Báo cáo

Cuộc thi: DataFlow
Đề tài: FORECASTING BUSINESS PERFORMANCE

Nhóm: **BÁNH ĐẬU XANH**

Thành viên: Trương Minh Hiếu

Phạm Văn Dũng

Trần Viết Kiên

Chu Thị Mai Duyên

Ngày 24 tháng 2 năm 2025

Mục lục

Phần 1: PHÂN TÍCH DỮ LIỆU

I. Tổng quan và tiền xử lý dữ liệu

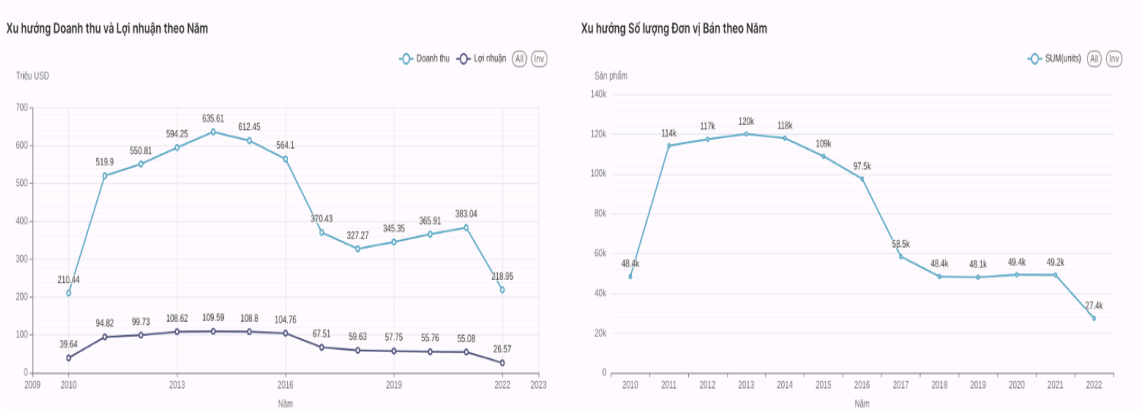
- Dữ liệu gồm 3 bảng chính: Product (2.412 dòng, 4 cột) chứa chi tiết sản phẩm, Geography (39.948 dòng, 5 cột) lưu thông tin địa lý, và SalesFact (976.243 dòng, 6 cột) ghi nhận giao dịch với ID sản phẩm, ngày bán, mã ZIP, số lượng, Revenue và COGS.

- Chất lượng dữ liệu tốt, trừ 41 giá trị thiếu ở cột Revenue của SalesFact. Ba bảng được gộp theo ProductID và Zip, giá trị khuyết được điền bằng **hồi quy tuyến tính** dựa trên COGS, nhờ tương quan mạnh (**95%**) với Revenue.

II. Khám phá dữ liệu

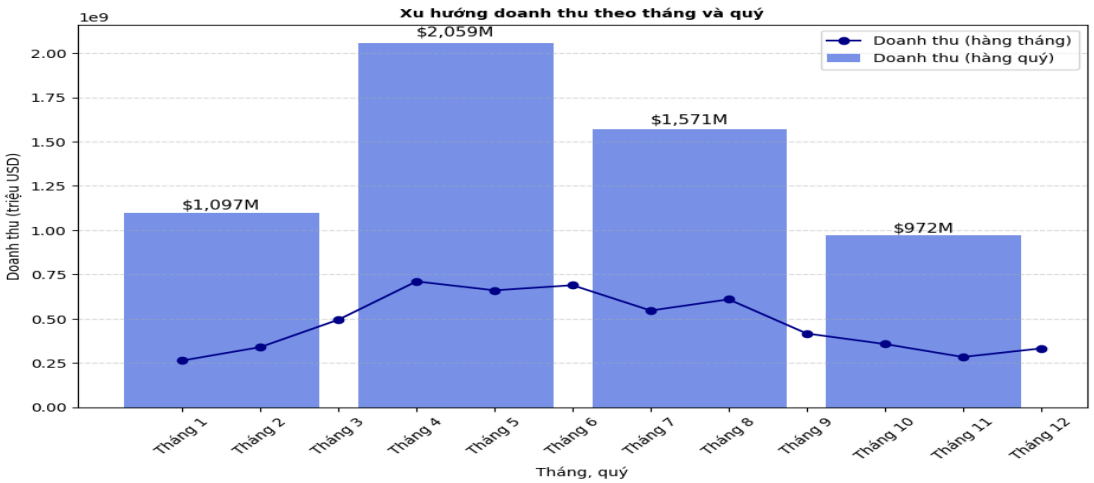
1. Phân tích theo thời gian

- Do doanh thu, lợi nhuận và số lượng bán có **mối tương quan thuận** (một biến tăng kéo theo các biến còn lại cũng tăng), nên nhóm tập trung vào phân tích yếu tố doanh thu.



