ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

ĐẠI HỌC KINH TẾ - LUẬT

ITB CLUB

BUSINESS INTELLIGENCE 6

BÁO CÁO VÒNG 3

KPIM DATATHON - PHÂN TÍCH THỐNG KÊ

Nhóm dự thi : The Arrow

ID : 034

Thành Phố Hồ Chí Minh, ngày 15 tháng 7 năm 2022

DANH SÁCH THÀNH VIÊN

STT	Họ và tên	Trường	Vai trò
1	Nguyễn Thảo Ngân	Đại học Kinh tế - Luật	Nhóm trưởng
2	Nguyễn Đỗ Thanh Thùy	Đại học Kinh tế - Luật	Thành viên
3	Nguyễn Thu Vân	Đại học Kinh tế - Luật	Thành viên

MỤC LỤC

1. Giới thiệu công ty KPIM

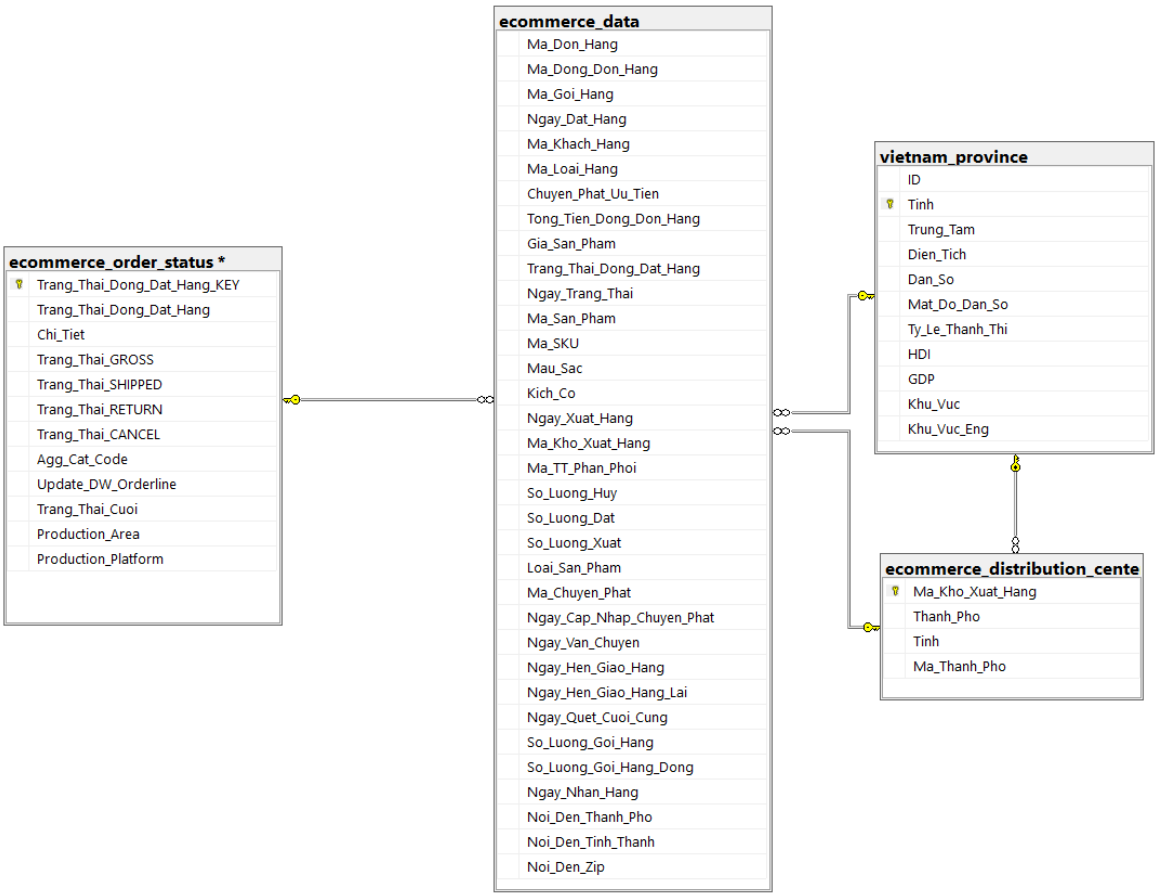
KPIM Ecommerce là một công ty thương mại điện tử cung cấp cho khách hàng nhiều loại sản phẩm đa dạng. Thị trường của KPIM Ecommerce trải rộng khắp các tỉnh thành toàn quốc. Công ty vận chuyển hàng triệu đơn hàng mỗi năm để nâng cao chất lượng đời sống của khách hàng. Tiêu chí số một của KPIM Ecommerce là luôn luôn đáp ứng khách hàng vượt trên sự mong đợi của họ và một trong những tiêu chí cần đáp ứng đó là tốc độ giao hàng.

2. Đồng bộ dữ liệu

Từ những bộ dữ liệu được ban tổ chức cung cấp, nhóm đã xác định được những dữ liệu cần thiết truy vấn phục vụ cho mục đích phân tích như sau:

- Dữ liệu đơn hàng trong Excel
- Dữ liệu đơn hàng trong SQL Server
- Dữ liệu kho hàng
- Dữ liệu trạng thái đơn hàng
- Dữ liệu các tỉnh thành Việt Nam

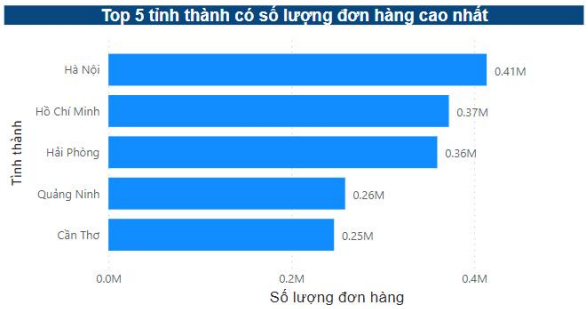
Sau khi đã xác định được các dữ liệu cần thiết, nhóm đã tiến hành xử lý dữ liệu và đồng bộ các dữ liệu từ các nguồn khác nhau thành 4 bảng: ecommerce_data, ecommerce_order_status, ecommerce_distribution_center và vietnam_province. Dựa trên các bảng nhóm thực hiện tạo khóa chính và tạo mối liên kết giữa các bảng như hình bên dưới.

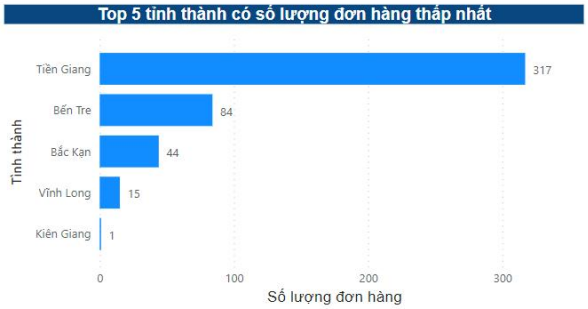


3. Phân tích dữ liệu

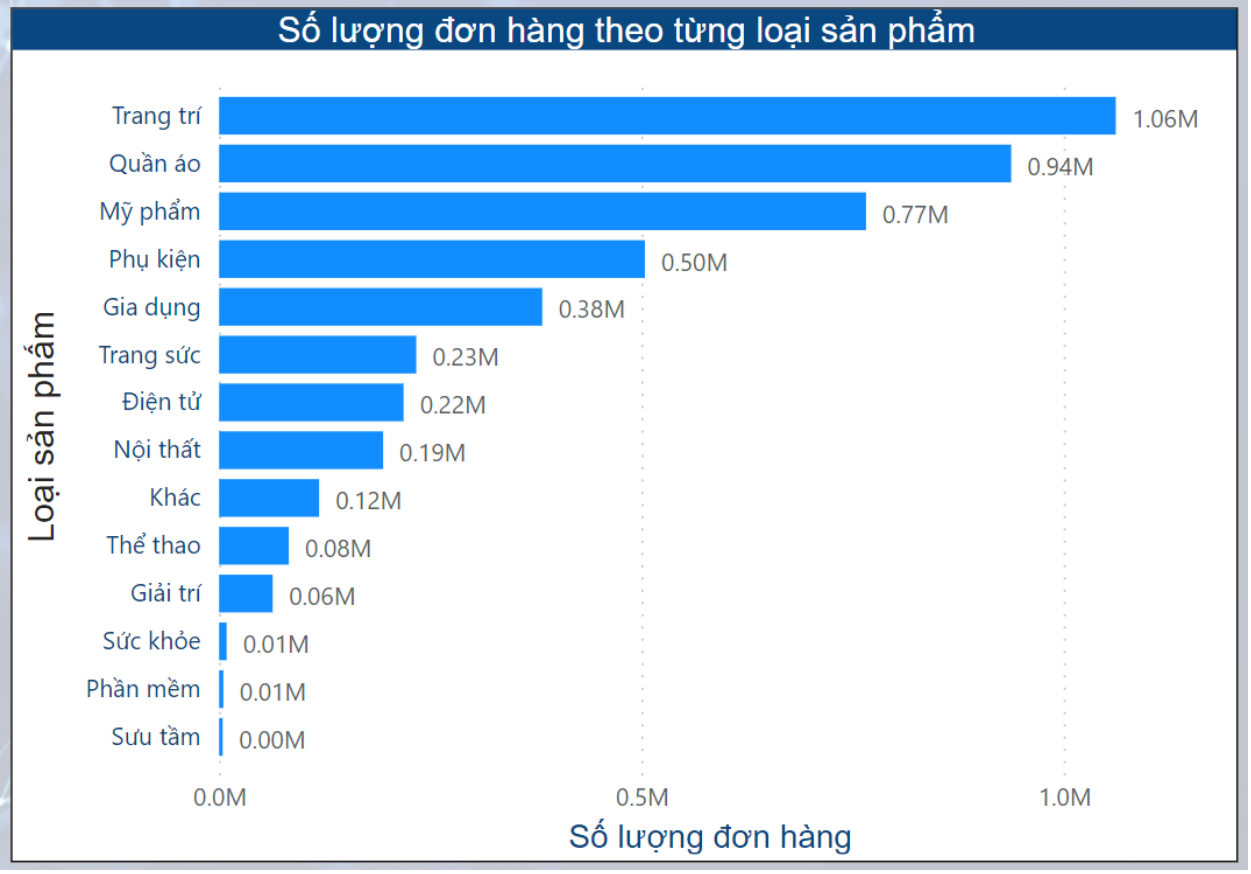
3.1 Phân tích đơn hàng của công ty

Trước tiên để xem được tổng quan về sự phân bố của đơn hàng, nhóm trực quan hóa dữ liệu về đơn hàng như sau:





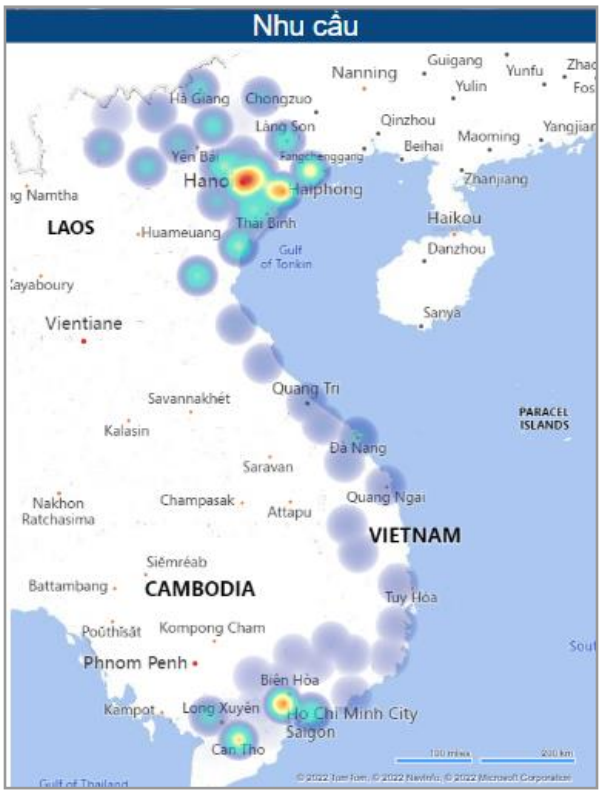
Dựa vào biểu đồ về sự phân bố cho thấy rằng khách hàng của KPIM phân bố đều khắp cả nước. Biểu đồ cột hiển thị rằng số lượng lớn khách hàng tập trung ở các thành phố lớn có nền kinh tế phát triển của Việt Nam như Hà Nội: 413731 đơn hàng, Hồ Chí Minh: 372336 đơn hàng và Hải Phòng: 359690 đơn hàng. Nhìn chung các tỉnh thành có số lượng đơn hàng thấp tập trung chủ yếu ở khu vực Đồng bằng sông Cửu Long như Tiền Giang: 317 đơn hàng, Bến Tre: 84 đơn hàng và Kiên Giang: 1 đơn hàng.

