DataFlow 2025

Mastering the Data Waves

Team Project First Round

FORECASTING BUSINESS PERFORMANCE [VN]

Authors:

Phạm Duy Anh (leader) anhpd.23BI14023@usth.edu.vn
Lê Quốc Anh anhlq.23BI14025@usth.edu.vn
Nguyễn Thái Nhất Anh anhntn.23BI14024@usth.edu.vn
Nguyễn Lâm Tùng tungnl.23BI14446@usth.edu.vn

Team No Name

2025/2026

Mục lục

1 Phân tích dữ liệu (EDA & Visualize).

1.1 Dữ liệu doanh số chung

Về thời gian

Phân phối thu nhập của công ty trong suốt 10 năm từ 2010 đến 2020 được thể hiện tại hình 2, chúng ta có thể thấy rõ xu hướng doanh thu đang giảm dần từ năm 2014 đến nay và khoảng doanh thu lớn nhất của công ty rơi vào quý II và các tháng 4, 6, 8 hàng năm.

Trên thực tế sau khi phân tích các thành phố thì sự sụt giảm của doanh số lại được giải thích phần lớn do có đến hơn 4000 thành phố ước tính chỉ được ghi lại doanh thu ở một khoảng thời gian nhất định (dưới 5 ngày) sau đó biến mất, điều này diễn ra trong khoảng từ năm 2013-2017, từ kết quả ở hình 3 và 4, các thành phố này chiếm một phần không nhỏ dẫn đến sự tăng mạnh doanh thu tổng trong khoảng thời gian này. Cùng với đó, doanh thu của các thành phố lớn vẫn giữ được sự ổn định và không thể hiện sự giảm sút rõ rệt.

Có thể thấy rằng mỗi năm đều xuất hiện các thành phố chỉ có doanh thu dưới 5 ngày, nhưng từ 2012 đến 2016 con số này tăng lên rõ rệt (gấp đôi so với các năm khác), lí giải trực tiếp cho doanh thu cao bất thường tại thời gian này.

Một lí do khác có thể giải thích cho sự sụt giảm doanh thu các năm gần đây là do dịch bệnh Covid-19 xuất hiện và ảnh hưởng đến thói quen sinh hoạt và tiêu dùng của người dân.

Về phân bố khu vực

Các tiểu bang có mức phân phối doanh thu cao (từ nhạt đến đậm, từ xanh da trời đến xanh lá cây) chủ yếu tập trung ở khu vực Đông Nam của Vùng Trung Tâm, với giá trị cao nhất được ghi nhận tại bang California.

Một số thành phố chiếm tỉ trọng doanh thu chính của công ty là Miami, Houston, Las Vegas, San Diego và Jacksonville. Theo số liệu từ hình 6.

1.2 Dữ liệu doanh số chi tiết

Ngày 1 tháng 6 (kỷ niệm Ngày lễ Thiếu nhi và đợt khuyến mãi hè), ngày 30 tháng 8 (bắt đầu năm học mới) và ngày 1 tháng 4 (khuyến mãi xuân) hàng năm đều ghi nhận mức doanh thu đạt kỷ lục.

Thương hiệu Maximus liên tục đạt doanh số bán hàng cao nhất qua các năm, vượt xa các thương hiệu khác. Theo sau là Natura và Aliqui, với mức doanh số đạt được chỉ dưới 50% so với Maximus. Xem quy mô doanh số nhãn hàng tại bảng 7.

Về Category, hình 8, doanh số của "Urban" vượt trội hẳn so với các loại khác.

Về Segment, hình 9, các sản phẩm có segment là "Convience", "Moderation", "Extreme" và "Productivity" có doanh số đứng đầu.

2 Phương pháp và mô hình dự báo.

2.1 Các mô hình dự báo đã lựa chọn.

Bài trình bày sẽ sử dụng 4 mô hình dự đoán với độ chính xác (Accuracy) tăng dần:

1. ARIMA
2. RNN
3. LSTM
4. XGBoost

Các chỉ số đánh giá gồm có:
R-Square (R^2), Mean Absolute Percentage Error **(MAPE)** và **RMSE**.