Hình 1: *Xu hướng doanh thu, lợi nhuận và số lượng bán theo năm*

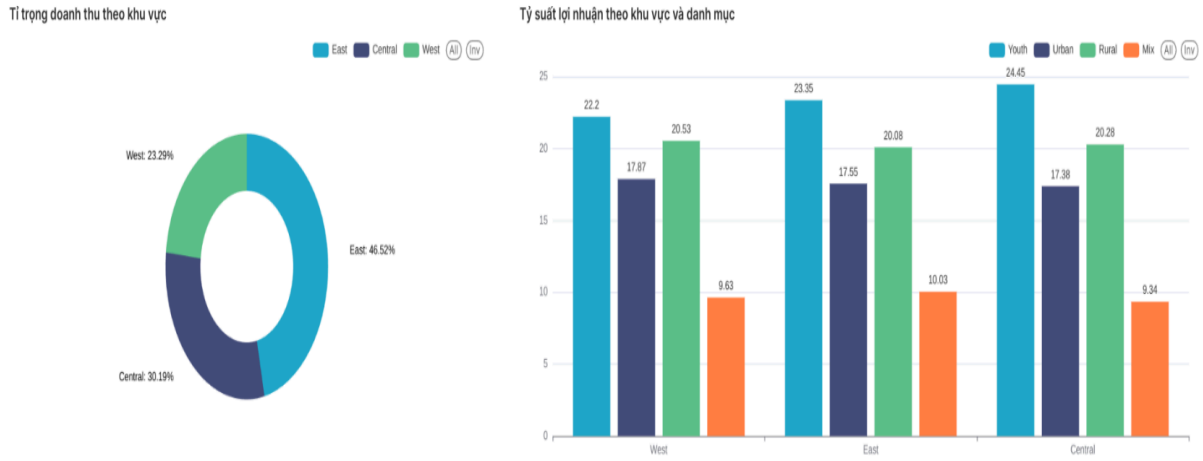
- Từ **2010-2014**, doanh thu tăng vọt (đỉnh điểm **147.06%** năm 2011) nhờ kinh tế phục hồi, thương mại điện tử và thời trang nhanh bùng nổ, nhưng đến **2015-2018** lại lao dốc (**-34.33%** năm 2017) do đối thủ mới với công nghệ vượt trội, sản phẩm lỗi thời và suy thoái kinh tế cục bộ, trước khi ổn định từ **2019-2022.**



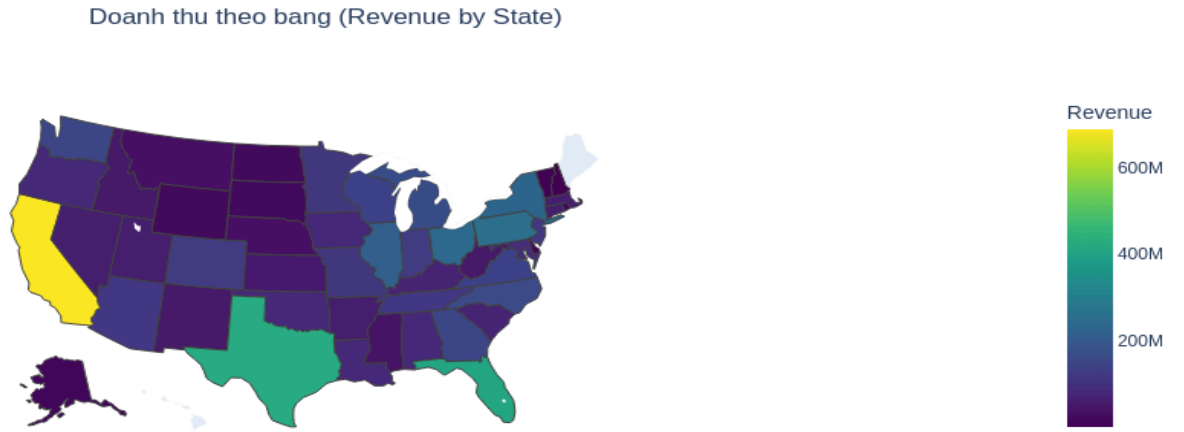
Hình 2: *Xu hướng doanh thu theo tháng và quý*

- **Xu hướng doanh thu theo tháng và quý** cho thấy **tính mùa vụ rõ rệt**, ổn định qua các năm, với **quý 2** (đặc biệt từ tháng 4 đến tháng 6) luôn là **đỉnh cao**, chiếm **35-40%** doanh thu cả năm nhờ các sự kiện mua sắm sôi động, trong khi **quý 4** thường **chạm đáy** do thời điểm nghỉ đông ít hoạt động mua sắm.