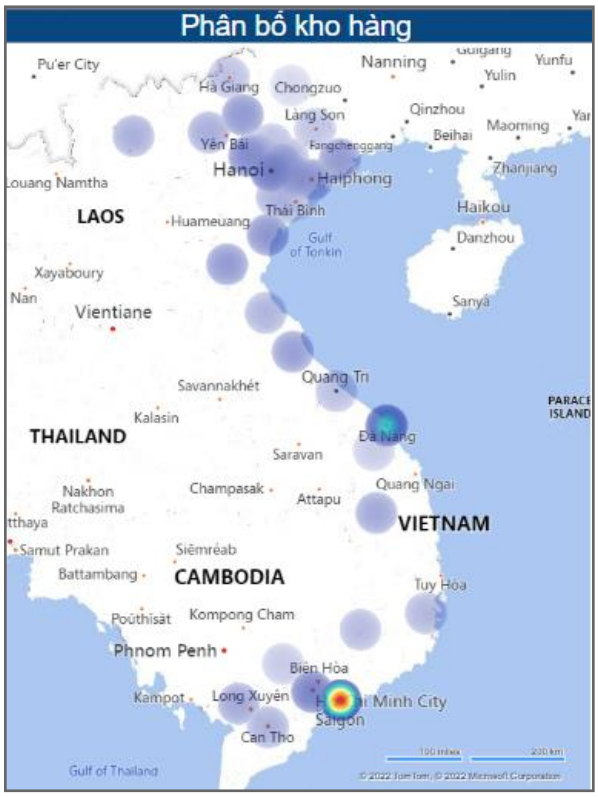


Biểu đồ cho thấy các sản phẩm được đặt hàng nhiều nhất của công ty KPIM lần lượt là: trang trí (23.18%), quần áo (20.47%) và mỹ phẩm (16.72%). Các sản phẩm thuộc loại sản phẩm sưu tầm, phần mềm, sức khỏe được đặt ít nhất. Dựa vào biểu đồ cột cho thấy sự phân bố của loại sản phẩm trang trí, quần áo và mỹ phẩm tập trung số lượng lớn ở các thành phố như Hà Nội, Hả Phòng, Hồ Chí Mình. Bên cạnh đó, nhìn chung sự phân bố của các gói theo các loại sản phẩm còn lại khá đồng đều giữa các tỉnh thành và có sự chênh lệch lớn hơn không đáng kể ở các thành phố lớn như Hà Nội, Hải Phòng và Hồ Chí Minh.

Số đơn hàng trễ ở tháng 7 cao (10.3K) và số lượng đơn đặt hàng ở tháng 8 và 9 có sự giảm. Số lượng đơn hàng trễ ở tháng 8 và 9 thấp (6.6 K và 6.9K) dẫn đến đơn hàng ở tháng 10, 11 tăng (0.84M và 0.88M). Tuy nhiên tại tháng 8 và 9 số lượng đơn hàng trễ tăng cao (13.3K và 13.9K) là nguyên nhân làm cho số lượng đơn hàng ở tháng 12 giảm (0.82M). Từ đó có thể thấy mức độ tăng trưởng của các đơn hàng ở tháng sau có mối quan hệ tỉ lệ nghịch với độ tăng của các đơn vận chuyển trễ của tháng liền trước.

3.2 Phân tích mạng lưới phân phối





Nhìn chung, số lượng kho hàng được phân bổ tương xứng với nhu cầu, trải đều khắp các tỉnh thành, và tập trung chủ yếu ở 2 vùng có nhu cầu lớn nhất là ở vùng Đông Nam Bộ và vùng Đồng Bằng Sông Hồng.

Tuy nhiên khi phân tích kỹ hơn, nhóm phát hiện một số vấn đề như sau:

- Một số kho mặc dù có tồn tại nhưng không hoạt động
- Các loại hàng hóa trong kho chưa được phân bổ đồng đều, mặc dù nhu cầu cho các loại mặt hàng ở các tỉnh thành là như nhau
- Còn tồn tại vấn đề giao hàng trễ, đặc biệt là những đơn hàng giao nội thành

4. Dự báo đơn hàng mới, tối ưu hóa vị trí kho hàng và đường vận chuyển hàng hóa.

4.1 Dự báo đơn hàng mới

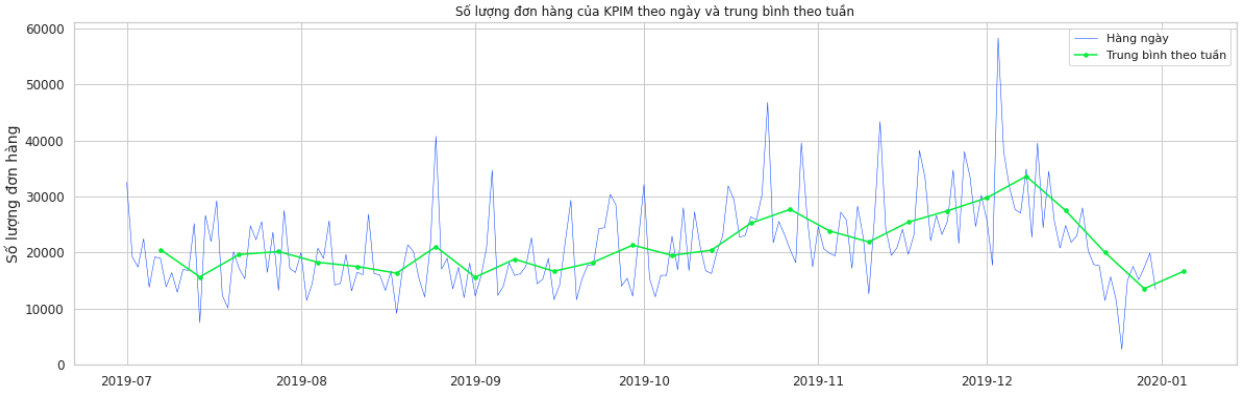
4.1.1 Thu thập dữ liệu

Từ dữ liệu ban đầu, nhóm thực hiện chọn ra dữ liệu phù hợp để thực hiện dự báo đơn hàng trong tương lai. Dữ liệu được chọn là dữ liệu gồm các đơn hàng đã nhận hàng thành công và loại bỏ các đơn hàng bị hủy hoặc hoàn lại - loại bỏ các đơn hàng có trạng thái giao hàng là: BX, CP, CX và RT. Tiếp theo đó tính tổng số đơn hàng theo ngày của công ty. Sau khi thực hiện nhóm được kết quả như sau:

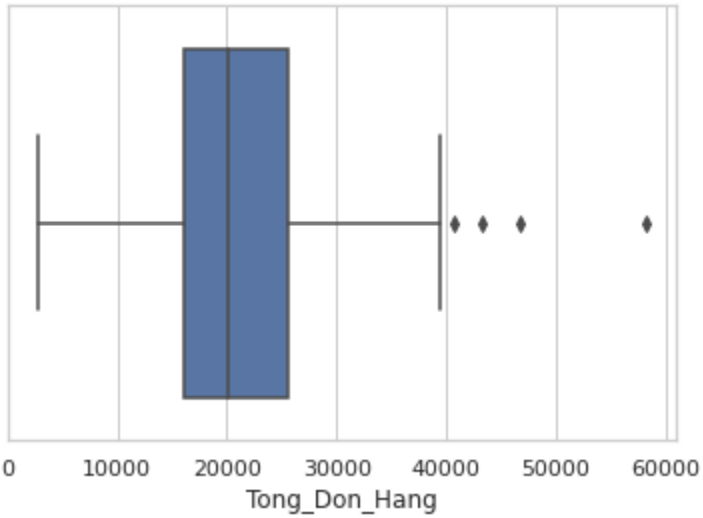
Ngay_Dat_Hang	Tong_Don_Hang
2019-07-01	32547
2019-07-02	19191
2019-07-03	17407
2019-07-04	22435
2019-07-05	13803

	Tong_Don_Hang
count	184.000000
mean	21315.157609
std	7765.139328
min	2744.000000
25%	16030.500000
50%	20041.000000
75%	25556.000000
max	58270.000000

Dữ liệu bao gồn đơn hàng của KPIM trong khoảng thời gian 6 tháng từ 01/07/2019 đến 31/12/2019.



4.1.2 Xác định các dữ liệu ngoại lai (Detect outliers)

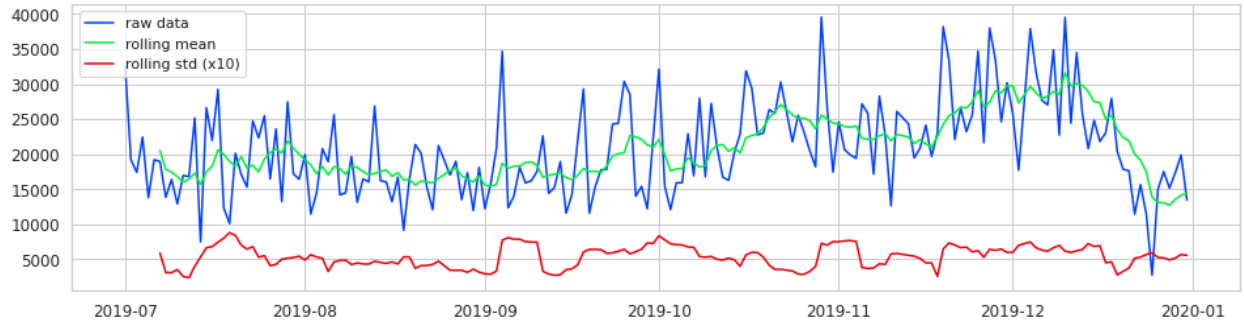


Hình trên cho thấy có 4 điểm dữ liệu nằm ngoài 2 đầu của boxplot, các điểm dữ liệu ngoài 40000 là outliers. Các outliers trong tập dữ liệu như sau:

Ngay_Dat_Hang	Tong_Don_Hang
2019-08-25	40710
2019-10-23	46776
2019-11-12	43351
2019-12-03	58270

4.1.3 Kiểm tra tính dừng của dữ liệu (Check for Stationarity)

Trực quan hóa dữ liệu (Visualization)



Vì giá trị trung bình và độ lệch chuẩn của dữ liệu có sự thay đổi lớn theo thời gian nên chưa thể kết luận tính dừng của dữ liệu bằng biểu đồ trên. Do đó, nhóm thực hiện một phương pháp khác để xác định tính dừng của dữ liệu.

Kiểm Định Dickey Fuller Bổ Sung (Augmented Dickey-Fuller Test)

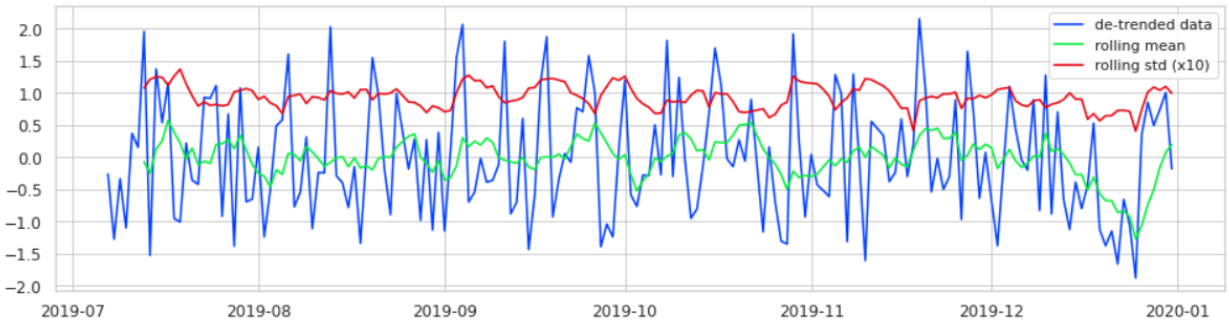
```
 > Is the raw data stationary ?
Test statistic = -1.716
P-value = 0.423
Critical values :
        1%: -3.468952197801766 - The data is not stationary with 99% confidence
        5%: -2.878495056473015 - The data is not stationary with 95% confidence
        10%: -2.57580913601947 - The data is not stationary with 90% confidence
```

Trực quan hóa dữ liệu và kiểm định ADF cho thấy dữ liệu không có tính dừng vì vậy bước tiếp theo cần chuyển sang dữ liệu có tính dừng.

