Retail Giant Sales Forecasting Case Study

Name:- Divit Karmiani
Email ID:- divitkarmiani1998@gmail.com

Problem Statement

Global Mart is an online supergiant store that has worldwide operations. This store takes orders and delivers across the globe and deals with all the major product categories — consumer, corporate and home office.
As a sales manager for this store, you have to forecast the sales of the products for the next 6 months, so that you have a proper estimate and can plan your inventory and business processes accordingly.

Task - Identify the most profitable market segment and forecast the sales of the products of the identified segment for the next 6 months

Data Description

The store dataset has the following 5 attributes and their data description is as given below:

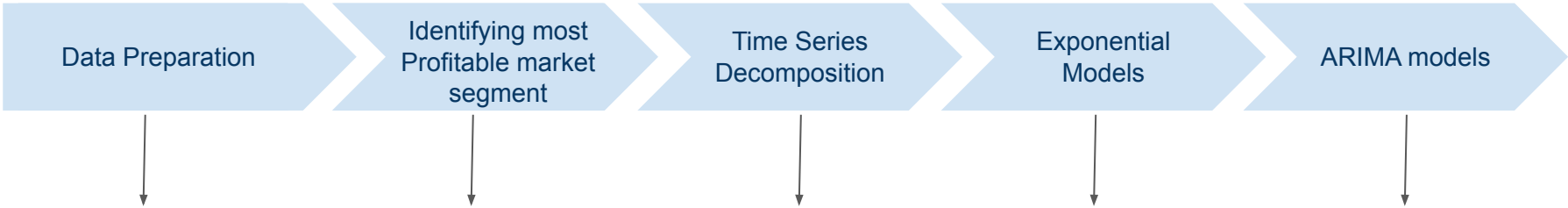
Attributes	Description
Order-Date	The date on which the order was placed
Segment	The segment to which the product belongs
Market	The market to which the customer belongs
Sales	Total sales value of the transaction
Profit	Profit made on the transaction

The store caters to 7 different geographical market segments and 3 major customer segments

Market	Segment
Africa	Consumer
APAC (Asia Pacific)	Corporate
Canada	Home Office
EMEA(Middle East)	
EU (European Union)	
LATAM (Latin America)	
US (United States)	

There are a total of 21 market segments

Analysis Approach



- Data Preparation involves various steps like aggregation of data at month level, creating the market segment column and split the data into train and test sets
- Here the train set is data for the first 42 months and test set is data for the last 6 months

- In order to find the most profitable market segment, we use the CoV method
- The coefficient of variation (COV) is the ratio of the standard deviation of a data set to the expected mean
- The market segment with the least CoV value for profit is the most consistently profitable

- This step is done to analyze the series as a combination of level, trend, seasonality, and noise components.
- This helps to get a better understanding of problems during time series analysis and forecasting.

- Exponential forecasts are equal to a weighted average of past observations and the corresponding weights decrease exponentially as we go back in time.

- AutoRegressive model forecasts correspond to a linear combination of past values of the variable
- In a Moving Average model the forecasts correspond to a linear combination of past forecast errors.

Data Preparation

- Step 1:- Convert Order Date to year-month
- Step 2:- Create Market_Segment column
- Step 3:- Aggregate profit data by month
- Step 4:- Split aggregated data into train and test
- Step 5:- Use train data to find the CoV value for various market segments