2.2 Xây dựng mô hình ARIMA.

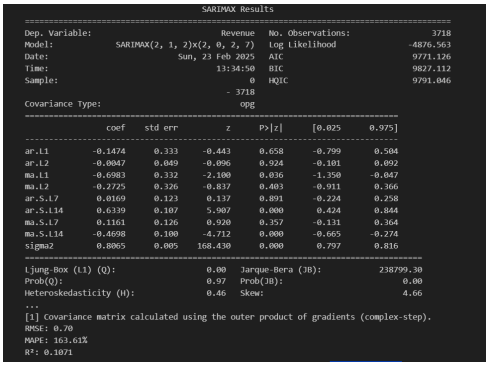
2.2.1 Xác định tham số p, d, q và kiểm tra Seasonality

- Sử dụng phương trình Augmented Dickey-Fuller (ADFuller) để dự đoán chỉ số d
- Vẽ AutoCorrelation (ACF) và Partial Auto-Correlation (PACF) để tính p và q
- Sử dụng auto_arima để xác định các tham số p, d, q cho ARIMA
- Sử dụng seasonal_compose để kiểm tra tính mùa vụ và auto_arima để tối ưu lại các tham số P, D, Q, m cho SARIMA

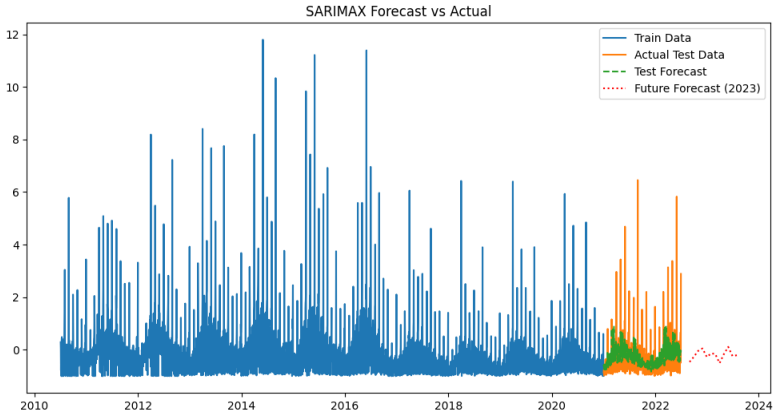
Kết quả tune parameter (tạm thời):

$$(p, d, q, P, D, Q, m) = (2, 1, 2, 2, 0, 2, 7)$$

2.2.2 Chạy Model và kiểm tra Performance



```
                               SARIMAX Results
==========================================================================================
Dep. Variable:                          Revenue   No. Observations:                 3718
Model:             SARIMAX(2, 1, 2)x(2, 0, 2, 7)   Log Likelihood               -4876.563
Date:                           Sun, 23 Feb 2025   AIC                           9771.126
Time:                                   13:34:50   BIC                           9827.112
Sample:                                        0   HQIC                          9791.046
                                            - 3718
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.1474      0.333     -0.443      0.658      -0.799       0.504
ar.L2         -0.0047      0.049     -0.096      0.924      -0.101       0.092
ma.L1         -0.6983      0.332     -2.100      0.036      -1.350      -0.047
ma.L2         -0.2725      0.326     -0.837      0.403      -0.911       0.366
ar.S.L7        0.0169      0.123      0.137      0.891      -0.224       0.258
ar.S.L14       0.6339      0.107      5.907      0.000       0.424       0.844
ma.S.L7        0.1161      0.126      0.920      0.357      -0.131       0.364
ma.S.L14      -0.4698      0.100     -4.712      0.000      -0.665      -0.274
sigma2         0.8065      0.005    168.430      0.000       0.797       0.816
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):            238799.30
Prob(Q):                              0.97   Prob(JB):                         0.00
Heteroskedasticity (H):               0.46   Skew:                             4.66
...
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
RMSE: 0.70
MAPE: 163.61%
R²: 0.1071
```



2.3 Mô hình RNN và LSTM.

2.3.1 Tiền xử lí và xác định tham số

- Tạo các biến để huấn luyện model RNN và LSTM (`day_of_week`, `day_of_month`, `month`).
- Chuyển đổi các biến ngày/tháng về dạng sin và cos để biểu diễn tính chu kỳ; đồng thời áp dụng hàm `log` cho biến `Revenue` nhằm giảm ảnh hưởng của các giá trị lớn.
- Sử dụng `MinMaxScaler` cho các biến đầu vào (`'Revenue_log'`, `'dow_sin'`, `'dow_cos'`, `'month_sin'`, `'month_cos'`, `'dom_sin'`, `'dom_cos'`); chia dữ liệu thành 2 phần: **training set** (trước năm 2021) và **test set** (trong năm 2021 và 2022).
- Thiết lập **'lookback window' = 30 ngày)** để tạo mẫu cho mô hình học chuỗi thời gian, đảm bảo mô hình có đủ thông tin về thời gian trong quá khứ để dự đoán tương lai.