4.1.4 Chuyển dữ liệu sang dữ liệu có tính dừng

Detrending

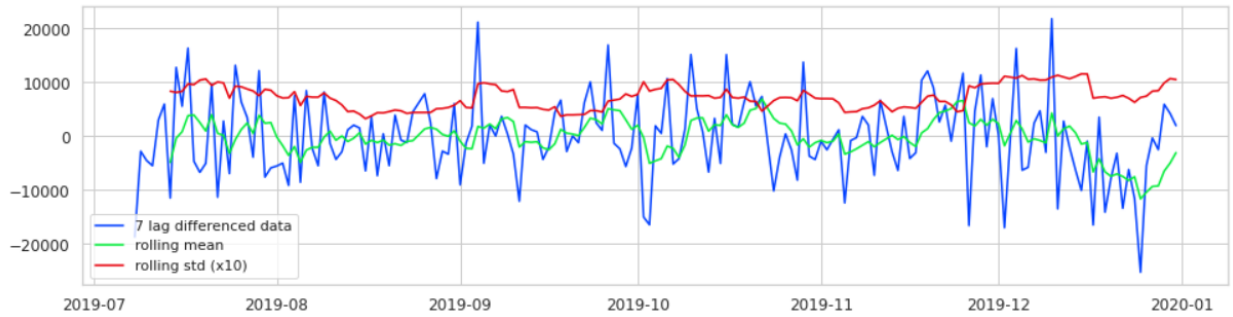
```
 > Is the de-trended data stationary ?
Test statistic = -12.848
P-value = 0.000
Critical values :
        1%: -3.4687256239864017 - The data is stationary with 99% confidence
        5%: -2.8783961376954363 - The data is stationary with 95% confidence
        10%: -2.575756341007O5 - The data is stationary with 90% confidence
```



Kết quả cho thấy dữ liệu có tính dừng, được biểu thị bằng kết quả kiểm định ADF.

Differencing

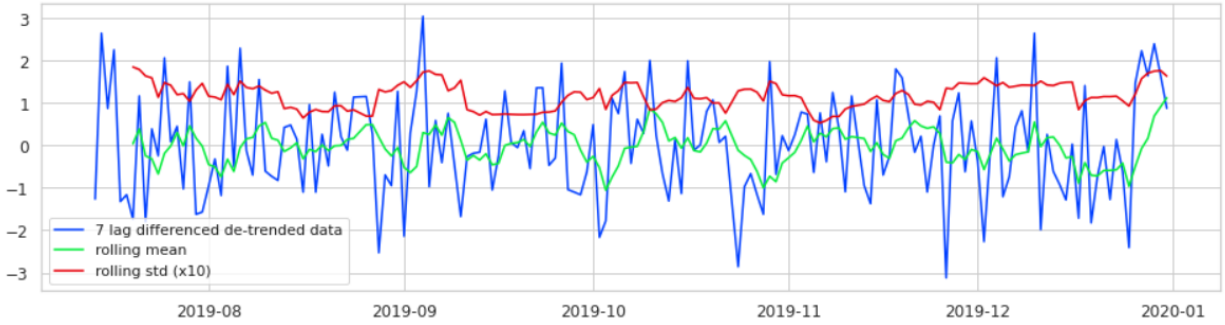
```
 > Is the 7 lag differenced data stationary ?
Test statistic = -4.054
P-value = 0.001
Critical values :
        1%: -3.472161410886292 - The data is stationary with 99% confidence
        5%: -2.8798954259680936 - The data is stationary with 95% confidence
        10%: -2.5765565828092245 - The data is stationary with 90% confidence
```



Kết quả cho thấy dữ liệu có tính dừng.

Kết hợp Detrending và Differencing

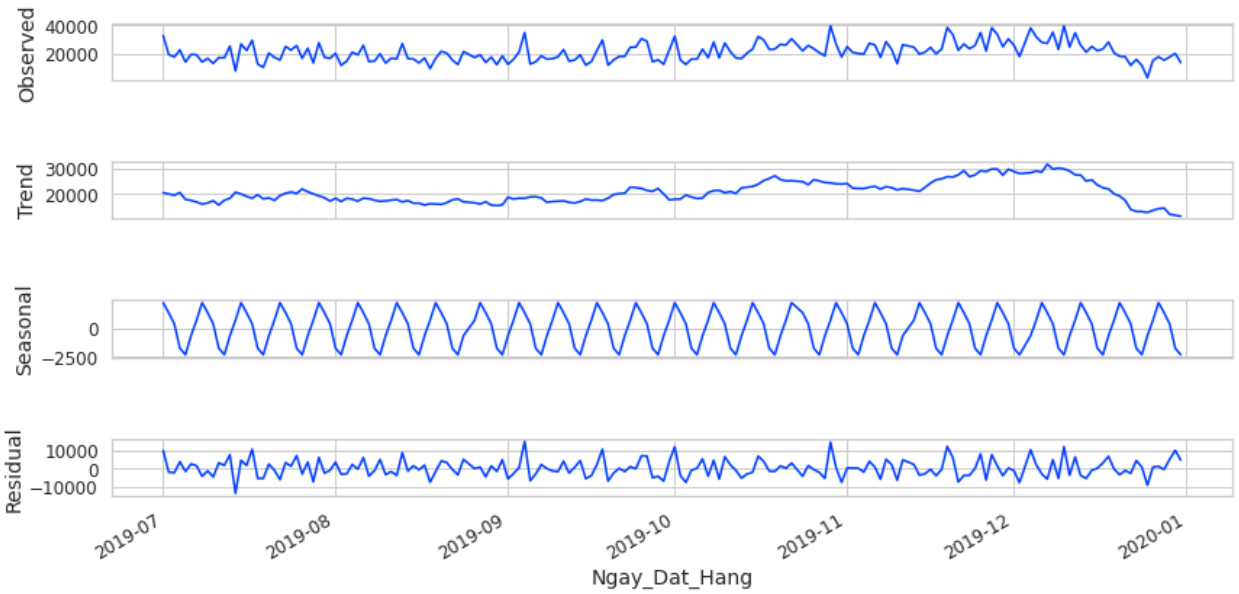
```
 > Is the 7 lag differenced de-trended data stationary ?
Test statistic = -5.714
P-value = 0.000
Critical values :
        1%: -3.473829775724492 - The data is stationary with 99% confidence
        5%: -2.880622899711496 - The data is stationary with 95% confidence
        10%: -2.5769448985432954 - The data is stationary with 90% confidence
```



Dựa vào biểu đồ về kết hợp của hai phương pháp cho thấy dữ liệu sau khi chuyển đổi có tính dừng. Nhóm sẽ sử dụng kết quả của quá trình này để tiếp tục thực hiện các bước tiếp theo.

4.1.5 Phân rã dữ liệu (Decomposing the data)

Biểu đồ dữ liệu đơn hàng ở trên có thể thấy xu hướng tăng sau đó giảm của dữ liệu nhưng hiển thị cụ thể về sự thay đổi theo mùa hoặc theo chu kỳ. Tiếp theo nhóm thực hiện phân rã dữ liệu để xem được các thành phần cấu thành nên chuỗi bao gồm: xu hướng (trend), mùa vụ (seasonal), phần dư (residual) như hình sau:



Các biểu đồ trên cho thấy tập dữ liệu có xu hướng tăng và đến gần giữa tháng 12 dữ liệu có xu hướng giảm và có tính mùa vụ hàng tuần. Tùy thuộc vào các thành phần các thành phần cấu thành nên chuỗi như: xu hướng, mùa vụ để lựa chọn mô hình phù hợp.

4.1.6 Xây dựng mô hình

Ở bước này, nhóm thực hiện bốn mô hình dự đoán: Simple Exponential Smoothing (SES), Holt, Seasonal Holt-Winters, và Seasonal ARIMA (SARIMA). Dựa trên kết quả đó để đánh giá các mô hình dự báo và tìm ra mô hình phù hợp cho tập dữ liệu.

Đầu tiên nhóm chia dữ liệu thành 2 tập: dữ liệu huấn luyện và dữ liệu kiểm tra:

+ Tập huấn luyện (Training dataset): dữ liệu đơn hàng từ 01/07/2019 đến 30/11/2019
+ Tập kiểm tra (Testing dataset): dữ liệu đơn hàng từ 01/12/2019 đến 31/12/2019

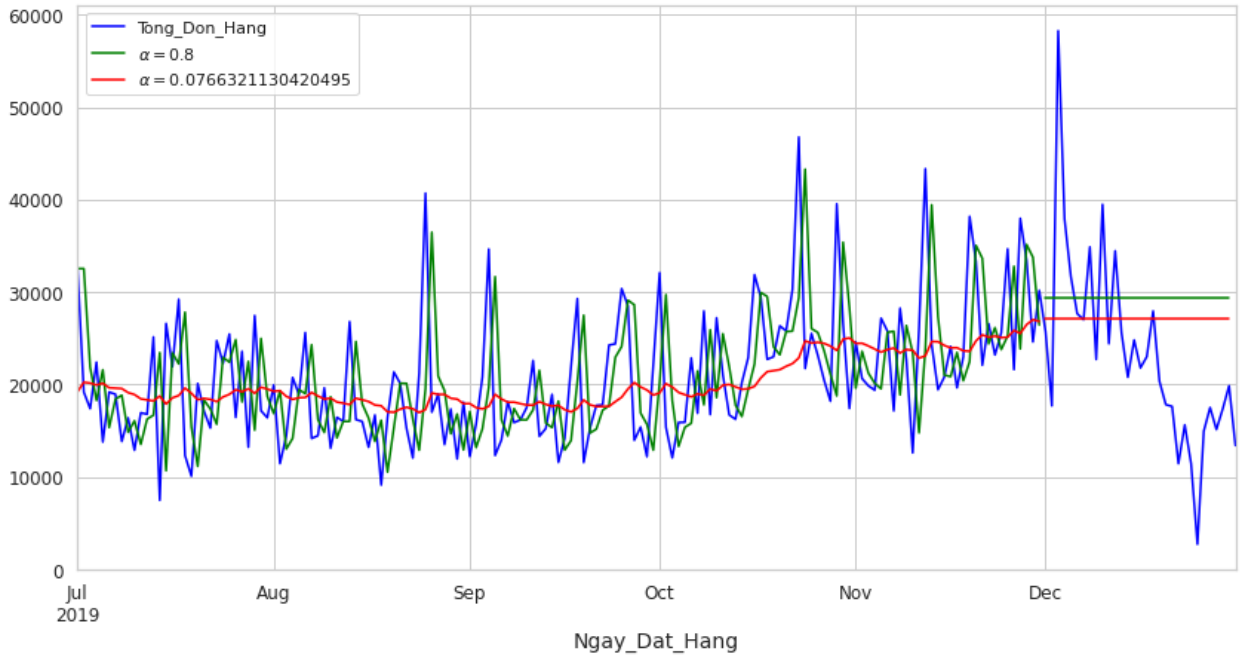
Tiếp theo nhóm tiến hàng xây dựng 4 model và đánh giá:

Simple Exponential Smoothing

Phương pháp này phù hợp với dữ liệu chuỗi thời gian không có xu hướng hoặc mùa vụ.

The Root Mean Squared Error of our forecasts with smoothing level of 0.8 is 11838.69

The Root Mean Squared Error of our forecasts with auto optimization is 10875.55



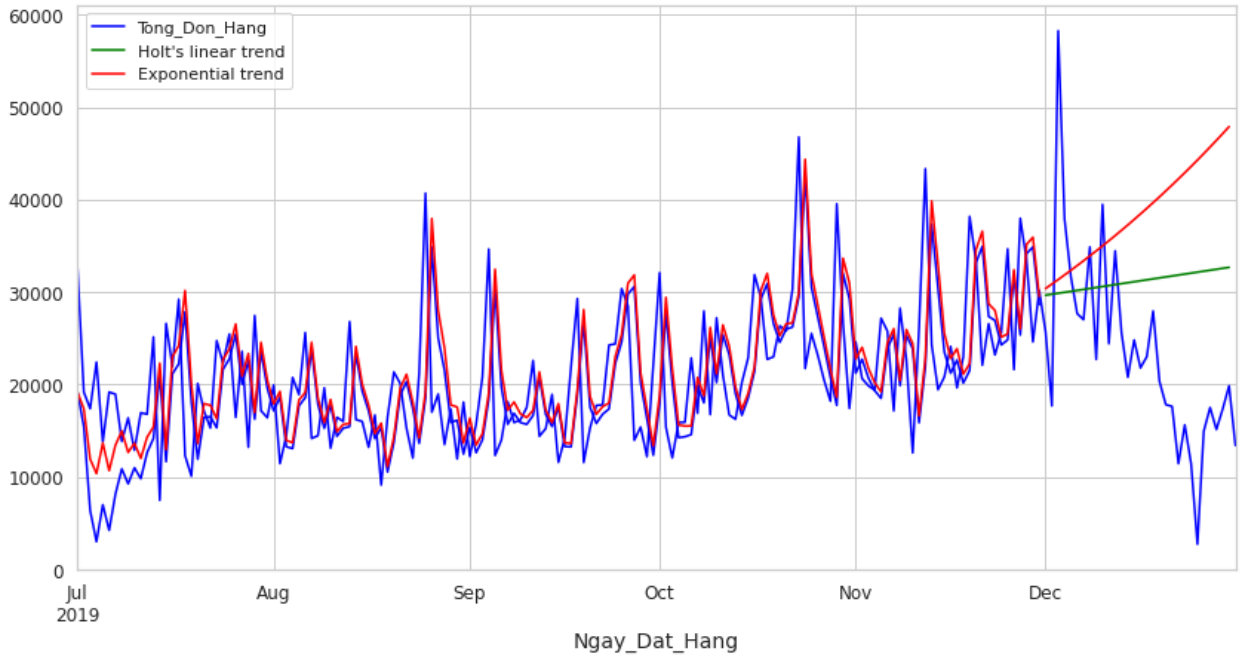
Kết quả của mô hình dự báo SES cho thấy sự khác biệt giữa α=0.8 (đường màu xanh lá) và α được tối ưu hóa tự động (đường màu đỏ). SES dự đoán kết quả là một đường thẳng vì nó sử dụng weighted averages. Từ đó cho thấy SES không dự đoán bất kỳ biến động nào. Vì hầu hết dữ liệu chuỗi thời gian đều tính xu hướng hoặc tính mùa vụ, nên mô hình này không thể được sử dụng để dự đoán trong tương lai.