2. Phân tích theo khu vực



Hình 3: *Tỷ trọng doanh thu và tỷ suất lợi nhuận theo khu vực và danh mục*



Hình 4: *Tỷ trọng doanh thu theo bang*

- **East** dẫn đầu về doanh thu và lợi nhuận, chiếm **46.7%** tổng doanh thu. Khu vực này có mật độ dân số cao, tập trung nhiều thành phố lớn (New York, Boston, Los Angeles…), du lịch phát triển và thu nhập bình quân cao, thúc đẩy nhu cầu tiêu dùng.

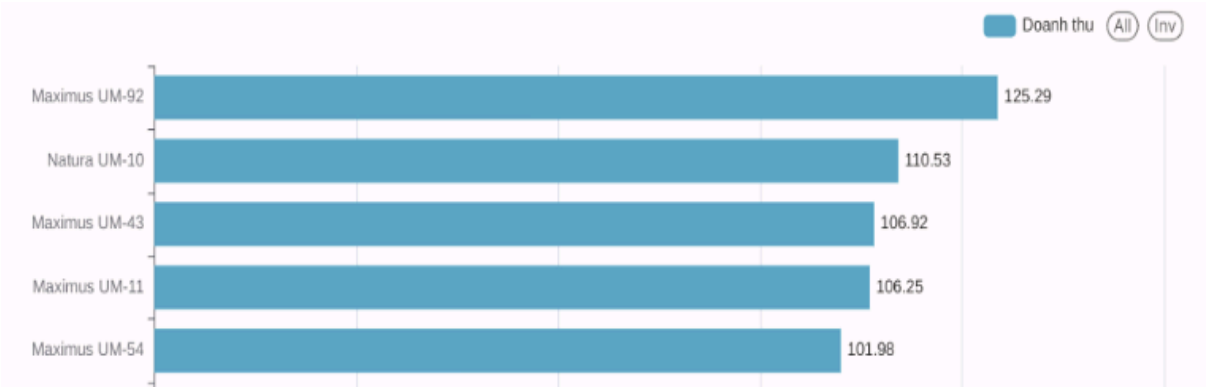
- **West** và **Central** có tỷ trọng thấp hơn, lần lượt **23.5%** và **29.9%**.

- Doanh thu phân bố không đồng đều ở các bang, cao nhất ở các bang lớn như **California, Texas, Florida**, trong khi khu vực Trung Tây và Đông Bắc ghi nhận mức thấp hơn.

3. Phân tích theo danh mục sản phẩm

- Danh mục **Urban** đóng vai trò chủ lực, chiếm **80%** tổng doanh thu với tỷ suất lợi nhuận ổn định ở mức **17.9%**. Trong khi đó, **Youth** là danh mục tiềm năng với tỷ suất lợi nhuận cao nhất (**23.3%**) nhưng doanh thu còn thấp. **Mix** cần được cải thiện do có tỷ suất lợi nhuận thấp nhất. **Rural** tuy doanh thu không cao nhưng vẫn duy trì ổn định. Hai danh mục **Mix** và **Youth** có tỷ trọng nhỏ hơn và tập trung chủ yếu ở khu vực **West** và **Central**.

- Top 5 sản phẩm doanh thu cao nhất:



Hình 5: *Top 5 sản phẩm doanh thu cao nhất*

Phần 2: MÔ HÌNH

I. Xây dựng

- Nhóm bắt đầu với mô hình thống kê truyền thống, áp dụng auto-arima cho dữ liệu chuỗi thời gian, sử dụng acf và pacf, dẫn đến ARIMA(3,1,3). Do nhận thấy tính mùa vụ trong doanh thu hàng tháng, nhóm mở rộng sang SARIMA, chọn ARIMA(1,1,0)(1,0,0)[12] để nắm bắt chu kỳ hàng năm, phù hợp với ngành thời trang.