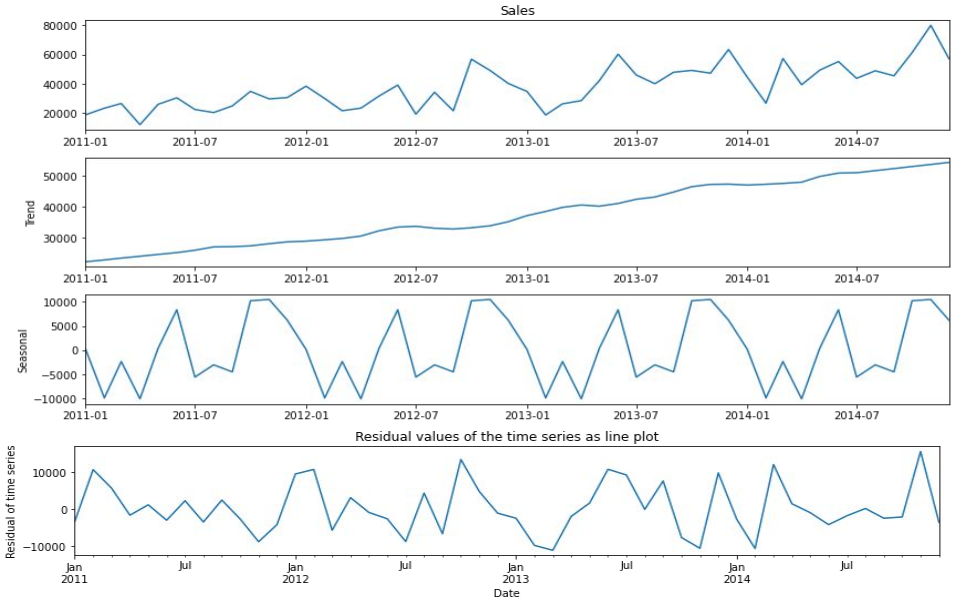
Identifying Most Profitable Segment

	segment	cov
0	APAC_Consumer	0.522725
1	APAC_Corporate	0.530051
12	EU_Consumer	0.595215
15	LATAM_Consumer	0.683770
13	EU_Corporate	0.722076
16	LATAM_Corporate	0.882177
14	EU_Home Office	0.938072
2	APAC_Home Office	1.008219
18	US_Consumer	1.010530
19	US_Corporate	1.071829
20	US_Home Office	1.124030
17	LATAM_Home Office	1.169693
6	Canada_Consumer	1.250315
3	Africa_Consumer	1.310351
7	Canada_Corporate	1.786025
4	Africa_Corporate	1.891744
5	Africa_Home Office	2.012937
8	Canada_Home Office	2.369695
9	EMEA_Consumer	2.652495
10	EMEA_Corporate	6.355024
11	EMEA_Home Office	7.732073

- To find the most consistently profitable market-segment we are using a measure called **"Coefficient of Variation (CoV)"**
- The coefficient of variation or CoV is the ratio of the standard deviation to mean for the data that it is being calculated for.
- We checked the CoV values calculated on profit for all the **21 market segments** and compare. You will find that these values vary a lot and hence it is not useful to compare the profits based on the standard deviation and their mean.
- A better metric to compare the variance between the segments is coefficient of variation which will normalise the standard deviation with the mean and give you a comparative figure on the basis of which you can identify the most profitable market segment.
- By using this, we find that the market segment with the least CoV value is **APAC_Consumer,** and hence the most profitable
- We then filtered the data for the selected segment to contain only sales information and aggregated by this filtered data by months to forecast its sales

Time Series Decomposition - Additive

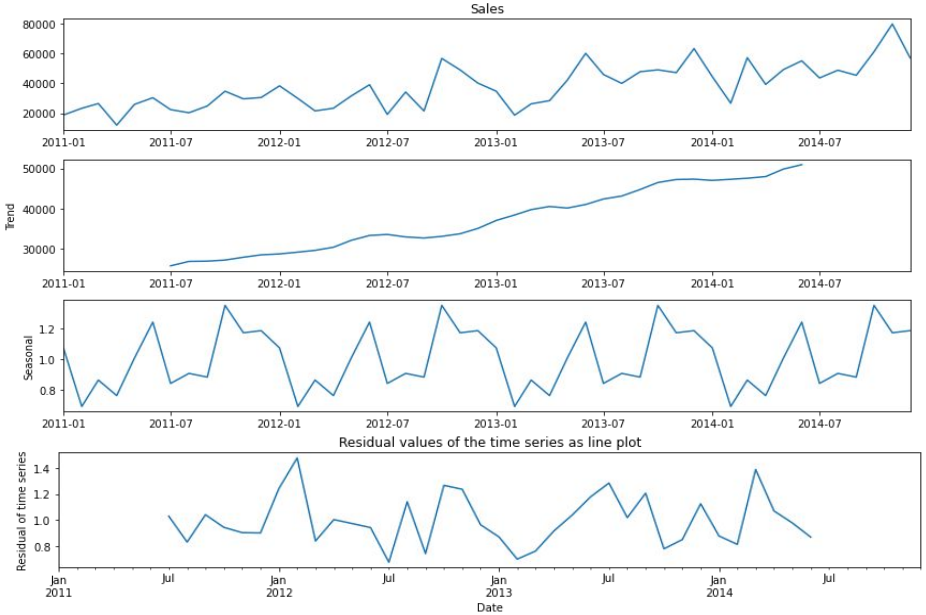
Additive decomposition argues that time series data is a function of the sum of its components trend-cycle component, seasonal component, and the remainder