Holt's Linear Trend Method

Phương pháp này phù hợp dữ liệu chuỗi thời gian có tính xu hướng nhưng không có mùa vụ.

The Root Mean Squared Error of Holts Linear trend 13297.78

The Root Mean Squared Error of Holts Exponential trend 20688.82



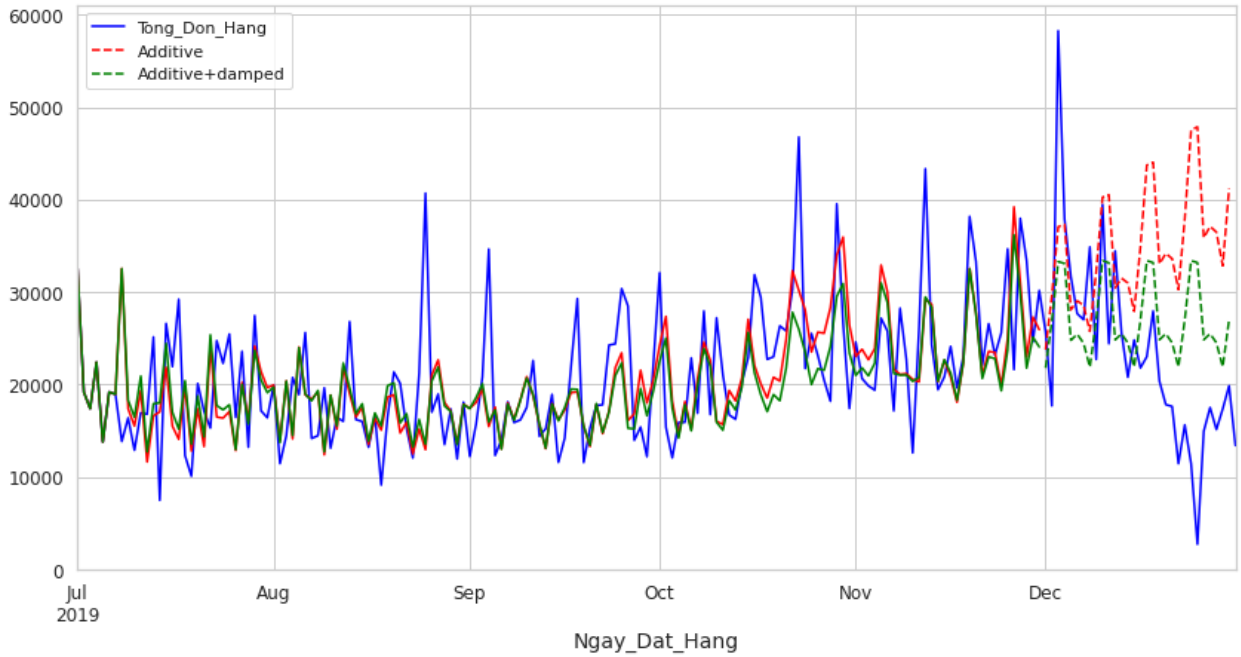
Kết quả của mô hình dự báo bằng phương pháp Holt, cho thấy sự so sánh giữa Holt's linear trend (đường màu xanh lá) và Exponential trend (đường màu đỏ) với nhau và với tổng đơn hàng. So với SES, Holt không dự đoán tốt được xu hướng của dữ liệu.

Holt-Winters' Seasonal Method

Phương pháp này phù hợp với dữ liệu chuỗi thời gian có tính xu hướng có hoặc không có tính mùa vụ.

The Root Mean Squared Error of additive trend, additive seasonal of period season_length=7 and a Box-Cox transformation 17237.96

The Root Mean Squared Error of additive damped trend, additive seasonal of period season_length=7 and a Box-Cox transformation 10766.63

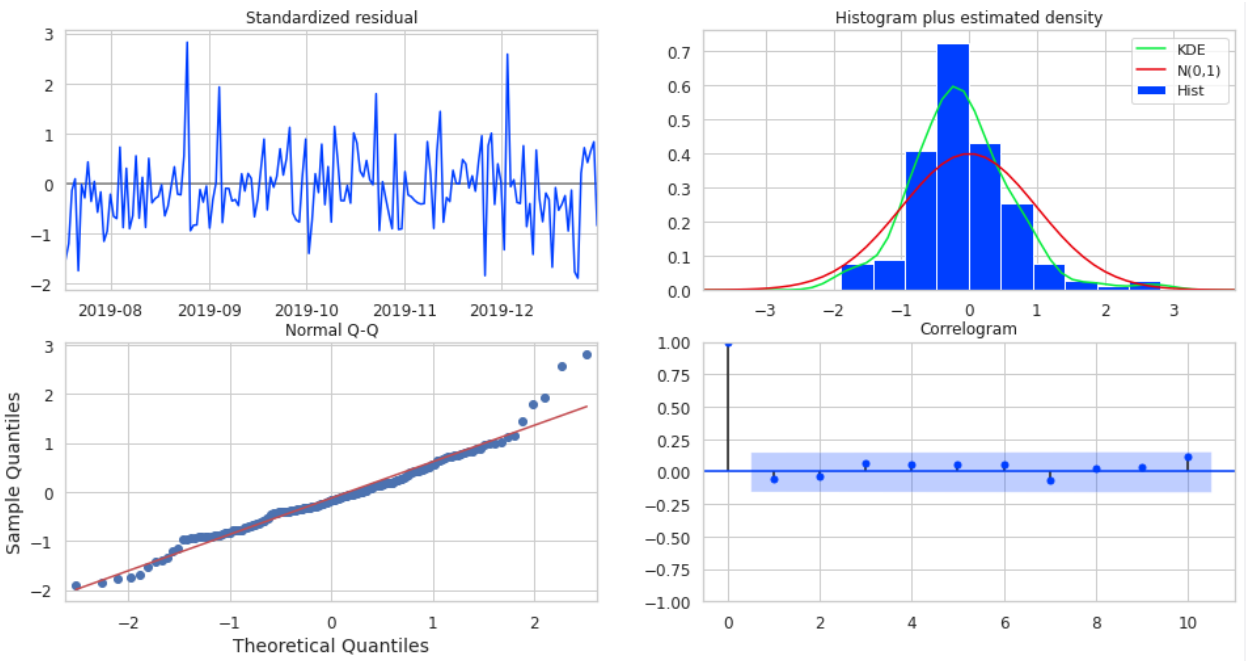


Kết quả của mô hình dự báo bằng phương pháp Holt-Winters' Seasonal cho thấy sự so sánh giữa Additive xu hướng phụ gia (đường màu đỏ) với Additive+dampes (đường màu xanh lá). Dựa vào biểu đồ trên cho thấy mô hình Holt-Winters' Seasonal dự đoán được kết quả phù hợp nhất so với dữ liệu thực tế. Tuy nhiên, RMSE lại cao hơn mô hình SES.

SARIMA

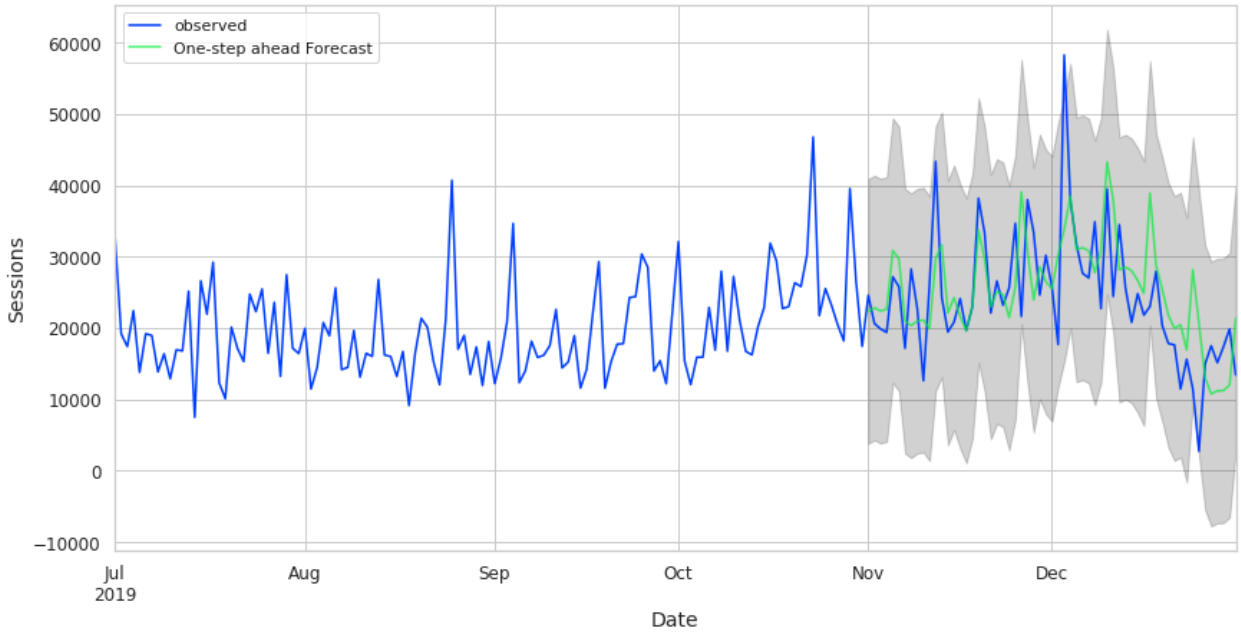
Phương pháp này phù hợp với dữ liệu chuỗi thời gian có tính xu hướng có hoặc không có tính mùa vụ.

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.8083	0.082	-9.838	0.000	-0.969	-0.647
ma.S.L7	-0.7689	0.097	-7.949	0.000	-0.958	-0.579
sigma2	8.944e+07	2.22e-10	4.02e+17	0.000	8.94e+07	8.94e+07

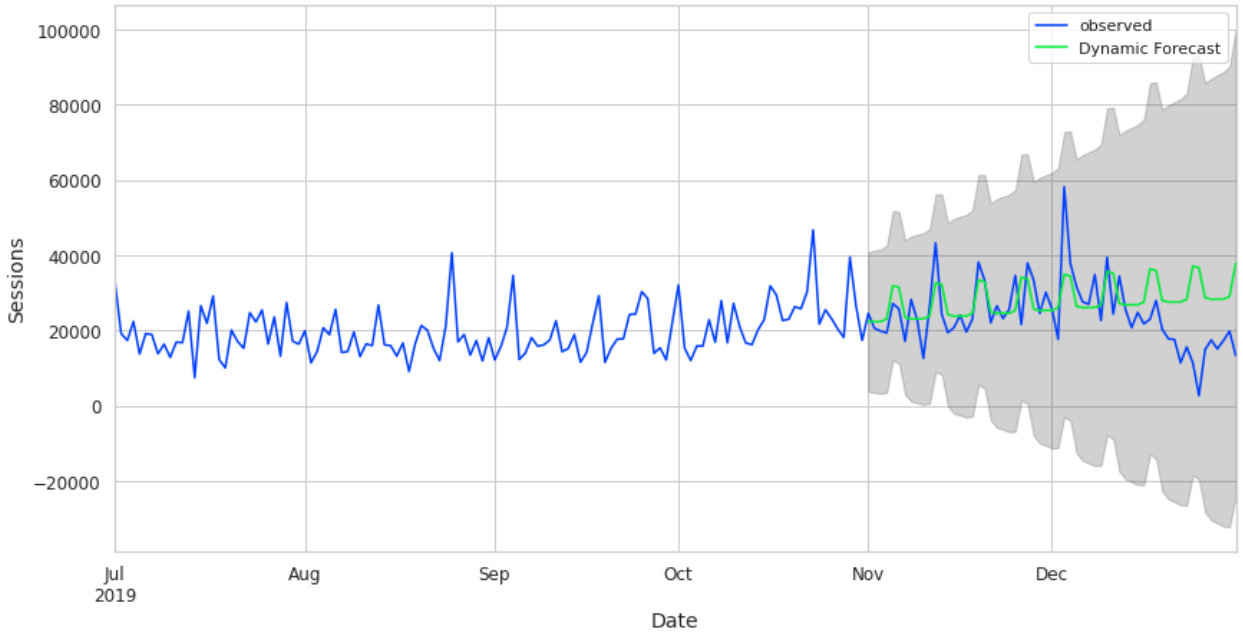


Biểu đồ cho thấy đã tìm ra được tính tính mùa vụ, xu hướng phù hợp và đã loại bỏ được dữ liệu bị nhiễu, phần dư được phân phối chuẩn và có sự tương quan thấp với chính nó.