2.3.2 Huấn luyện mô hình và đánh giá

- Hai mô hình được xây dựng dựa trên thư viện Keras.
- Mô hình RNN gồm 300 đơn vị, hàm kích hoạt: ReLU, hàm tối ưu: Adam, tối ưu theo MSE.
- Mô hình LSTM gồm 300 đơn vị, hàm kích hoạt: ReLU, hàm tối ưu: Adam, tối ưu theo MSE.
- Cả 2 mô hình đều được huấn luyện trên **training set**, epochs = 38, batch_size = 20.
- Kết quả của cả 2 mô hình được thể hiện ở Bảng 1. Trong đó LSTM vượt trội về chỉ số R2, cho thấy mô hình này giải thích sự biến thiên của biến phụ thuộc ('Revenue') tốt hơn so với RNN. Các chỉ số MAPE, RMSE đều thấp cho thấy các sai số dự báo nhỏ và dao động của các sai số này cũng ít, từ đó thể hiện độ ổn định của mô hình.

2.3.3 Dự đoán 12 tháng tiếp bằng LSTM

Sau khi huấn luyện mô hình trên dữ liệu từ năm 2010 đến 2020, dữ liệu từ năm 2021–2022 được sử dụng để kiểm tra và tinh chỉnh các siêu tham số, cụ thể được mô tả trong hình 10. Sau khi đạt được độ chính xác tối ưu, mô hình được huấn luyện lại 11 trên toàn bộ dữ liệu từ 2010–2022 để dự báo cho năm 2023.

2.4 Mô hình XGBoost và Random Forest

2.4.1 Tiền xử lý

Tiền xử lý dữ liệu cho XGBoost và Random Forest gồm: chuyển đổi cột 'Date' sang dạng datetime và tạo các đặc trưng thời gian ('Year', 'Month', 'Quarter', 'DayOfWeek'); tạo biến đặc trưng quan trọng như 'ProfitMargin', 'RevenuePerUnit' và các đặc trưng lag để tránh rò rỉ dữ liệu; mã hóa biến phân loại bằng LabelEncoder; chọn các biến quan trọng qua ma trận tương quan; và chuẩn hóa dữ liệu bằng RobustScaler để giảm ảnh hưởng của ngoại lai.

2.4.2 Xác định tham số cho mô hình mô hình XGBoost

XGBoost

$$\mathbf{XG} = \begin{bmatrix} \text{n\_estimators} & 50 \\ \text{learning\_rate} & 0.07 \\ \text{max\_depth} & 1 \\ \text{subsample} & 0.6 \\ \text{colsample\_bytree} & 0.6 \\ \text{min\_child\_weight} & 15 \\ \text{gamma} & 1.0 \\ \text{reg\_lambda} & 2.0 \\ \text{reg\_alpha} & 0.5 \\ \text{random\_state} & 42 \end{bmatrix}$$