- The trend seems to be almost linearly increasing with few negligible crests and troughs
- The seasonal pattern, though not evident from the actual Sales plot, has been captured but doesn't look to be a strong pattern
- There is no pattern observed in Residuals graph

Time Series Decomposition - Multiplicative

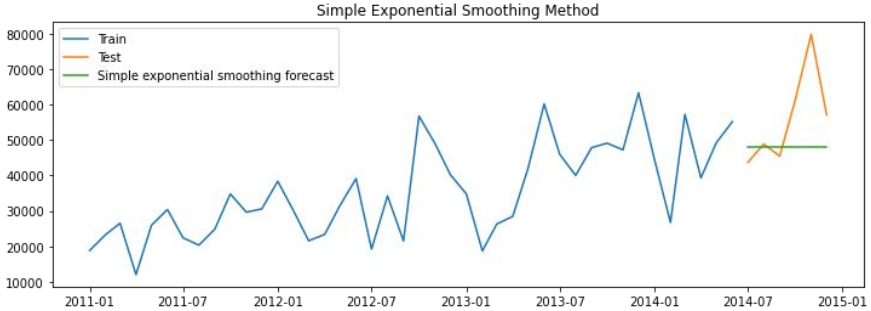
Multiplicative decomposition argues that time series data is a function of the product of its components trend-cycle component, seasonal component, and the remainder



- The trend seems to be almost linearly increasing with few evident crests and troughs
- The seasonal pattern, though not evident from the actual Sales plot, has been captured but doesn't look to be a strong pattern
- There is no pattern observed in Residuals graph

Exponential Models - 1

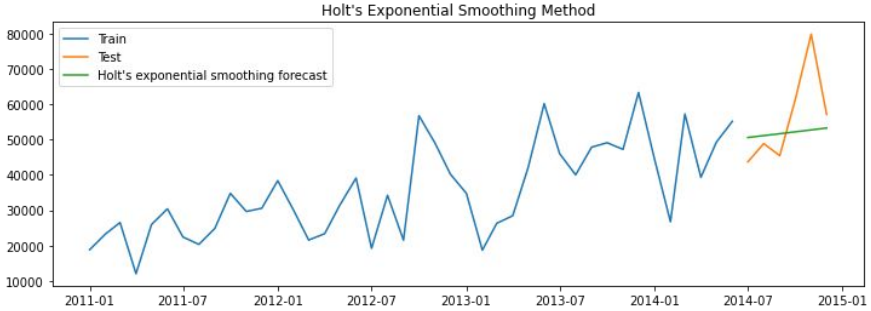
Simple Exponential Model



RMSE	MAPE
14627.34	15.74

The simple exponential smoothing method has captured the level in the time series data. But other components like trend and seasonality isn't captured. Hence high RMSE and MAPE values

Holt's Exponential Model

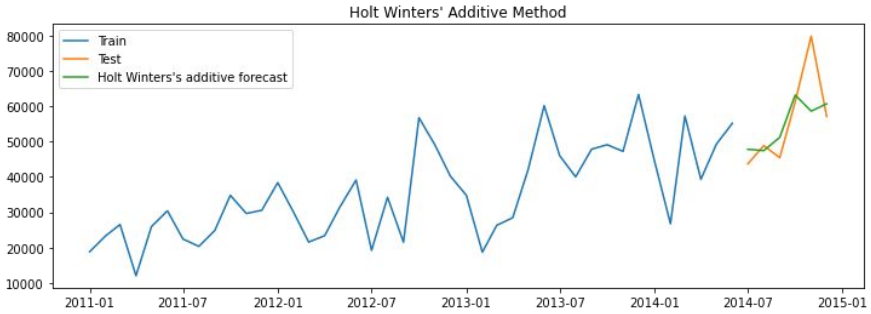


RMSE	MAPE
12403.84	14.93

The Holt's exponential smoothing method has captured the level and trend in the time series data. Hence lower RMSE and MAPE values than Simple exponential smoothing method. But still seasonal component hasn't been captured.