The Root Mean Squared Error of SARIMA with season_length=7 and dynamic = False
8836.36



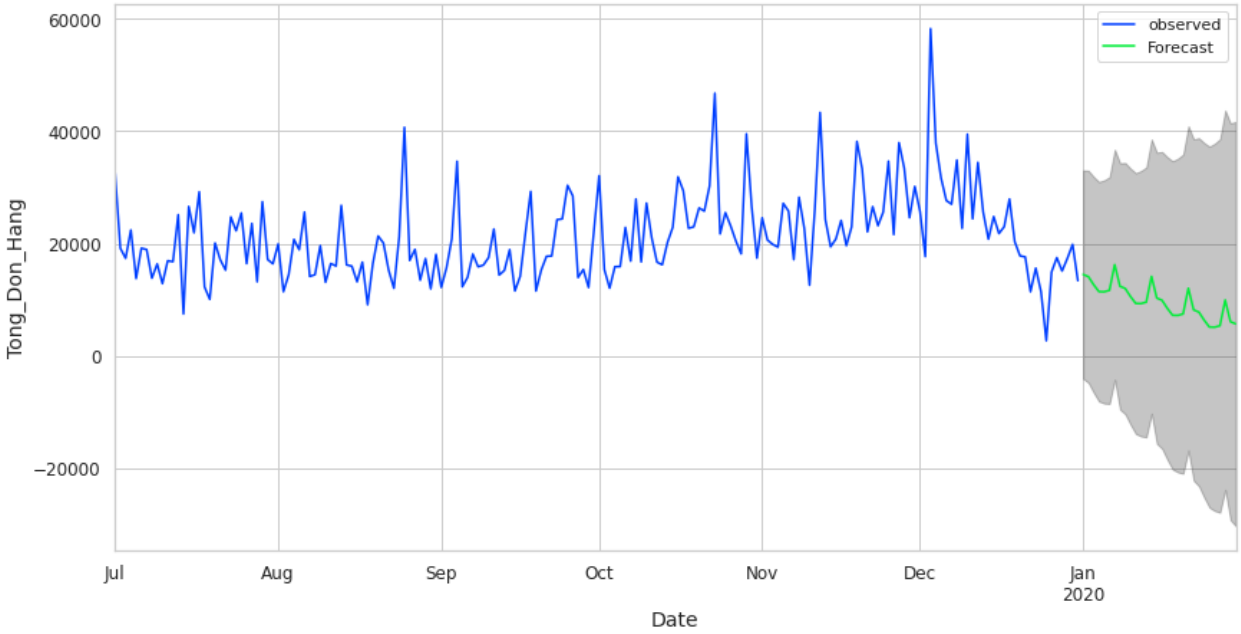
The Root Mean Squared Error of SARIMA with season_length=7 and dynamic = True
12737.29



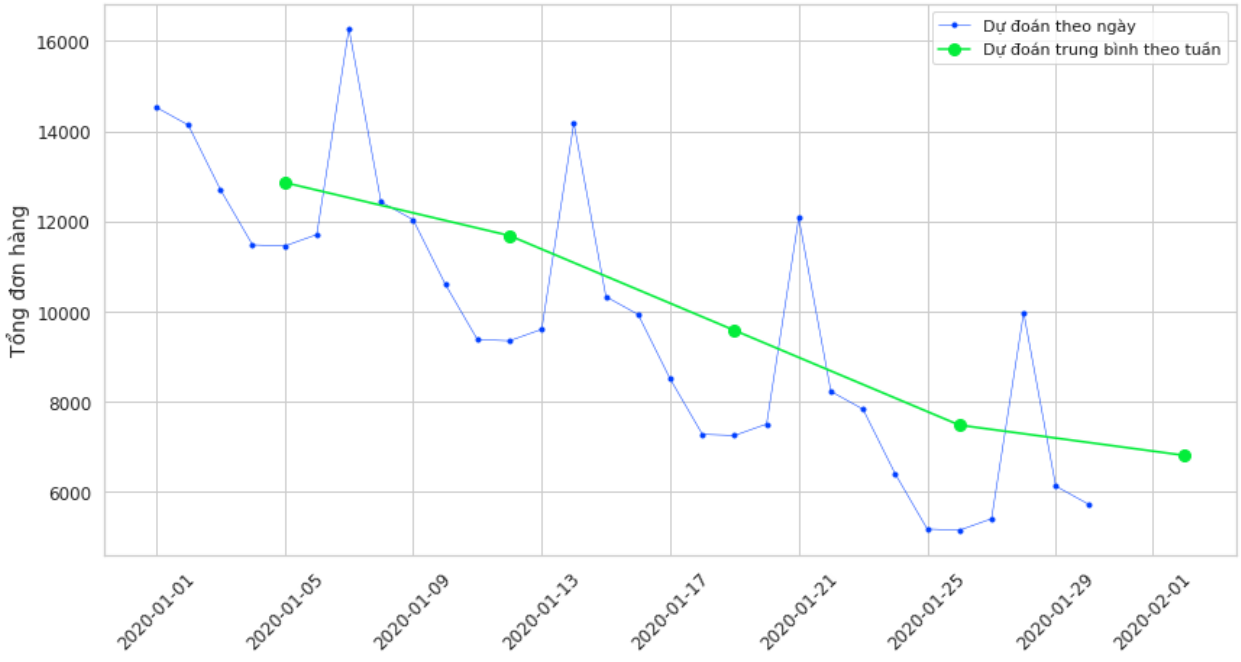
Dựa vào kết quả của mô hình SARIMA cho thấy: so với kết quả của tất cả các mô hình trước đó, mô hình SARIMA dự đoán tốt khi tập dữ liệu có tính thời vụ và xu hướng. Kết quả dự đoán của mô hình này gần nhất với số đơn hàng thực tế nhất.

Nếu chỉ lựa chọn mô hình dựa trên việc lựa chọn giá trị RMSE thấp nhất thì mô hình SES phù hợp nhất. Nhưng trong trường hợp này, để phục vụ cho việc dự báo dài hạn và dữ liệu có tính chu kỳ và tính mùa vụ thì mô hình SARIMA là mô hình phù hợp nhất.

4.1.7 Dự báo đơn hàng trong tương lai



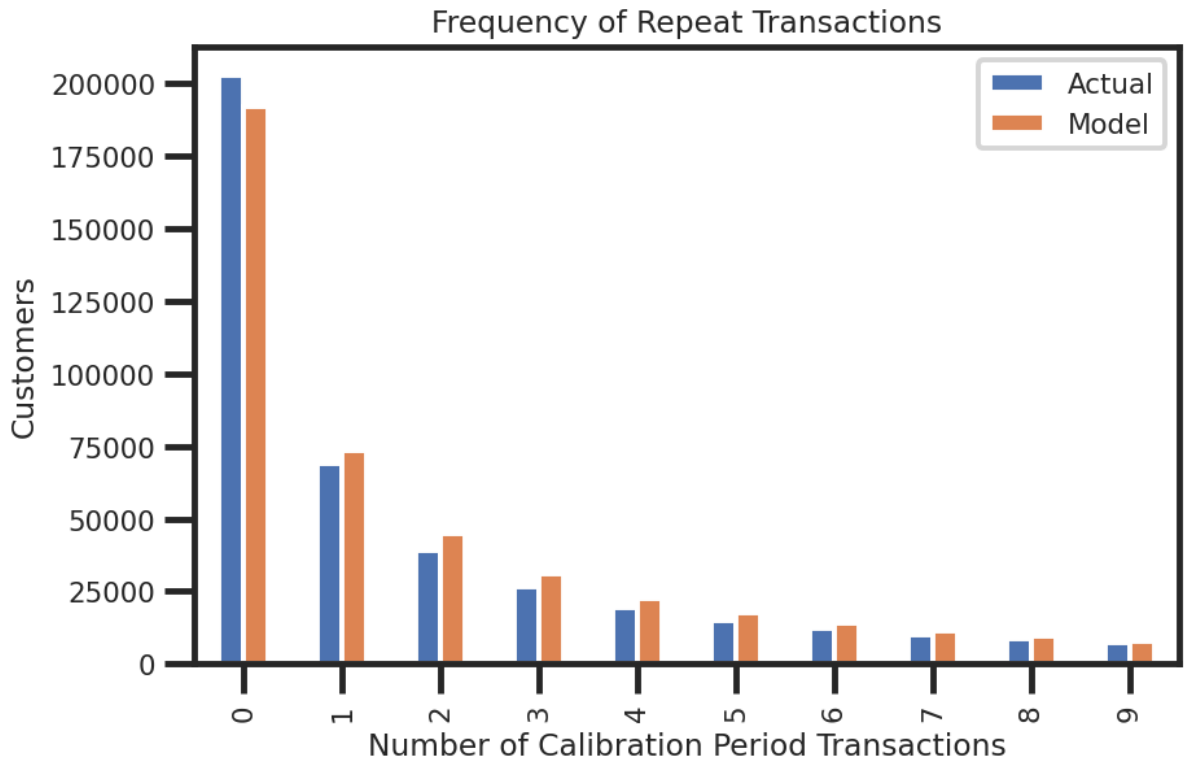
Đường màu xanh lá thể hiện số đơn hàng dự đoán trong 30 ngày tiếp theo trong tương lai dựa trên mô hình đã xây dựng.



Hình trên cho thấy chi tiết đơn hàng dự báo cho tháng tiếp theo. Nhìn chung ở tháng tiếp theo tổng đơn hàng của KPIM có xu hướng giảm

4.2 Dự đoán Customer Lifetime Value

Sau khi xây dựng model, nhóm sử dụng biểu đồ để xem lại độ chính xác của model so với thực tế.

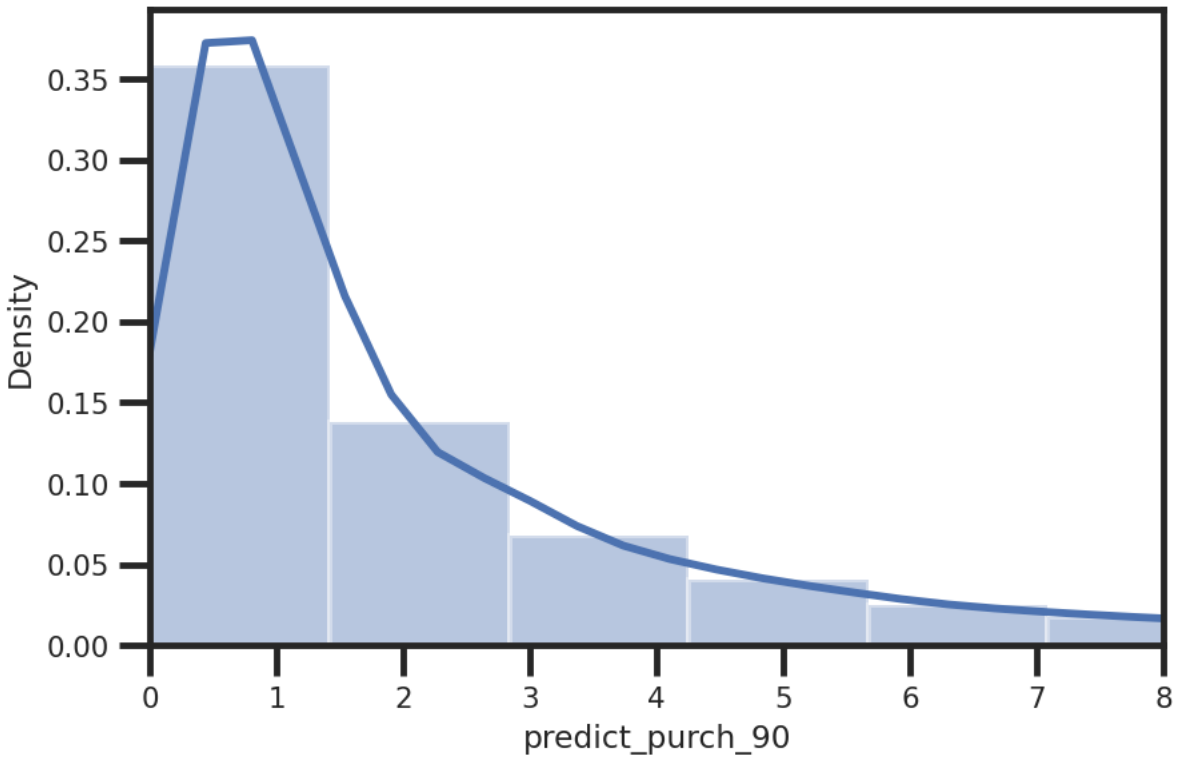


Biểu đồ trên cho thấy model có sự chênh lệch nhỏ giữa số dự đoán và thực tế, model dự đoán khá tốt. Từ đó nhóm sử dụng model đã xây dựng được để dự đoán số lượng khách hàng trong tương lai.

