# DEMAND FORECASTING FOR E-COMMERCE

-- Presented By

ARANYA PAL

19.07.2024

# PROBLEM STATEMENT

- **Enhanced Marketing Efficiency :** Identify periods of high demand for targeted marketing campaigns, optimizing resource allocation.

- **Data-Driven Decision Making :** Reliable forecasts provide a basis for business decisions, such as pricing adjustments or product promotions.

- **Accurate Demand Predictions :** Implement a forecasting model that achieves high accuracy in predicting future demands thereby improving customer service levels.

# CONTENT

# INTRODUCTION

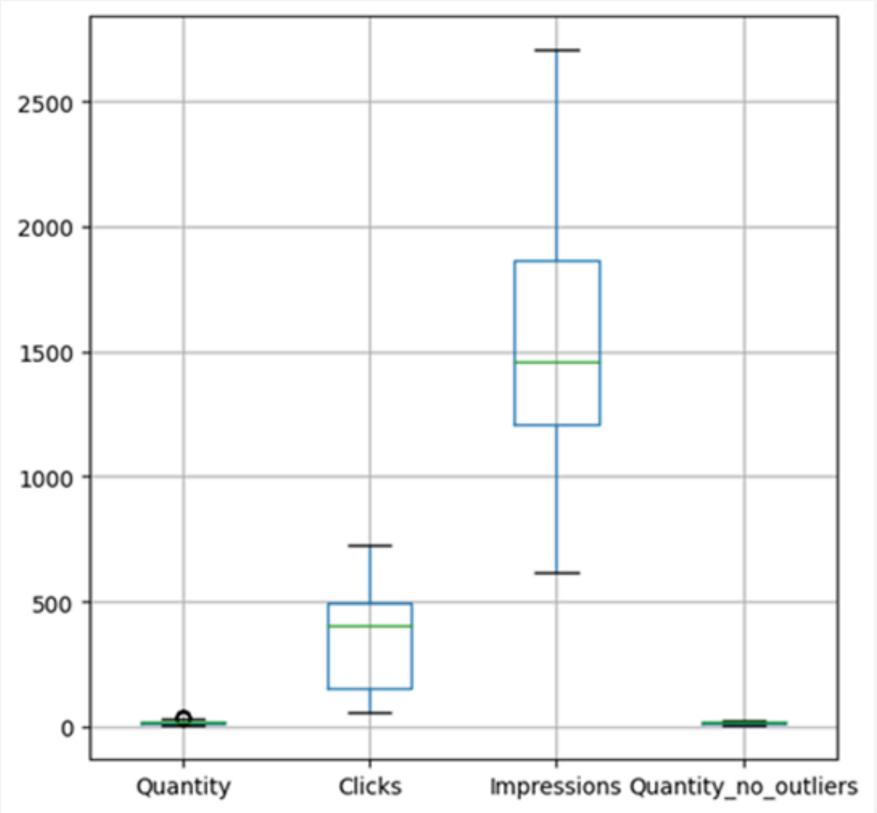
The primary goal of this project is to accurately forecast product demand, which is crucial for inventory management, sales strategy, and overall business planning.

The dataset used in this project includes the below features help to capture the influence of marketing activities on product demand.

- **Sales Data :** Contains daily sales quantity of a product.
- **Google Clicks :** Daily click data from Google
- **Facebook Impressions :** Daily impression data from Facebook.
- **Total Records :** 212
- **Time Span :** Covers 7 months to capture trends and seasonality.
- **Granularity :** Daily