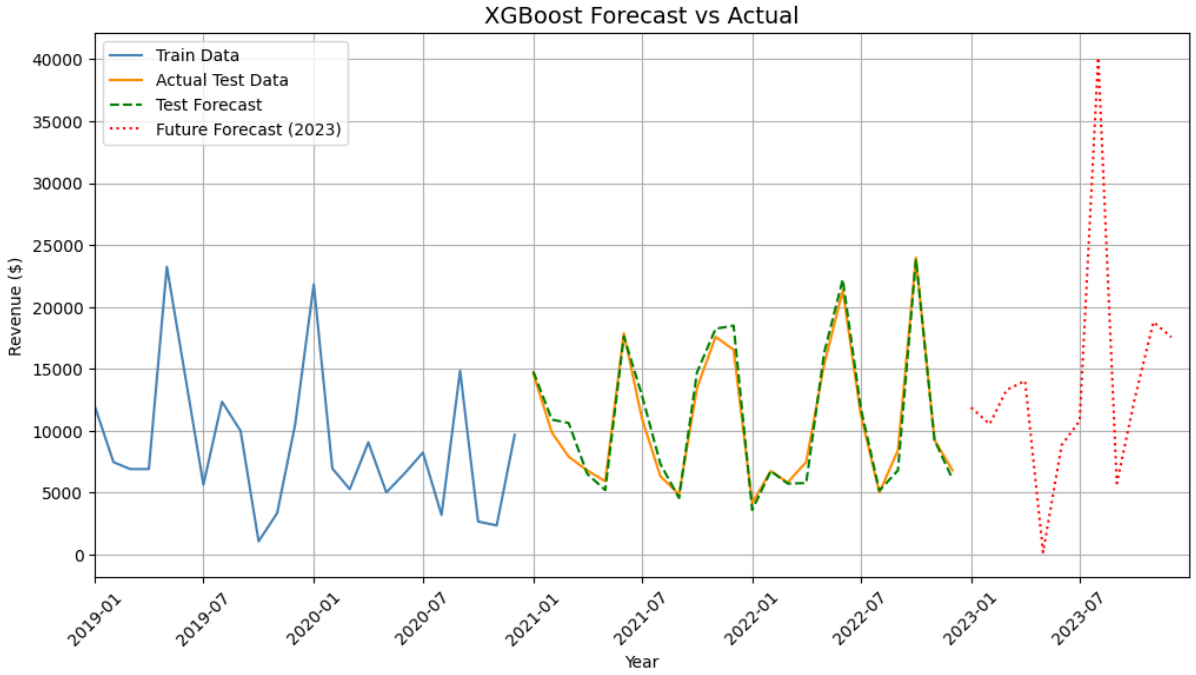
Chú thích:

- Số lượng cây (n_estimators = 50)
- Tốc độ học (learning_rate = 0.07)
- Độ sâu của cây (max_depth = 1)
- Tỷ lệ lấy mẫu hàng (subsample = 0.6) – Thêm ngẫu nhiên khi lấy mẫu hàng, giúp mô hình tổng quát tốt hơn.
- Tỷ lệ lấy mẫu đặc trưng (colsample_bytree = 0.6) – Chọn đặc trưng ngẫu nhiên trong mỗi cây để giảm tương quan.
- Trọng số tối thiểu để chia nhánh (min_child_weight = 15) – Hạn chế chia nhánh dựa trên biến động nhỏ để tránh overfitting.
- Ngưỡng tạo nhánh mới (gamma = 1.0) – Yêu cầu mức lợi ích cao hơn khi tạo nhánh mới.

2.4.3 Chạy và đánh giá Performance

Huấn luyện mô hình XGBoost với các tham số đã được tối ưu. 1

Đánh giá kết quả: XGBoost đạt độ chính xác cao hơn so với ARIMA và LSTM do khả năng khai thác tốt các đặc trưng phi tuyến tính trong dữ liệu.



Hình 1: Training and Testing by XGBoost.

3 Kết quả, đánh giá và chiến lược .

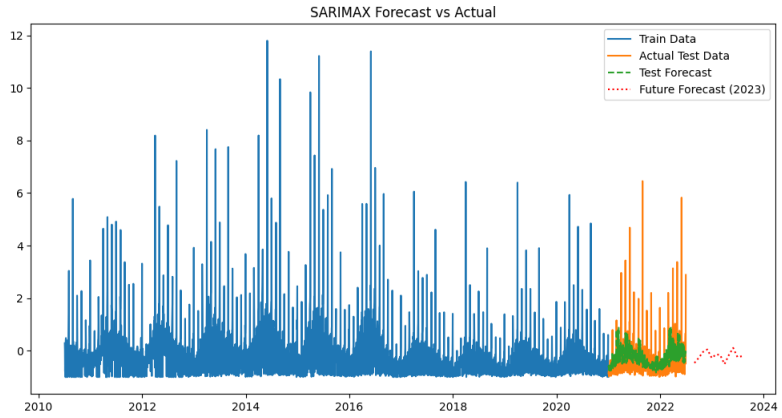
3.1 So sánh hiệu suất các mô hình đã chọn.