top 10 customers, by their predicted purchases over next 30 days

Ma_Khach_Hang	frequency	recency	T	monetary_value	predict_purch_10	predict_purch_30	predict_purch_60	predict_purch_90
21779170	152.0	182.0	183.0	6,722,220.4	7.7	23.1	46.1	68.9
18008790	148.0	182.0	182.0	4,039,370.9	7.6	22.6	45.1	67.4
49402770	146.0	182.0	182.0	13,783,674.9	7.5	22.3	44.5	66.5
22710390	146.0	183.0	183.0	4,464,549.3	7.4	22.2	44.3	66.2
510319460	143.0	182.0	182.0	5,176,782.5	7.3	21.9	43.6	65.1
51986480	140.0	183.0	183.0	3,328,115.6	7.1	21.3	42.5	63.5
36483980	138.0	183.0	183.0	6,095,677.1	7.0	21.0	41.9	62.6
1981990	138.0	183.0	183.0	7,426,679.1	7.0	21.0	41.9	62.6
515927240	135.0	179.0	181.0	2,461,923.0	6.9	20.7	41.3	61.8
34599730	134.0	181.0	182.0	5,972,443.8	6.8	20.5	40.8	61.0

Dựa trên các chỉ số Frequency, Rencency, T, Monetary, model dự đoán được danh sách 10 khách hàng sẽ đặt hàng nhiều nhất trong 30 ngày tới.



Hình trên cho thấy biểu đồ có dạng phân phối lệch phải, những khách hàng đã chi nhiều tiền để mua sắm thì sẽ có khuynh hướng đặt hàng nhiều hơn 1 lần trong tháng.

4.3. Phân cụm khách hàng với RFM và K-means

Để có thể gia tăng đơn hàng và doanh thu cũng như niềm tin của khách hàng trong tương lai, công ty cần phải hiểu rõ khách hàng của mình thông qua hành vi tiêu dùng của họ.

- Khách hàng đã chi tiêu bao nhiêu cho các sản phẩm của công ty?
- Khách hàng có quay lại mua hàng không và tần suất mua hàng của họ như thế nào?
- Khách hàng có mua hàng trong thời gian gần đây không?

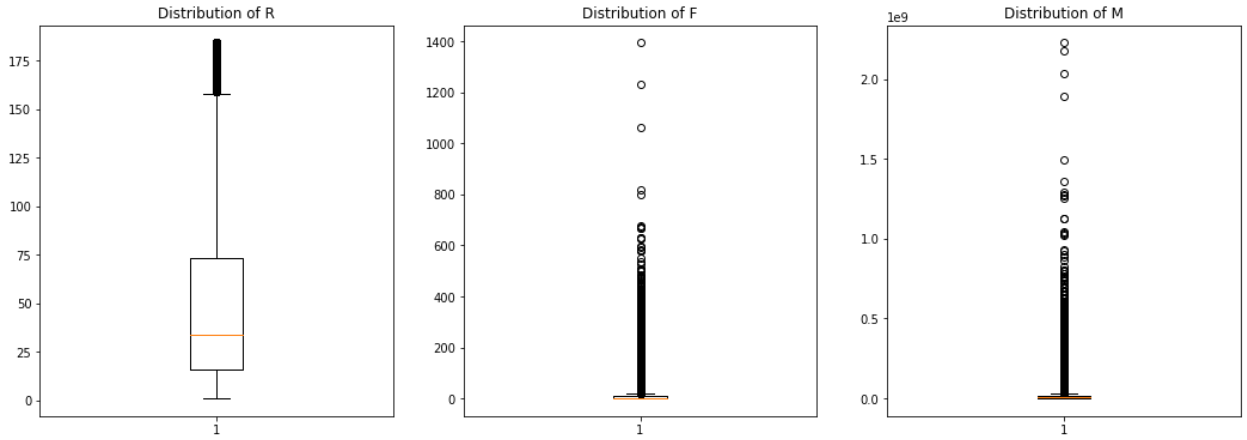
Từ những câu hỏi đó, nhóm tiến hành phân cụm khách hàng dựa trên mô hình RFM (Recency, Frequency, Monetary)

- Recency là khoảng thời gian tính từ hiện tại đến lần mua hàng gần nhất của khách hàng. Chỉ số này càng cao thì khách hàng càng lâu chưa quay lai mua hàng, tỉ lệ khách hàng rời bỏ cao.
- Frequency là tần suất mua hàng của khách hàng. Khách hàng mua hàng càng nhiều thì tần suất này sẽ càng cao.
- Monetary là tổng số tiền mà khách hàng đã chi tiêu, tác động trực tiếp tới doanh thu của công ty.

	R	F	M
count	475,350	475,350	475,350
mean	51	8	12,533,833
std	47	18	26,741,162
min	1	1	0
1%	1	1	552,124
2%	1	1	617,952
5%	3	1	786,266
10%	6	1	1,072,956
25%	16	1	1,964,188
50%	34	3	4,789,709
75%	73	8	12,689,709
90%	131	20	29,406,327
95%	158	34	47,701,173
98%	173	59	81,357,326
99%	179	83	115,695,811
max	184	1,393	2,226,240,344

Thống kê của ba giá trị Recency, Frequency và Monetary

Sau khi tính toán 3 giá trị R, F và M dựa trên dữ liệu thu được của 470350 khách hàng, có 2% khách hàng mua hàng gần nhất cách 1 ngày kể từ ngày cuối cùng của bộ dữ liệu, 75% khách hàng đã hơn 1 tháng chưa quay lại mua hàng và khách hàng có khoảng thời gian chưa mua hàng cao nhất là 184 ngày. Bên cạnh đó, có 25% khách hàng chỉ mới mua hàng một lần, đây là những khách hàng mới của của công ty, và đa số những khách hàng còn lại có tần suất mua hàng trên 2. Xét về chi tiêu của các khách hàng, trung bình chi tiêu rơi vào khoảng 12500000 và khách hàng chi tiêu nhiều nhất lên đến hơn 2,2 tỷ.



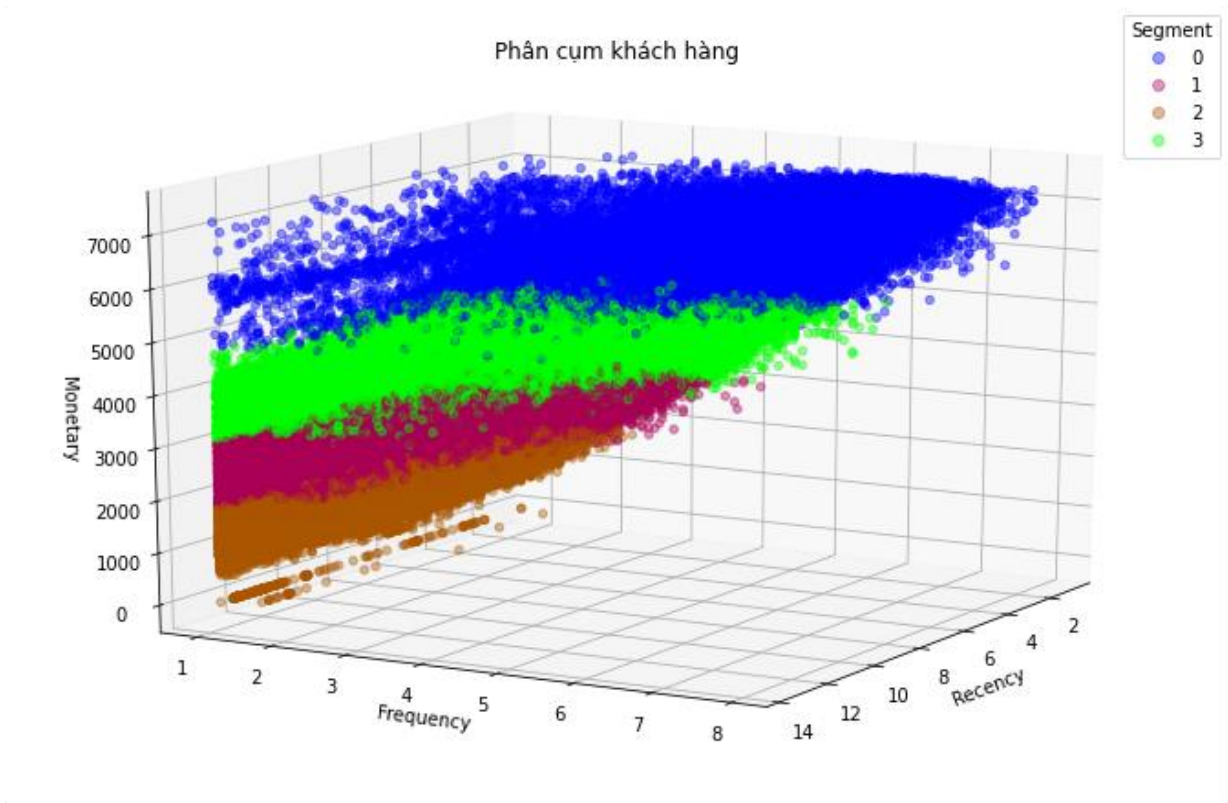
Phân phối các giá trị R, F, M

Từ hình trên có thể thấy có khá nhiều các giá trị nằm ở ngoài 2 đầu của boxplot. Nhóm đã sử dụng z-score để xác định các biến ngoại lai với ngưỡng (threshold) là 3, các giá trị có z-score lớn hơn 3 sẽ được xem là biến ngoại lai.

```
Observations considered as outliers in R
 R    0
F    0
M    0
dtype: int64
----
Observations considered as outliers in F
 R    8242
F    8242
M    8242
dtype: int64
----
Observations considered as outliers in M
 R    7439
F    7439
M    7439
dtype: int64
```

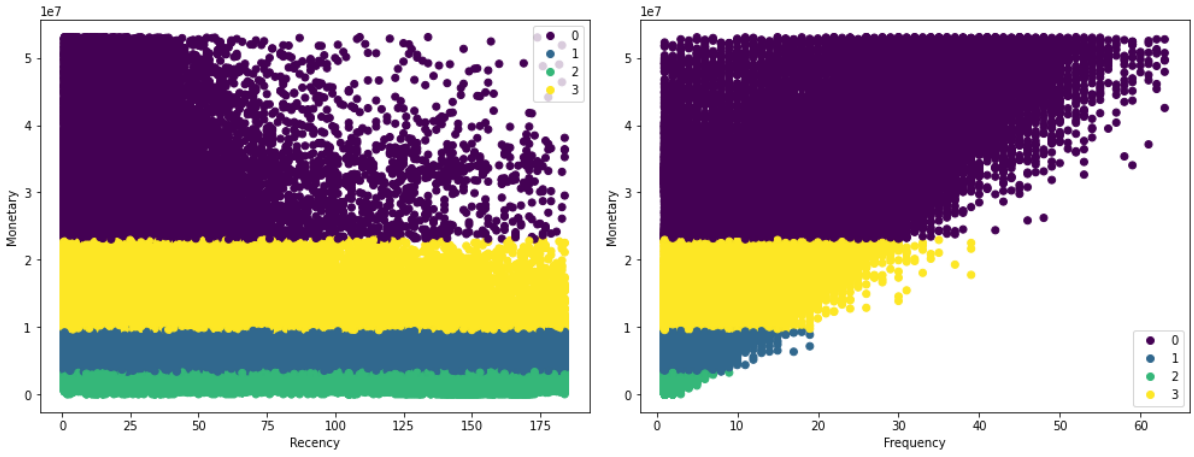
Những quan sát này có thể làm ảnh hưởng đến độ chính xác khi xác định centroid của K-means và gây mất thời gian khi phân cụm. Thế nên, nhóm quyết định loại các outlier này.

Để xác định số cụm k tối ưu cho mô hình, nhóm sử dụng phương pháp Elbow để xác định số k và xác định được số k = 4. Dưới đây là các điểm dữ liệu được gom lại thành 4 cụm.