Exponential Models - 2

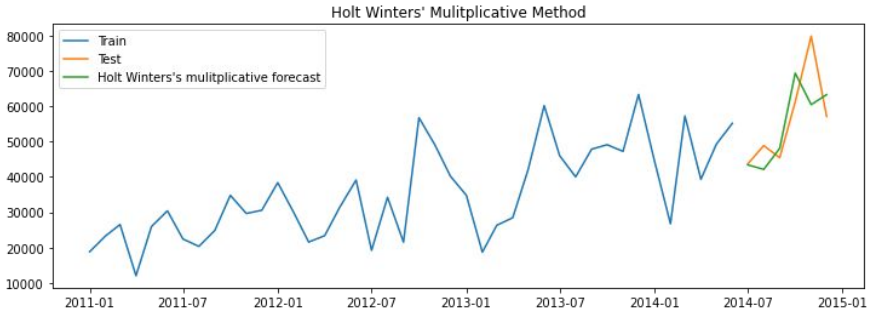
Holt Winters' additive



RMSE	MAPE
9306.82	10.17

Since Holt Winters' method captures all three components level, trend and seasonality, the RMSE and MAPE values are the lowest as compared to others. Since the seasonality pattern was not so strong, the forecast doesn't follow the actual values to the point.

Holt Winters' multiplicative



RMSE	MAPE
9423.23	11.43

Since Holt Winters' method captures all three components level, trend and seasonality, the RMSE and MAPE values are lower than others except for Holt Winters' additive method. Since the seasonality pattern was not so strong, the forecast doesn't follow the actual values to the point.

ARIMA Models - Testing Stationarity of Data

Two types of tests are done for this purpose:

1. **Augmented Dickey-Fuller (ADF) test -**
   - ADF Statistic: -3.376024
   - Critical Values @ 0.05: -2.93
   - p-value: 0.011804

   The series is stationary as p-value is less than 0.05.

2. **Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test -**
   - KPSS Statistic: 0.577076
   - Critical Values @ 0.05: 0.46
   - p-value: 0.024720

   The series is not stationary as p-value is less than 0.05

ARIMA Models - Types of Stationarity

Let us understand the different types of stationarities and how to interpret the contradictory results of the above tests(not contradictory in actual as indicated below).

1. **Strict Stationary:** A strict stationary series satisfies the mathematical definition of a stationary process. For a strict stationary series, the mean, variance and covariance are not the function of time. The aim is to convert a non-stationary series into a strict stationary series for making predictions.

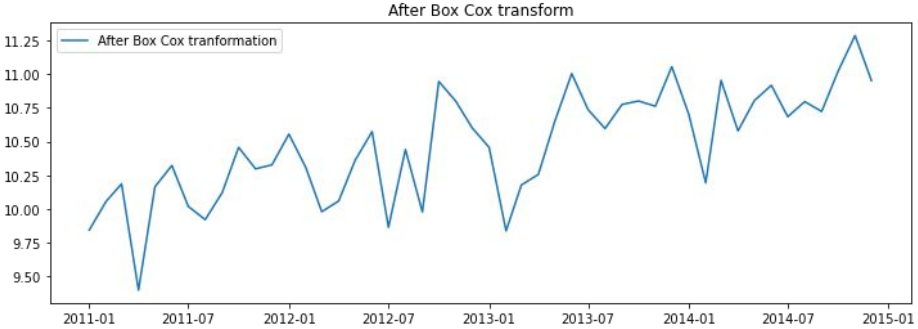
2. **Trend Stationary:** A series that has no unit root but exhibits a trend is referred to as a trend stationary series. Once the trend is removed, the resulting series will be strict stationary. The KPSS test classifies a series as stationary on the absence of unit root. This means that the series can be strict stationary or trend stationary.

3. **Difference Stationary:** A time series that can be made strict stationary by differencing falls under difference stationary. ADF test is also known as a difference stationarity test. It's always better to apply both the tests, so that we are sure that the series is truly stationary. Let us look at the possible outcomes of applying these stationary tests.