Mô hình	Chỉ số đánh giá			Vấn đề phát sinh	Hướng phát triển
	R2	MAPE	RMSE		
ARIMA	0.1071	163.61%	0.7	Nhạy cảm với dữ liệu lớn và chưa tối ưu hết được parameter do quá tốn dung lượng chạy	Tìm cách tune lại với m lớn hơn, tìm thêm các biến độc lập quan trọng
RNN	0.4787	21.20%	0.1236	Biến động của giá trị dự đoán thường giảm dần về 0. (Vanishing gradient)	Điều chỉnh lại số lượng neuron, hàm kích hoạt, hoặc thêm lớp RNN.
LSTM	0.5018	21.92%	0.1208	Thời gian huấn luyện mô hình lâu (>10 phút).	Điều chỉnh kiến trúc (nhiều tầng LSTM, Bi-LSTM) và giảm số lượng đơn vị (<300), thử các mô hình lai (khái quát tốt hơn và độ phức tạp giảm).
XGBoost	0.9268	14.19	0.1048	Cần điều chỉnh các siêu tham số nhiều lần để tránh hiện tượng overfitting, do XGBoost không tự động nắm bắt quan hệ thời gian. Do đó, cần xây dựng các đặc trưng thủ công (Lag Features). Ngoài ra, XGBoost không tối ưu cho dự báo dài hạn và rất nhạy cảm với các siêu tham số.	Tối ưu các tham số (learning rate, max depth), áp dụng lựa chọn đặc trưng (feature selection) để giảm nhiễu, và kết hợp với các mô hình khác nhằm xây dựng mô hình lai (hybrid model) giúp cải thiện độ chính xác dự báo.

Bảng 1: Bảng đánh giá mô hình

3.2 Đánh giá và dự đoán doanh số 2023

Cùng với xu hướng giảm của doanh số các năm trước đó thì năm 2023 được dự báo sẽ tiếp tục chuỗi giảm của thu nhập, ước tính rơi vào khoảng:



Month	Predicted Revenue
1	435641.557910
2	498564.981042
3	572076.255626
4	601715.841162
5	496638.456666
6	540848.192542
7	533724.080097
8	421735.676016
9	530639.008547
10	620318.290729
11	508112.582527
12	518004.285986)

3.3 Đưa ra chiến lược đầu tư sắp tới

- Đẩy mạnh đầu tư vào các bang Texas, Washington, Florida và khu vực Đông Nam (các thành phố ghi nhận doanh thu cao và ổn định trong dữ liệu lịch sử), trong đó kích thích sản xuất ra mắt thêm các sản phẩm đang bán chạy trong quá khứ.
- Tiếp tục đầu tư vào các nhãn hàng đang chiếm top đầu doanh thu Maximus, Natura, Aliqui cùng các sản phẩm đã bán chạy cũng như các sản phẩm mới phù hợp với thị hiếu khách hàng.
- Phát triển, tăng cường bán hàng qua nền tảng số online và phát triển thêm cách chiến dịch Marketing để phổ biến công ty.
- Tìm cách xử lý với các nguyên nhân làm ảnh hưởng doanh số các năm đươc ghi nhận (Lạm phát, Covid-19, xuất hiện các xu hướng mới, giảm nhu cầu chi tiêu vào quần áo) và có chiến lược chuẩn bị và đối phó.

Tài liệu

[1] Neusser, K. *Time Series Econometrics*. Springer, 2016.

[2] Hilmer, C. E., and Hilmer, M. J. *Practical Econometrics: Data Collection, Analysis, and Application*. McGraw-Hill, 2014.

[3] Auffarth, B. *Machine Learning for Time-Series with Python: Forecast, Predict, and Detect Anomalies with State-of-the-Art Machine Learning Methods*. Packt Publishing, 2021.

[4] YouTube Video: Forecasting with Machine Learning.

[5] YouTube Video: Time Series Analysis Tutorial.

[6] YouTube Playlist: Time Series and Forecasting Techniques.

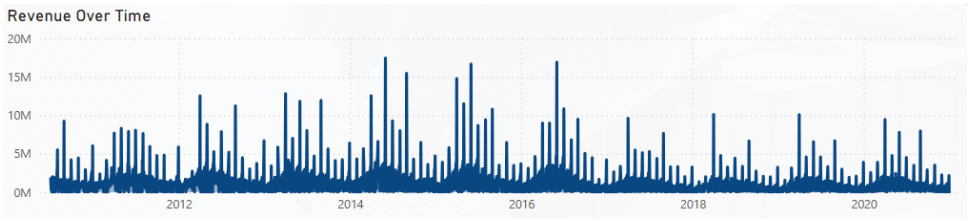
[7] Brownlee, J. "XGBoost for Time Series Forecasting." *Machine Learning Mastery*, 2024. Available at: https://machinelearningmastery.com/xgboost-for-time-series-forecasting/.