Biểu diễn 3D các cụm RFM

Để dễ nhìn hơn ta có thể xem giá trị M dựa trên F và R như bên dưới.



Biểu diễn các cụm theo 2D

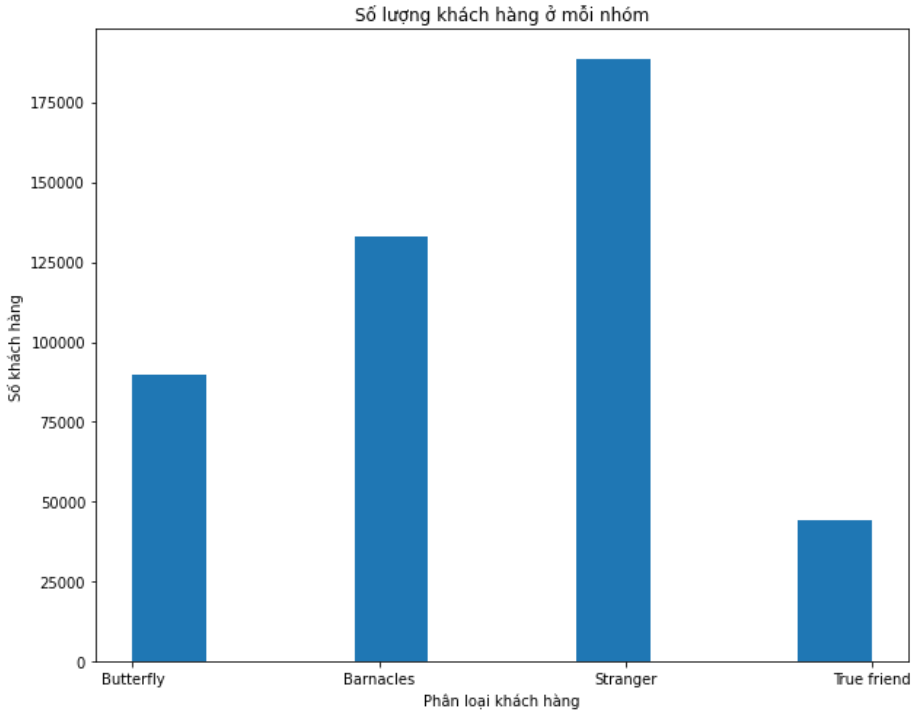
Có thể thấy giá trị của Recency ở các cụm khá giống nhau ngoại trừ cụm 0 thì các giá điểm dữ liệu có Recency chủ yếu từ thấp đến trung bình.

Bên cạnh đó khi quan sát sự phân bố của Monetary và Frequency, các khách hàng ở cụm 1 và cụm 2 có chi tiêu thấp và tần suất mua hàng cũng không nhiều. Cụm 3 thì có chi tiêu ở mức trung bình, cụm 4 có đóng góp nhiều vào doanh thu của công ty tuy nhiên tần suất mua của các khách hàng ở cụm 4 trải từ thấp đến cao.

Dựa trên các cụm và ma trận quan hệ khách hàng (customer relationship matrix), các khách hàng của công ty KPIM Ecommerce được phân thành 4 nhóm và mỗi phân khúc nên áp dụng những cách tiếp cận khách nhau:

- Cụm 0: True friends. Đây là những khách hàng trung thành nhất của công ty, họ đóng góp nhiều vào doanh thu công ty, cũng như có tần suất mua hàng cao và mua hàng gần đây. Có thể xem như đây là nhóm khách hàng VIP cần được chăm sóc với những chương trình khách hàng thành viên để họ cảm nhận được họ là có những giá trị nhất định khi là khách hành trung thành với công ty.
- Cụm 1: Barnacles. Đây là những khách hàng chi tiêu ít tuy nhiên nhưng hay quay lại mua hàng. Họ là những khách hàng trung thành nhưng không mang lại nhiều lợi nhuận. Thế nên công ty có thể upsell, giới thiệu thêm các các sản phẩm khác để khuyến khích nhóm này mua thêm hàng hoặc đưa ra các khuyến mãi phù hợp.
- Cụm 2: Strangers. Những khách hàng này không đóng góp nhiều vào doanh thu công ty và cũng không có sự gắn kết với công ty. Nhóm khách hàng này như những vị khách vãng lai, không tương thích nhiều những gì mà công ty mang lại, có thể cân nhắc không đầu tư gì nhiều vào nhóm khách hàng này. Tuy nhiên, trong trường hợp của KPIM Ecommerce, số lượng nhóm khách hàng này khá nhiều. Công ty có thể chạy các chiến dịch quảng cáo trên mạng xã hội để có thể tiếp cận tới nhiều khách hàng hơn và đề xuất các sản phẩm bán chạy để có thể dễ dàng tương thích với nhu cầu của nhóm khách hàng hơn.
- Cụm 3: Butterflies. Butterflies là những khách hàng tiêu dùng nhiều nhưng không gắn kết và không trung thành với thương hiệu, họ sẵn sàng rời đi khi thấy những ‘deal’ hời từ các thương hiệu khác. Đây thường là những khách hàng ngắn hạn và

khó để trở thành khách hàng trung thành, công ty nên tận dụng đặc tính thích các món hời của nhóm này để kiếm lợi nhuận.



Số lượng khách hàng ở mỗi phân khúc

Có thể thấy nhóm được kì vọng nhiều nhất True Friend lại có số lượng khách hàng ít nhất, trong khi đó số lượng khách hàng thuộc nhóm Stranger chiếm nhiều nhất. Công ty cần đưa ra chiến lược phù hợp để tăng lượng khách hàng trung thành.

4.4. Tối ưu vị trí kho hàng

Dựa trên những phân tích trên, công ty vẫn chưa tận dụng tối ưu số lượng kho hàng mà mình đang có. Hiện công ty đang có 1906 kho hàng phân bố ở khắp các tỉnh thành nhưng thời gian vận chuyển vẫn còn khá cao, trong đó thời gian chuẩn bị hàng khá lâu.

Để tối ưu hóa vị trí các kho hàng hiện tại dựa theo nhu cầu đặt hàng của khách hàng, nhóm sử dụng Weighted Kmeans để tìm ra vị trí tối ưu của các kho hàng nhằm tiết kiệm chi phí thuê mướn kho, nhân công quản lý cũng như nâng cao chất lượng dịch vụ giao

hàng. Vị trí các kho hàng là tâm (centriod) của cụm mà ở đó các tỉnh thành có càng nhiều đơn hàng thì tâm sẽ dịch chuyển dần về phía đó.

Nhận định rằng:

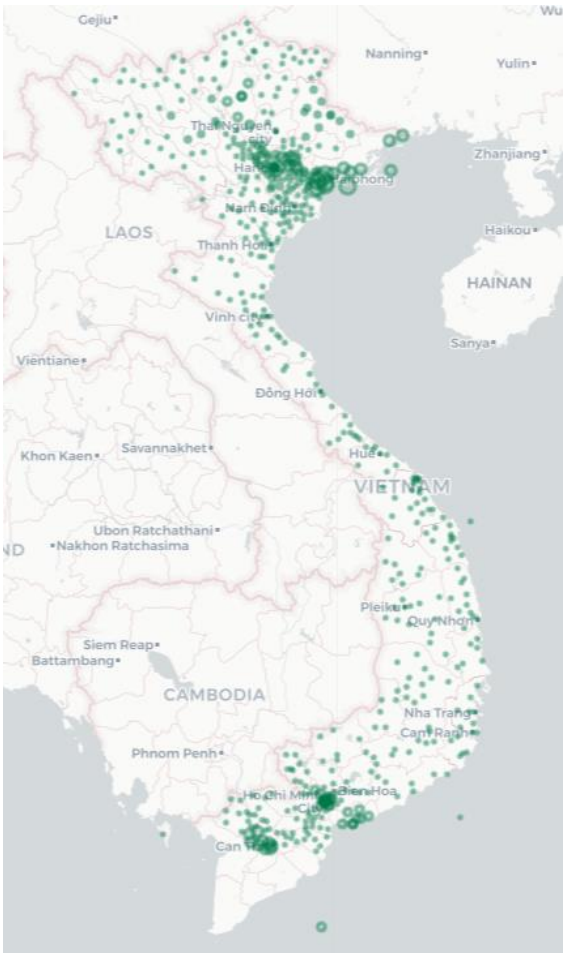
- Công ty vận chuyển chủ yếu bằng đường bộ.
- Mỗi kho đều có đủ các loại sản phẩm (trang sức, quần áo, trang trí,...).
- Các kho của công ty nhiều nhưng chưa tối ưu khi thời gian chuẩn bị hàng.

Đầu tiên nhóm xử lý các dữ liệu các đơn hàng, nhóm lại theo tỉnh thành, thành phố nhận hàng và tính thời gian nhận hàng trung bình ở các tỉnh và tổng số đơn hàng mà người dân ở tỉnh đó dã đặt. Bên cạnh đó, các dòng sau đây sẽ bị loại khỏi tập dữ liệu:

- Trạng thái hủy như CX, BX, CP đều được loại khỏi tập dữ liệu.
- Ngày nhận hàng (Ngay_Nhan_Hang) null.
- Nơi đến tỉnh thành (Noi_Den_Tinh_Thanh) null.

	Noi_Den_Tinh_Thanh	Noi_Den_Thanh_Pho	Thoi_Gian_Van_Chuyen	Ma_Don_Hang	lat	lng
0	An Giang	An Châu	7.12	9205	10.45	105.39
1	An Giang	An Phú	6.96	9012	10.82	105.09
2	An Giang	Chợ Mới	6.95	9122	10.55	105.40
3	An Giang	Cái Dầu	6.87	9114	10.57	105.23
4	An Giang	Long Xuyên	7.06	9174	10.37	105.42
...	...	...	...	...	...	...
551	Đồng Tháp	Mỹ Thọ	12.40	144	10.44	105.70
552	Đồng Tháp	Sa Rài	12.41	142	10.87	105.47
553	Đồng Tháp	Sa Đéc	12.01	162	10.31	105.74
554	Đồng Tháp	Thanh Bình	14.39	143	10.56	105.48
555	Đồng Tháp	Tràm Chim	13.30	142	10.67	105.56

Kinh độ, vĩ độ của các thành phố được lấy từ nguồn dữ liệu bên ngoài để phục vụ cho việc phân cụm dễ dàng hơn.



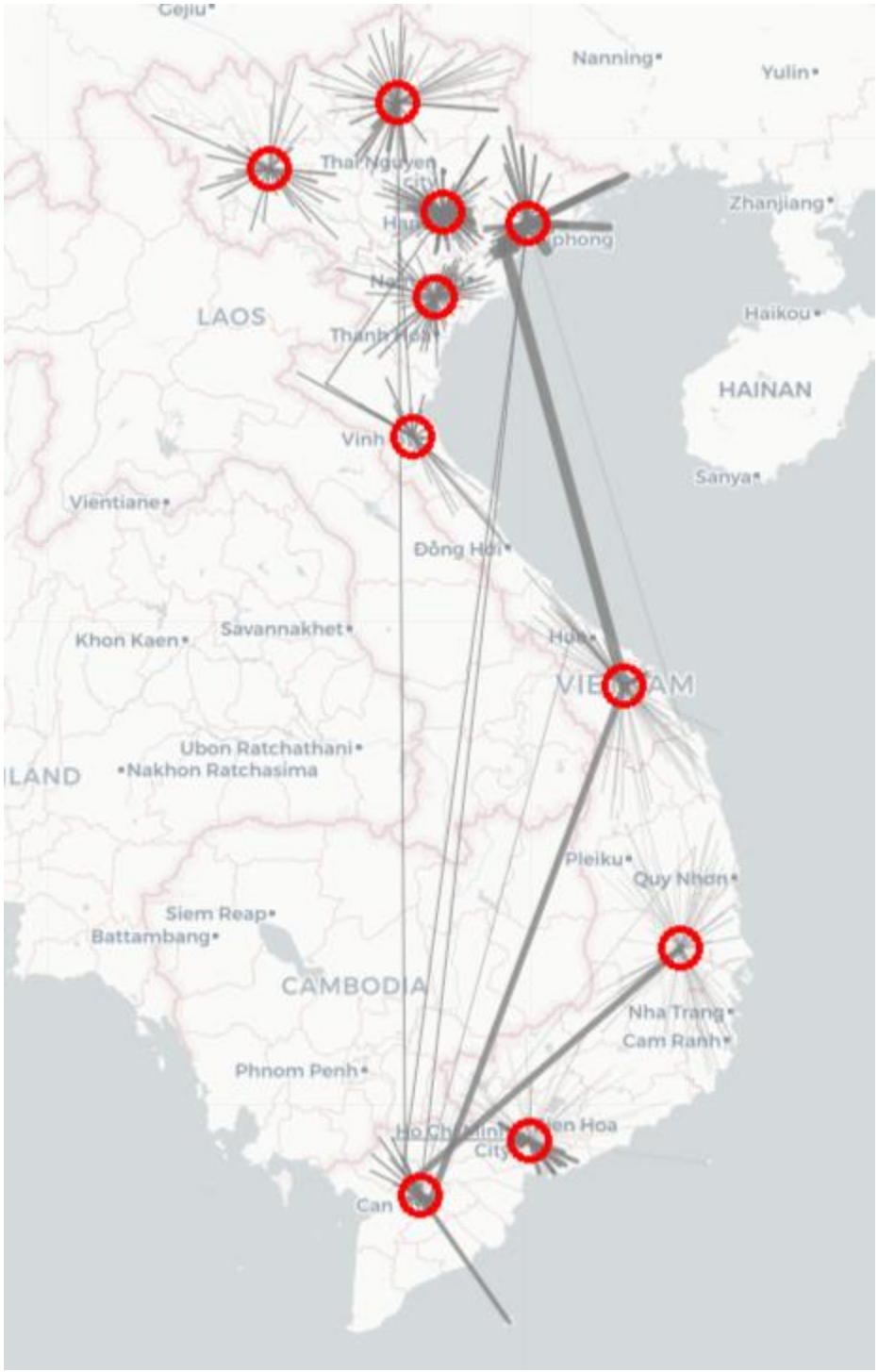
Nhu cầu mua hàng ở các thành phố trong 6 tháng cuối năm 2019

Có thể thấy ở khu vực phía Bắc và phía Nam có có khá nhiều đơn hàng được đặt, ngược lại ở miền Trung còn khá rải rác.