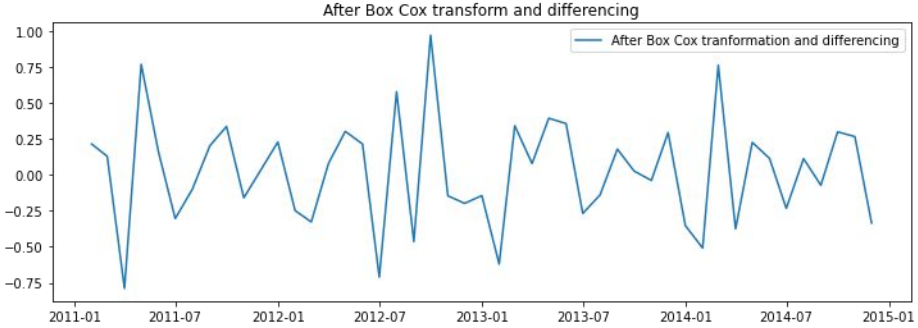
Case 1: Both tests conclude that the series is not stationary -> series is not stationary
Case 2: Both tests conclude that the series is stationary -> series is stationary
Case 3: KPSS = stationary and ADF = not stationary -> trend stationary, remove the trend to make series strict stationary
Case 4: KPSS = not stationary and ADF = stationary -> difference stationary, use differencing to make series stationary

For our data, Case 4 applies.

ARIMA Models - Box Transformations



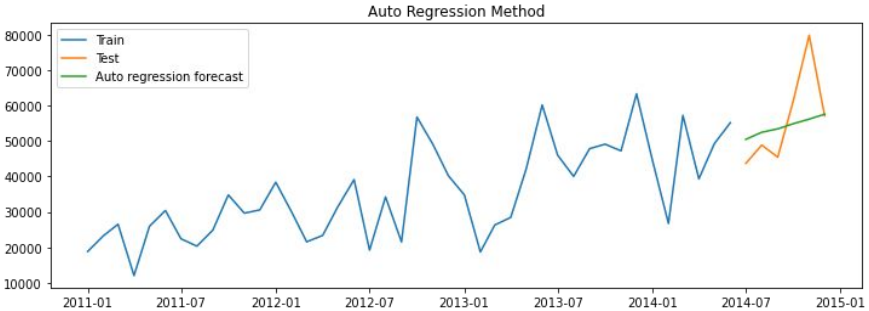
Box transformations are done to make the data stationary in order to implement ARIMA models



The time series data seems to have constant mean after differencing. The series seems to be stationary now.

ARIMA Models - 1

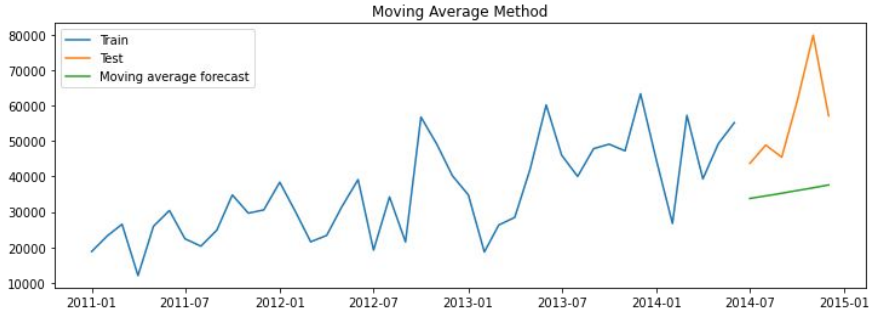
Auto Regression



RMSE	MAPE
10985.28	13.56

This model has captured the trend well. Though we have not used PACF to determine lag order 'p' and chosen by default as 1, the model seems to have performed well as seen by low MAPE value.