# EXPLORATORY DATA ANALYSIS

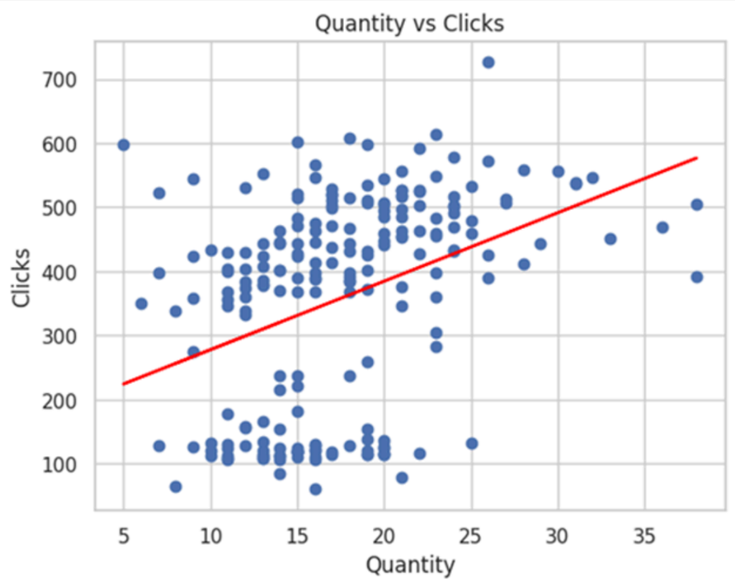
- In the dataset no missing value is found.
- There is no duplicate value in the dataset.
- There is a strong positive correlation (0.38) between Quantity and Clicks.
- Impressions also have slight relation to Quantity.
- We performed Feature Engineering to include some additional features like Day of the week (Monday,Tuesday etc.), Months (January,February etc.), IsWeekend, IsWeekdays etc.
- Outliers are detected in the dependent column(Quantity). They are replaced with 95th Percentile.




The boxplot shows that ‘New_Quantity’ doesn’t contain any value with outliers.

# VISUALIZATION

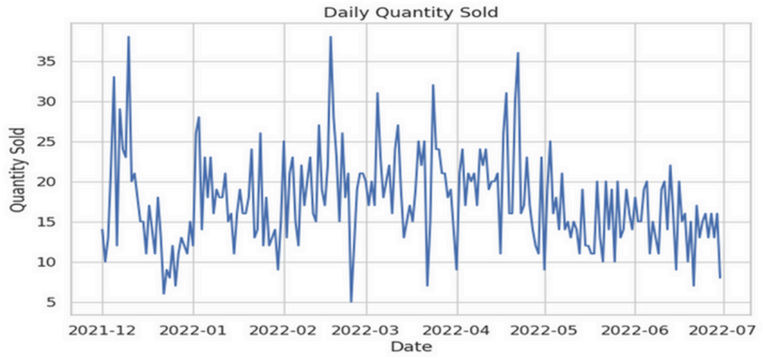
## Linear Regression Analysis





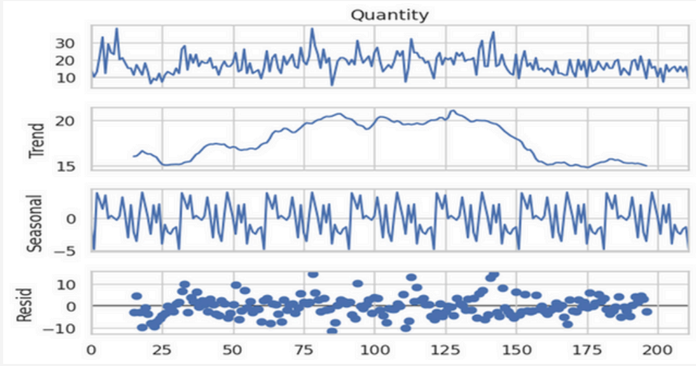
- The analysis between the number of clicks and the quantity sold suggests that increasing the number of clicks can significantly boost sales.
- The analysis between impressions and the quantity sold indicates a positive but relatively weak relationship.

## Daily Quantity Sold




- The plot reveals the daily fluctuations in the quantity of the items sold over the period from December 2021 to July 2022.
- Peaks and troughs are visible throughout the period, suggesting potential seasonality or other cyclical patterns

## Seasonality Test

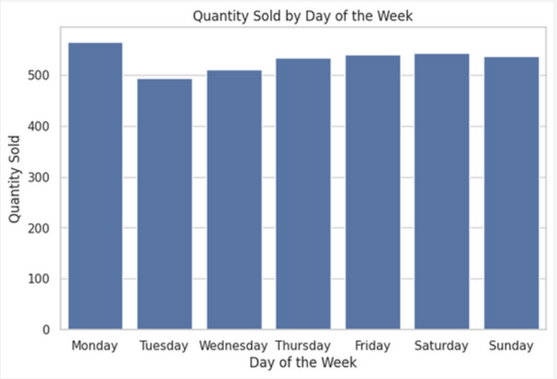



- The decomposition highlights the presence of a clear trend and seasonal patterns in the data, with cyclical monthly fluctuations.