[8] Analytics Vidhya. "XGBoost for Time Series Forecasting." *Analytics Vidhya*, 2024. Available at: https://www.analyticsvidhya.com/blog/2024/01/xgboost-for-time-series-forecasting/.

[9] GeeksforGeeks. "XGBoost Overview." *GeeksforGeeks*, 2024. Available at: https://www.geeksforgeeks.org/xgboost/.

Appendices

A List of EDA



Hình 2: Doanh số qua thời gian.

```
city_counts = df['City'].value_counts()

# Find the number of cities that appear only once
unique_city_count = (city_counts < 5).sum()
print(f'City appears less than 5 times: {unique_city_count} cities')
```

✓ 0.1s

```
City appears less than 5 times: 4127 cities
```

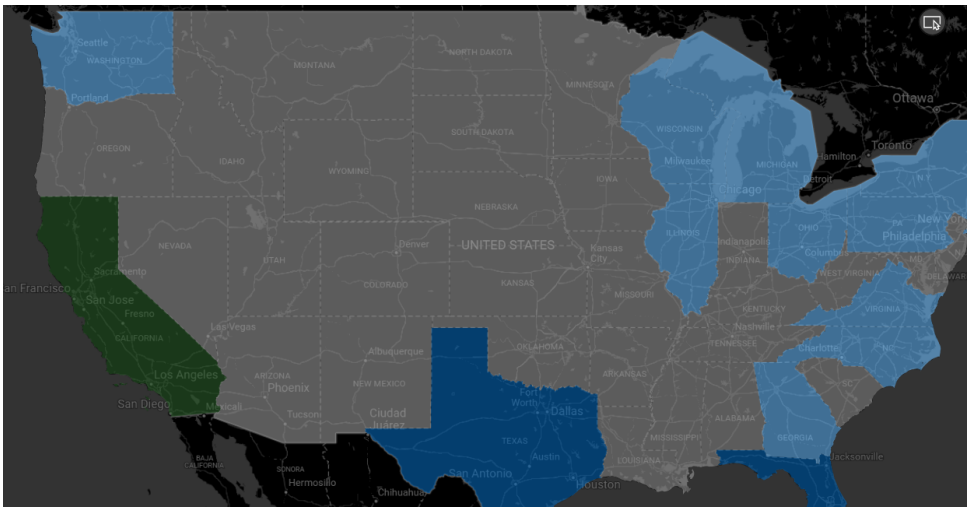
Hình 3: Tổng số thành phố ghi nhận doanh thu dưới 5 ngày.

```
city_counts = df['City'].value_counts()

# Get cities that appear < 5 times
cities_less_than_5 = city_counts[city_counts < 5].index

# Filter the original DataFrame
filtered_df = df[df['City'].isin(cities_less_than_5)]

# Count distinct cities per year
yearly_city_count = filtered_df.groupby('Year')['City'].nunique().reset_index()


# Sort the result
yearly_city_count = yearly_city_count.sort_values(by='Year')

print(yearly_city_count)
```

✓ 0.2s

	Year	City
0	2010	329
1	2011	754
2	2012	793
3	2013	833
4	2014	877
5	2015	824
6	2016	761
7	2017	444
8	2018	398
9	2019	375
10	2020	421
11	2021	379
12	2022	241