Moving Average

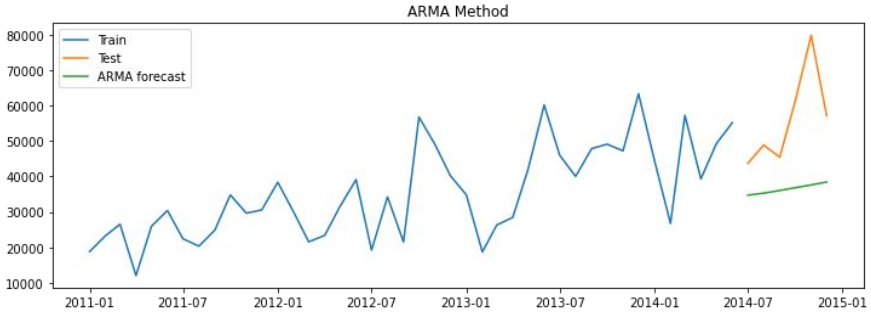


RMSE	MAPE
23360.02	33.93

This model has captured the trend well but not the level. The model has not performed well as seen by high RMSE and MAPE value.

ARIMA Models

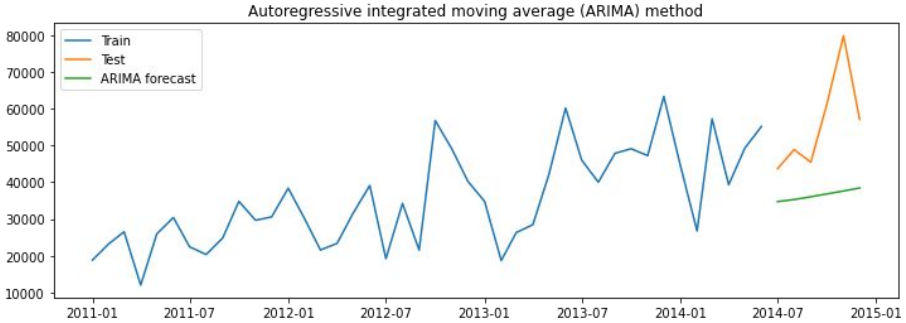
Auto Regression Moving Average



RMSE	MAPE
22654.32	32.40

Due to combination of MA with AR to give ARMA, the model has not been able to capture the level well. This model has performed poorly.

Auto regressive integrated moving average

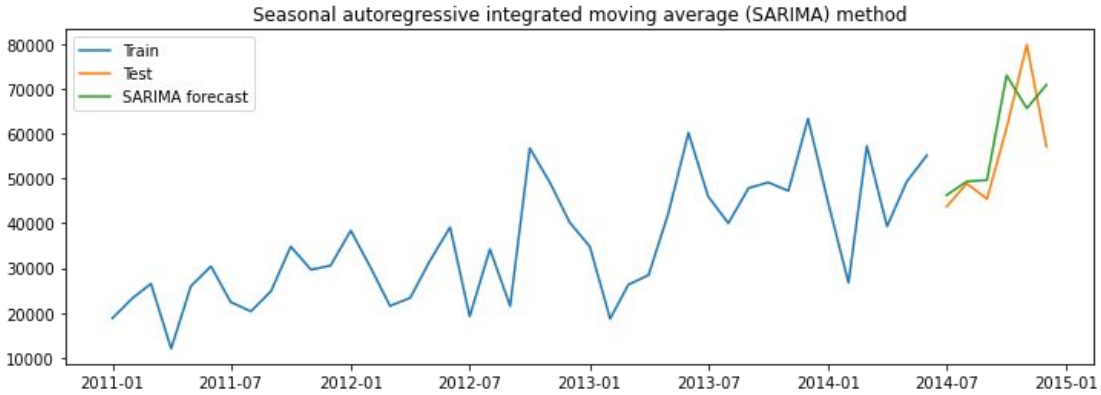


RMSE	MAPE
22654.32	32.40

ARIMA is similar to ARMA with the difference being that differencing of time series is performed by ARIMA model itself. This model has same poor performance as that of ARMA.

ARIMA Models

Seasonal Autoregressive integrated moving average



RMSE	MAPE
9616.49	12.87

SARIMA is Seasonal ARIMA. The level, trend and seasonal components have been captured well. This model has performed best among ARIMA models because it captured seasonal component well.

Findings & Conclusion

- We tested different Exponential Models and ARIMA models on the Retail Giant Sales dataset.
- Among exponential models, **Holt Winters' additive model** performed the best as seasonal component is captured in this model and was captured well.
- Among ARIMA models, **SARIMA model** performed the best as seasonal component is captured in this model and was captured well.
- All in all, even if trend and level components were captured by some models, **the capture of seasonal component decreased the RMSE and MAPE to a great extent.**