## Monthly Quantity Sold


- March month has the highest and June has the lowest sold quantity.

## Quantity Sold by Days of Week



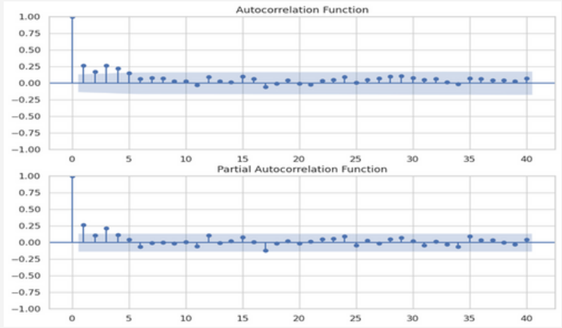

- Monday has the highest and Tuesday has the lowest sale.

# TIME SERIES HYPOTHESIS

## The Augmented Dickey-Fuller (ADF) to check Stationarity

- A stationary series has a constant mean and variance over time, making it easier to model and forecast.
- In the result of the test, p-value : 0.0002. So, sales quantity data is stationary.

## Autocorrelation and Partial Autocorrelation Function




- The ACF plot shows a significant spike at lag 1, followed by smaller spikes that die off quickly.
- The PACF plot cuts off after lag 1.

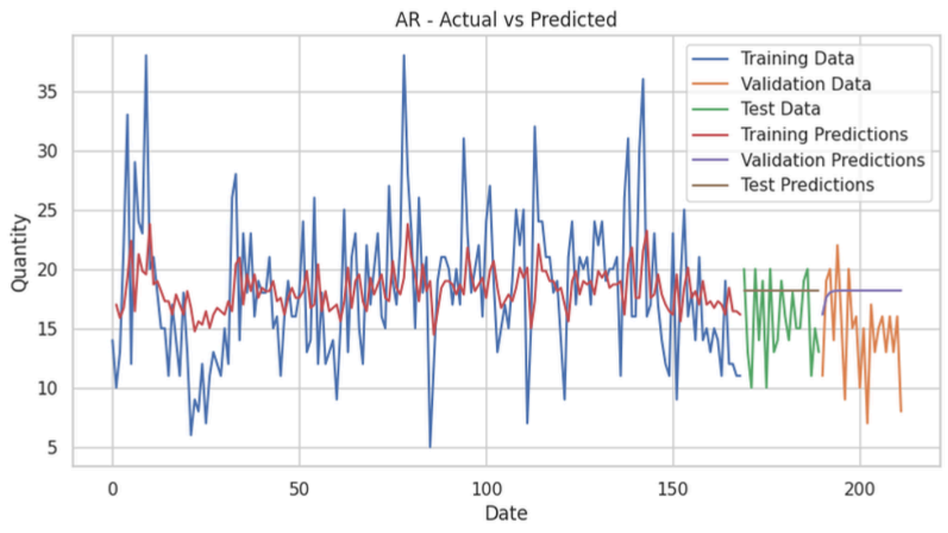
## Determined Parameters

- p = 1 : There are one autoregressive terms.
- d = 0 : The data is already stationary, so no differencing is needed.
- q = 1 : There is one moving average term.