- Nhóm cũng dùng **XGBoost** để dự đoán doanh thu, thêm đặc trưng tổng hợp tháng (doanh thu, số lượng sản phẩm, chỉ số tài chính), cùng các đặc trưng lag và rolling mean. Mô hình được tối ưu hóa (max_depth, learning_rate, n_estimators) và kiểm định chéo theo chuỗi thời gian, đạt RMSE và MAE thấp, đặc biệt sau khi loại bỏ tháng dữ liệu không đồng nhất.

- Ngoài ra, nhóm sử dụng mô hình học sâu **LSTM** với 3 lớp (192 đơn vị mỗi lớp), dropout 0,3, và learning rate 0,003459, tối ưu bằng **Keras Tuner**. Mô hình **Transformer** cũng được xây dựng với embedding 64, 2 attention heads, dropout 0,4, và một lớp dense 128 đơn vị để xử lý chuỗi dài.

- Cuối cùng, thử nghiệm mô hình **Prophet** của Facebook, định dạng lại dữ liệu và điều chỉnh riêng cho doanh thu và doanh số, tăng cường khả năng dự báo.

II. So sánh

Mô hình	RMSE	MAPE	R^2
Arima	9386847.1	42.67	0.41
Sarima	8075974.92	36.16	0.56

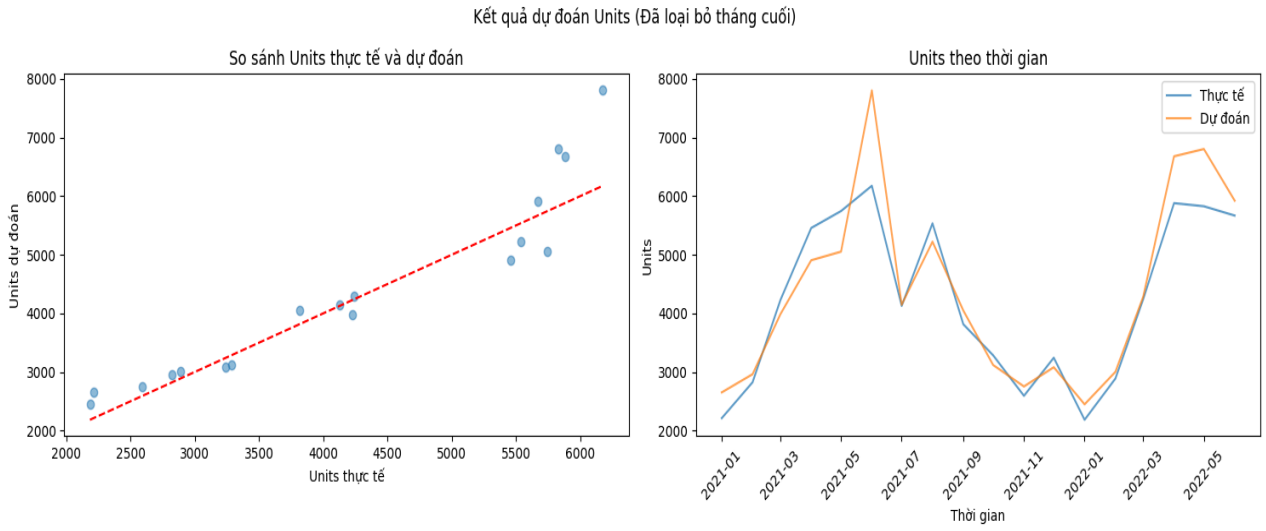
XGBoost	**1532869.01**	**2.69**	**0.97**
LSTM	9495544.97	49.26	0.39
Transformers	13636417.26	71.77	-0.25
Prophet	12600356.11	57.2	-0.07

***Bảng 1**: So sánh dự đoán doanh thu*

Mô hình	RMSE	MAPE	R^2
Arima	1448.97	47.42	0.11
Sarima	914.94	35.46	0.65
XGBoost	**561.41**	**8.85**	**0.83**
LSTM	943.05	34.59	0.62
Transformers	1291.99	46.39	0.3
Prophet	2476.5	71.49	-1.59

***Bảng 2**: So sánh dự đoán số lượng bán hàng*

- XGBoost vượt trội trong dự đoán doanh thu và số lượng bán hàng, vượt xa các mô hình khác. Ngược lại, Transformers thất bại do kiến trúc phức tạp gây overfitting khi dữ liệu không đủ lớn hoặc không phản ánh đặc thù ngành thời trang (mùa vụ, xu hướng). Prophet kém hiệu quả vì phương pháp phân rã cộng tính không phù hợp với dữ liệu phi tuyến tính và biến động mạnh trong bán lẻ thời trang.



***Hình 6:** Trực quan hóa kết quả dự đoán của mô hình XGBoost*