cải thiện khả năng học xu hướng dữ liệu. Ngoài ra, có thể thử nghiệm **mô hình hybrid (Transformer + LSTM)** để tận dụng lợi thế của cả hai mô hình. Việc **tối ưu preprocessing**, như thêm các đặc trưng quan trọng hơn (giá cả, xu hướng theo mùa, khuyến mãi), cũng có thể giúp mô hình dự đoán chính xác hơn.