# TIME SERIES MODEL

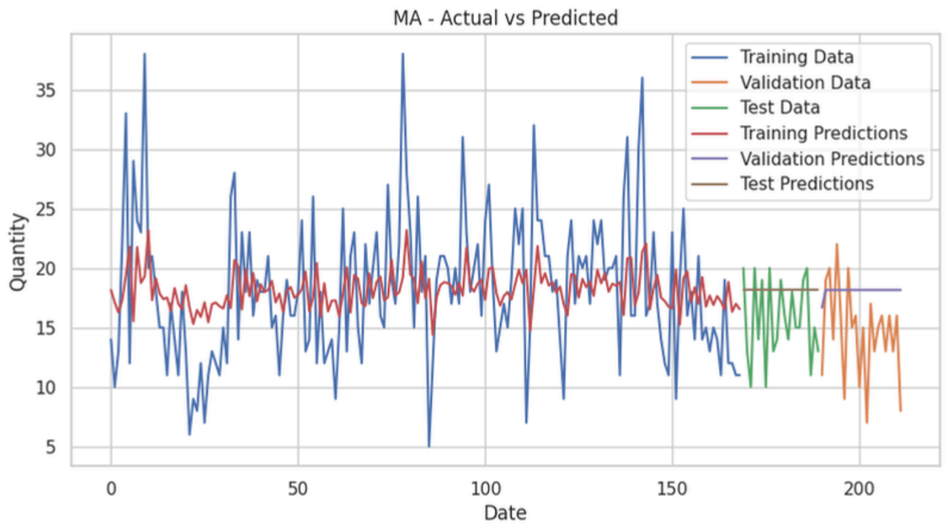
- To forecast the sales quantity, we have used various time series models including AR, MA, ARIMA, SARIMA, ARIMAX, and SARIMAX, as well as multivariate regression models.
- Performance metrics such as $R^2$, adjusted $R^2$, MAPE, RMSE, and MAE were used to evaluate and compare the models' accuracy.

## AR [Autoregressive] Model :



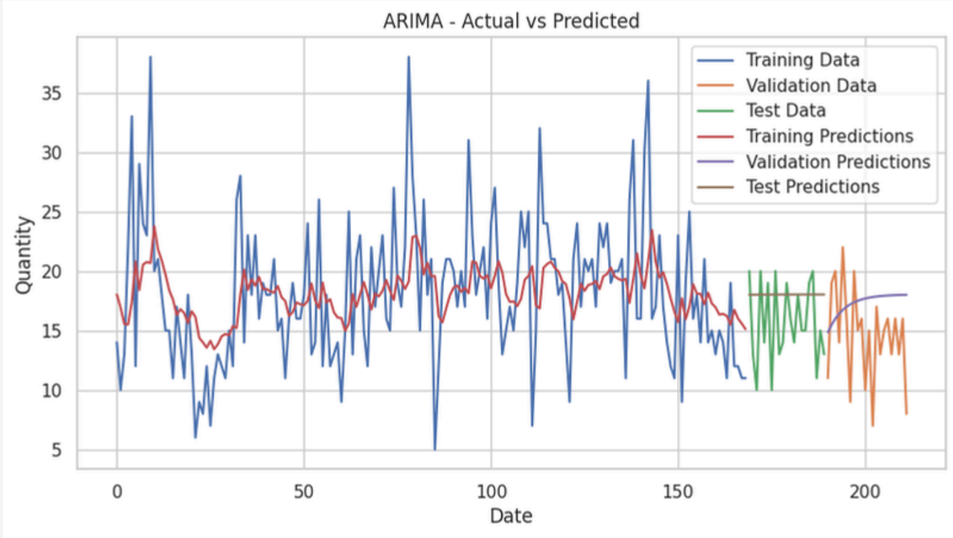

| Error Metrics | TEST SET | VALIDATION SET |
|---|---|---|
| RMSE | 4.19 | 5.15 |
| MAE | 3.49 | 4.28 |
| MAPE | 26.94% | 38.61% |
| R2 | -0.6 | -0.81 |
| Adjusted R2 | 6.33 | 8.58 |