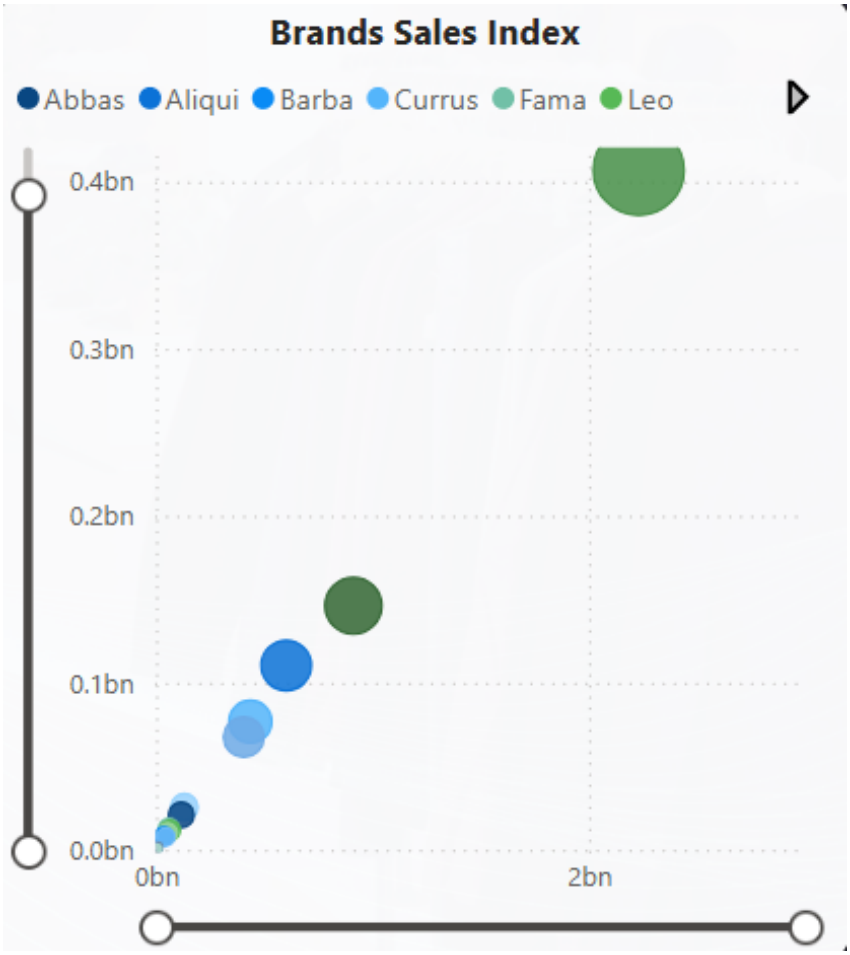
Hình 4: Số thành phố ghi nhận doanh thu dưới 5 ngày mỗi năm



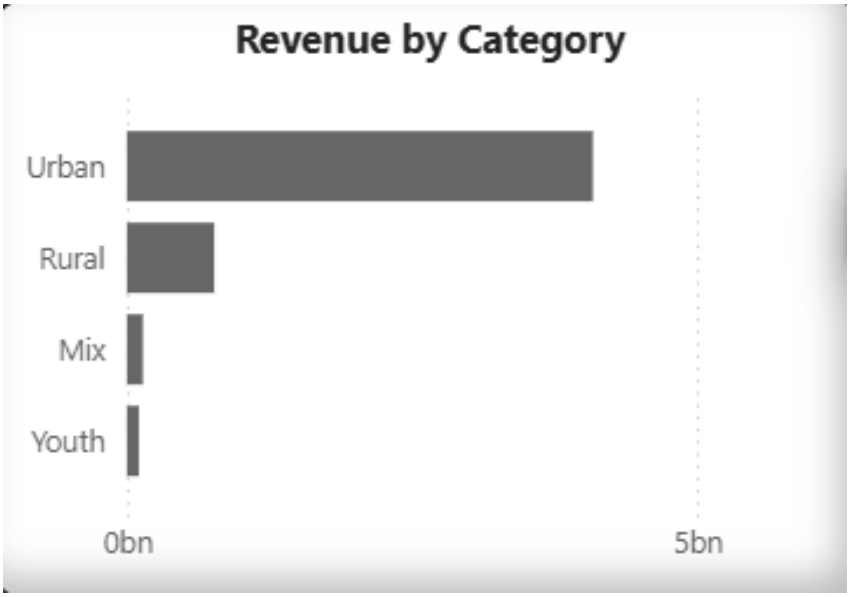
Hình 5: Phân bố khu vực phân phối.

City	Revenue	Units Sold	Profit
Miami	26,472,570.39	4690	4157662.46
Houston	25,115,716.08	4190	4354357.80
Las Vegas	24,190,889.94	4268	4300447.70
San Diego	21,560,166.81	3992	3912525.64
Jacksonville	21,108,853.71	3425	3664535.67