	0	1	ClusterID	location
0	21.61	103.82	1	Mường La , Tỉnh Sơn La
1	12.97	108.62	2	Ea Kar , Đắk Lắk
2	10.14	105.57	3	Huyện Cờ Đỏ Xã Thới Hưng, Thành phố Cần Thơ
3	20.23	105.75	4	Nho Quan , Tỉnh Ninh Bình
4	15.92	107.95	5	Đại Lộc , Tỉnh Quảng Nam
5	21.14	105.84	6	Thành phố Hà Nội,
6	10.76	106.85	7	Huyện Nhơn Trạch Xã Long Tân, Tỉnh Đồng Nai
7	21.02	106.83	8	Quảng Yên ,
8	22.34	105.31	9	Lâm Bình , Tỉnh Tuyên Quang
9	18.69	105.50	10	Nam Đàn , Tỉnh Nghệ An

Vị trí các kho hàng đề xuất

Sau khi gom cụm ta xác định được 10 vị trí kho hàng như ở các địa phương như Sơn La, Đắk Lắk, Hà Nội,...



Các kho hàng có vị trí tối ưu

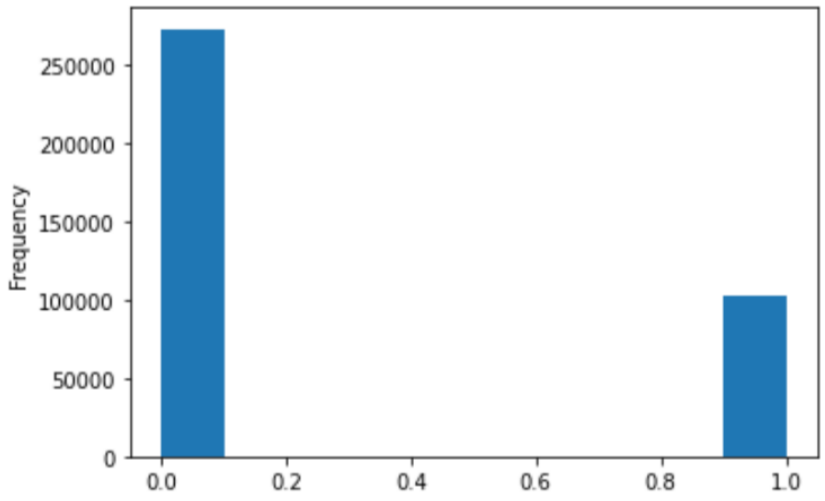
Để dễ dàng hình dung vị trí các kho hàng hơn, nhóm đã đánh dấu vị trí các kho hàng trên bản đồ. Theo như gom cụm, ta sẽ có 5 kho ở phía Bắc, 2 kho miền Trung và 3 kho ở phía Nam. Các đường thẳng màu xám thể hiện kho này thể hiện các kho đề xuất này có thể vận chuyển tới những thành phố, tỉnh thành nào. Đường thẳng càng đậm nghĩa là nhu cầu mua hàng của những vị trí đó trong 6 tháng 2019 càng cao.

Công ty có thể xem xét giảm các nhà kho ít xuất hàng đi và đặt các nhà ở ở các vị trí như trên để tối ưu việc vận chuyển cũng như tiết kiệm được chi phí thuê kho bãi, nhân công và quản lý kho.

4.5. Dự báo đơn hàng trễ

Để nâng cao chất lượng dịch vụ, gia tăng niềm tin khách hàng. KPIM cần xem xét đến việc giảm thiểu số lượng đơn hàng bị giao trễ. Nhóm sử dụng các thuật toán học giám sát và phân loại như Logistic Regression, Random Forest,XG Boost,... với những bước chính như sau:

Kiểm tra sự phân phối của biến phụ thuộc



Kết quả cho thấy dữ liệu bị mất cân bằng, tỉ lệ đơn hàng giao đúng hẹn (0) nhiều hơn đơn hàng bị giao trễ (1). Sự mất cân bằng này có thể dẫn tới việc phân loại kém chính xác đối với những đơn hàng giao trễ. Bởi đa phần kết quả dự báo ra thường thiên về 1

nhóm là nhóm giao hàng đúng hẹn. Để cải thiện kết quả dự báo chúng ta cần những điều chỉnh thích hợp để mô hình đạt được một độ chính xác cao trên nhóm thiểu số.

Mã hóa các biến phân loại

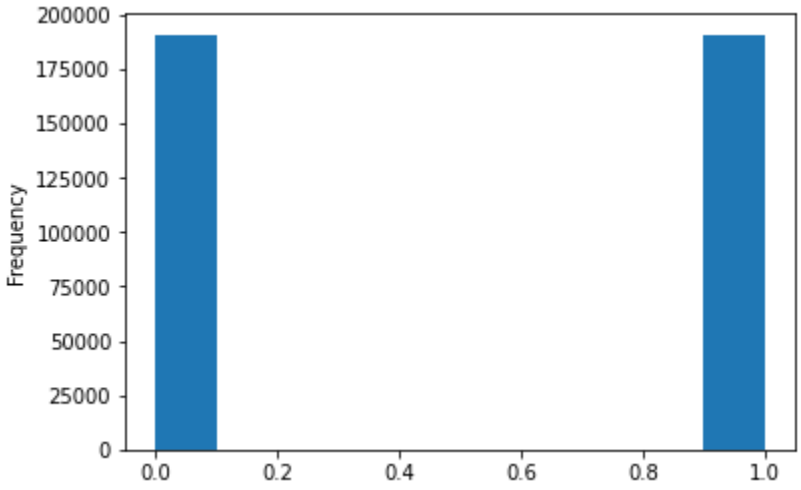
Trước khi áp dụng SMOTE để cân bằng dữ liệu, chúng ta cần xử lý các biến phân loại. Có hai cách để thực hiện quá trình này:

- Mã hóa nhãn (Label encoding): gán nhãn cho một biến phân loại với một số nguyên. Không tạo ra cột mới
- Mã hóa một lần (One-hot encoding) : tạo một cột mới cho từng danh mục trong một biến phân loại. Mỗi quan sát nhận được 1 trong cột cho danh mục tương ứng của nó và một 0 trong tất cả các cột mới khác.

  Trong bài này, nhóm sẽ sử dụng mã hóa một lần cho các biến phân loại

Xử lý dữ liệu không cân bằng với SMOTE

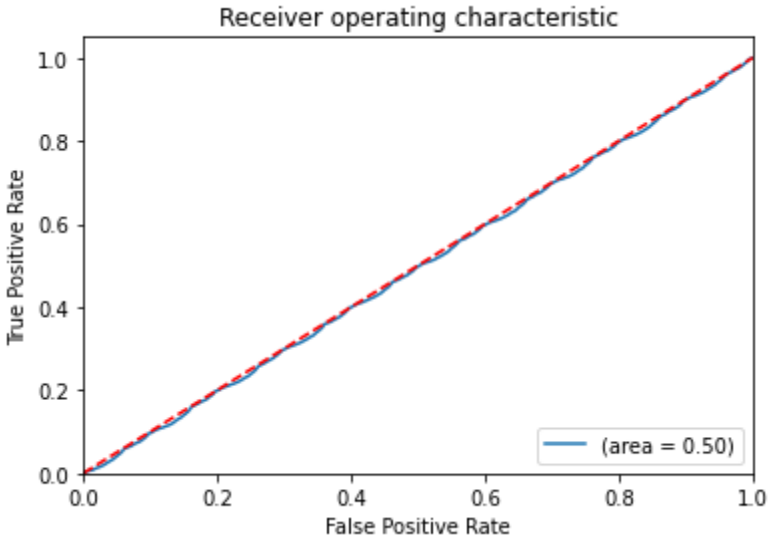
Nhóm chỉ oversampling trên dữ liệu đào tạo, sẽ không có thông tin nào được đưa từ dữ liệu kiểm tra vào SMOTE. Hình bên dưới là kết quả sau khi chạy:



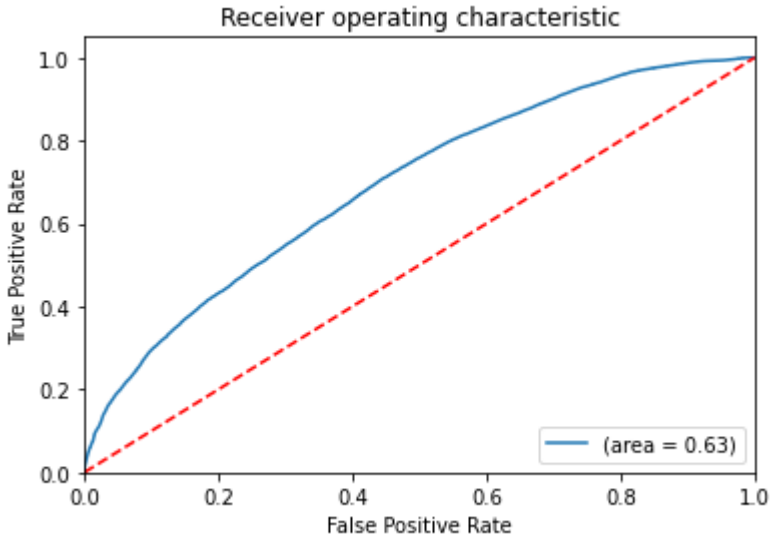
Áp dụng thuật toán, đánh giá và lựa chọn mô hình

- **Thuật toán Logistic Regression**

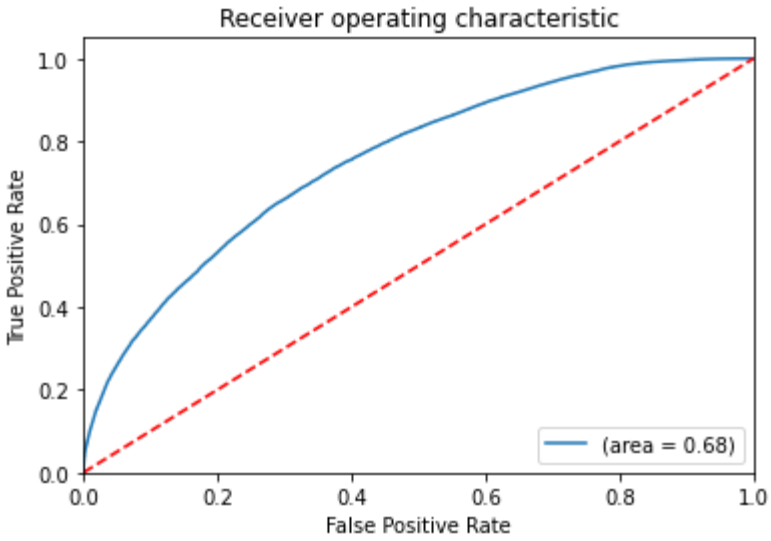
  Sau khi chạy mô hình. Kết quả đánh giá như sau:



- **Thuật toán Random Forest**



- **Thuật toán XG Boost**



Dựa trên bảng kết quả ROC của từng model, ta so sánh độ hiệu quả và đưa ra kết luận: Mô hình XG Boost cho kết quả dự đoán chính xác cao nhất (68%).