## MA [Moving Average] Model :



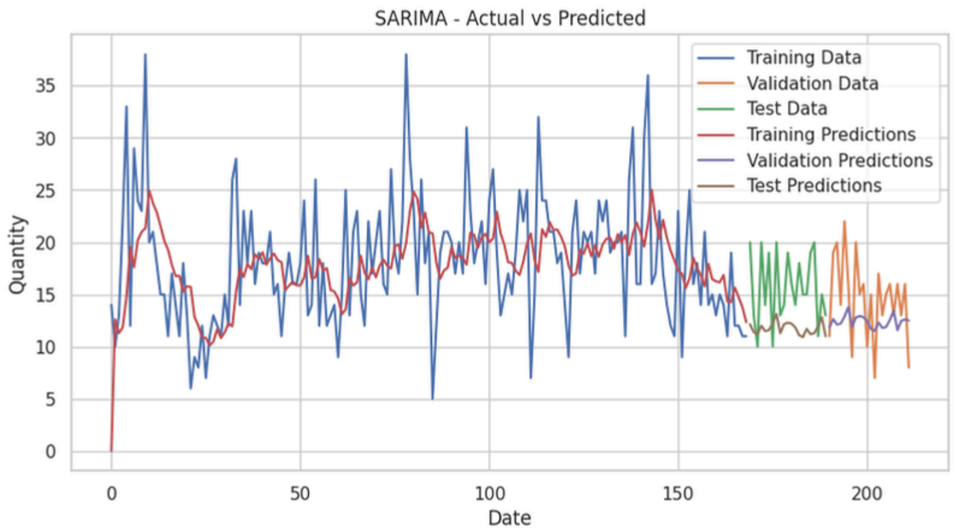

| Error Metrics | TEST SET | VALIDATION SET |
|---|---|---|
| RMSE | 4.17 | 5.15 |
| MAE | 3.48 | 4.26 |
| MAPE | 26.84% | 38.51% |
| R2 | -0.58 | -0.8 |
| Adjusted R2 | 6.28 | 8.58 |

## ARIMA [Autoregressive Integrated Moving Average] Model :



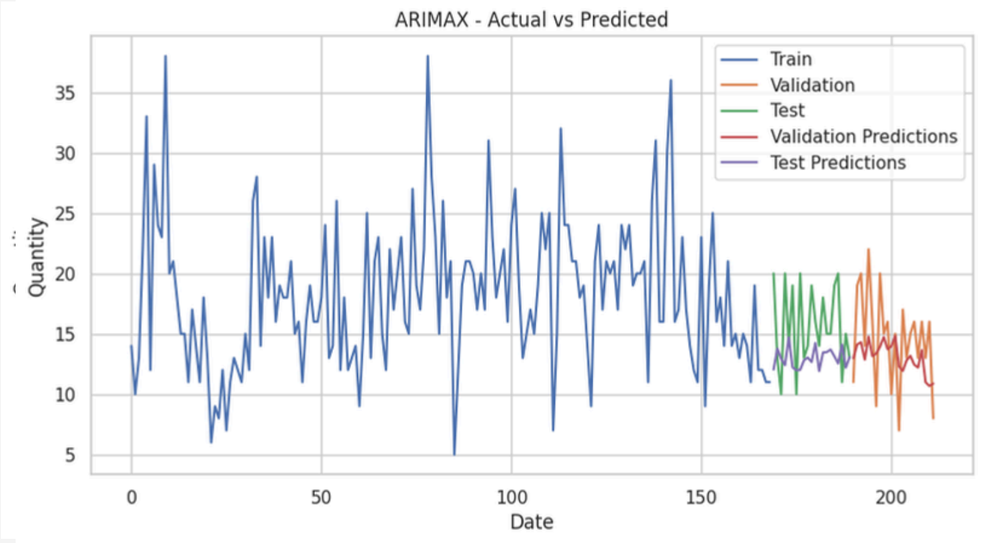

| Error Metrics | TEST SET | VALIDATION SET |
|---|---|---|
| RMSE | 4.1 | 4.95 |
| MAE | 3.44 | 4.13 |
| MAPE | 26.41% | 36.47% |
| R2 | -0.52 | -0.67 |
| Adjusted R2 | 6.08 | 8 |