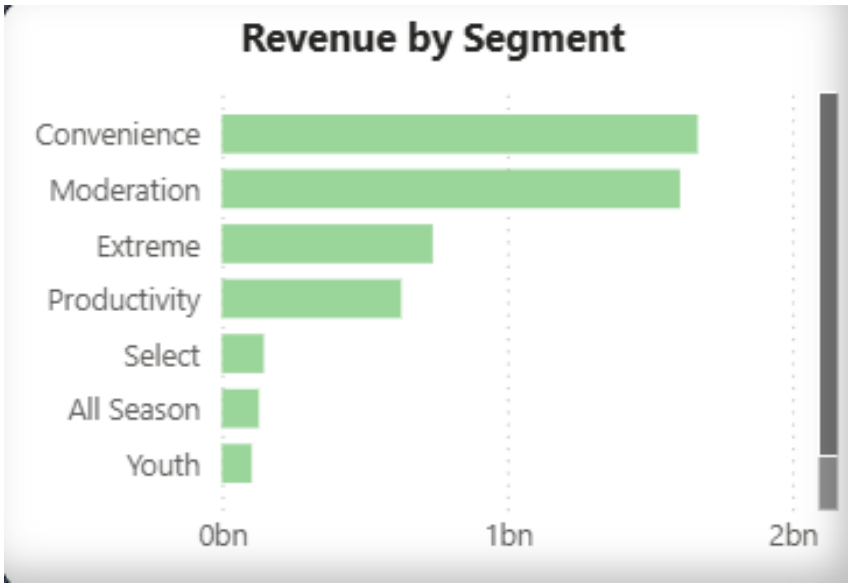
Hình 6: Top 5 thành phố chiếm tỷ trọng cao



Hình 7: Quy mô doanh số của các nhãn hàng

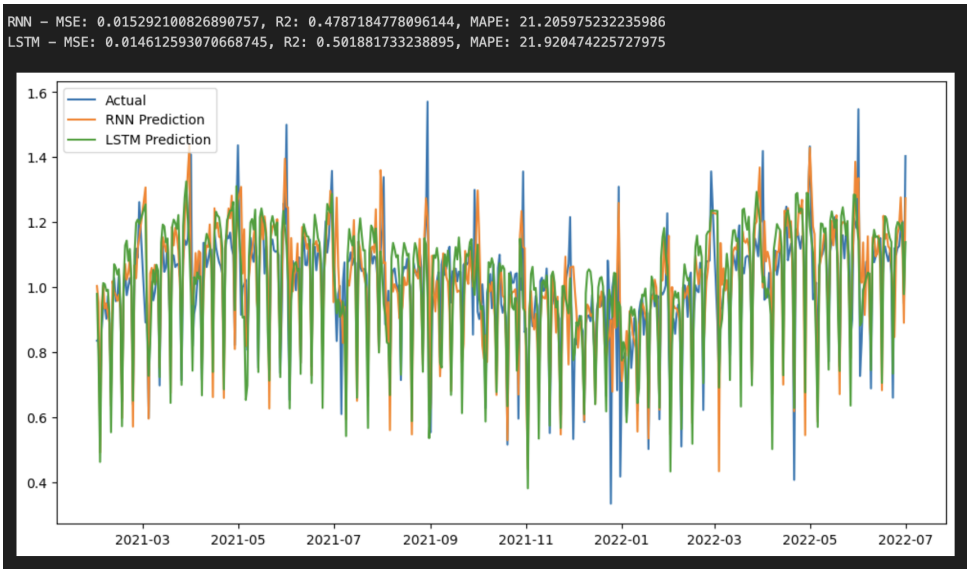


Hình 8: Doanh số theo Category

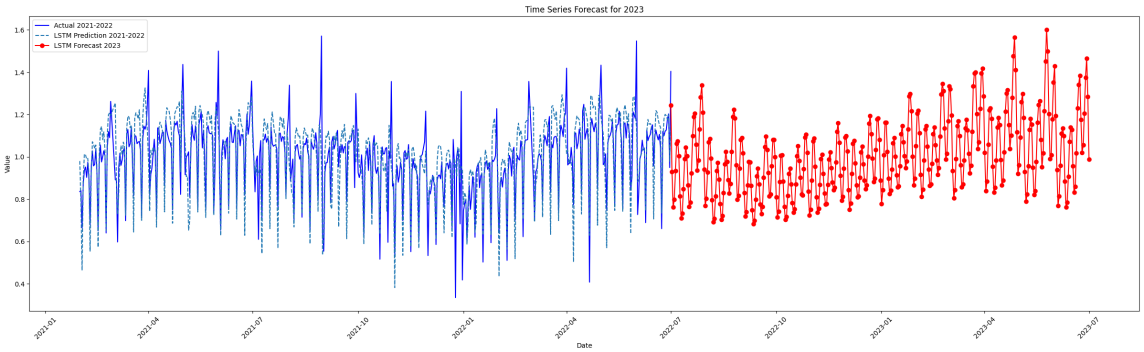


Hình 9: Doanh số theo Segment

B Model



Hình 10: Training and Testing Data



Hình 11: Prediction of 2023 by LSTM