## SARIMA [Seasonal ARIMA] Model :



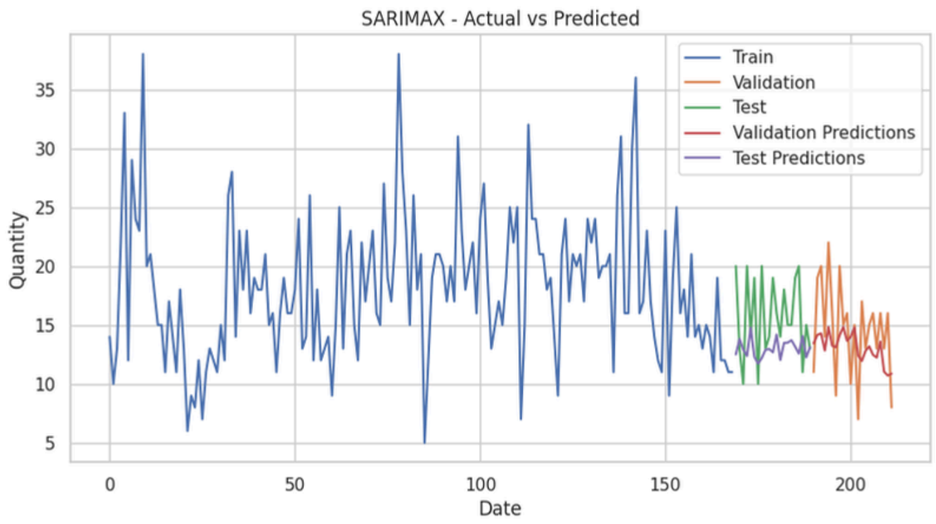

| Error Metrics | TEST SET | VALIDATION SET |
|---|---|---|
| RMSE | 5.05 | 4.28 |
| MAE | 4.28 | 3.62 |
| MAPE | 25.21% | 25.15% |
| R2 | -1.32 | -0.25 |
| Adjusted R2 | 8.74 | 6.24 |

## ARIMAX [ARIMA with Exogenous Variables] Model :



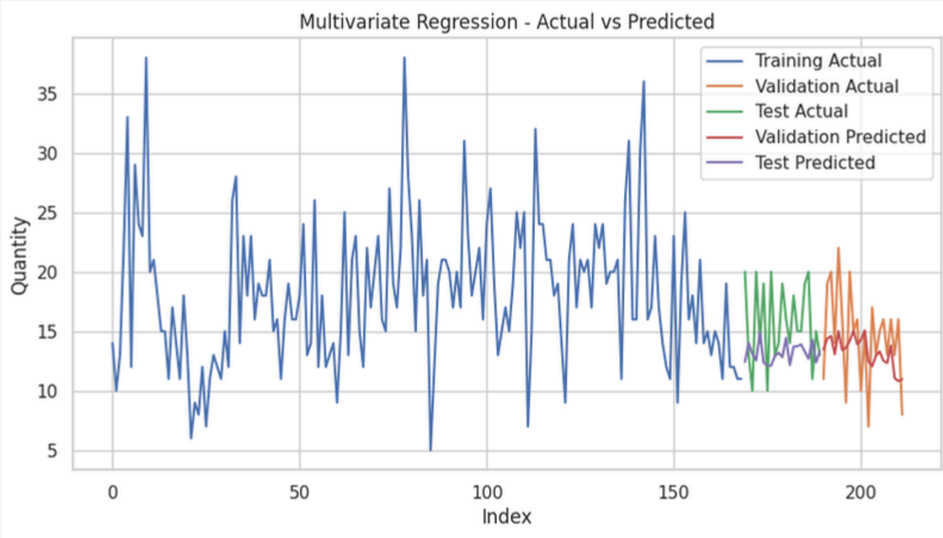

| Error Metrics | TEST SET | VALIDATION SET |
|---|---|---|
| RMSE | 4.5 | 3.79 |
| MAE | 3.56 | 3.19 |
| MAPE | 21.00% | 23.35% |
| R2 | -0.84 | 0.02 |
| Adjusted R2 | -2.35 | -0.71 |

## SARIMAX [Seasonal ARIMA with Exogenous Variables] Model :




| Error Metrics | TEST SET | VALIDATION SET |
|---|---|---|
| RMSE | 4.43 | 3.77 |
| MAE | 3.51 | 3.19 |
| MAPE | 20.64% | 23.38% |
| R2 | -0.78 | 0.03 |
| Adjusted R2 | -2.24 | -0.69 |

## Multivariate Linear Regression Model :




| Error Metrics | TEST SET | VALIDATION SET |
|---|---|---|
| RMSE | 4.39 | 3.72 |
| MAE | 3.48 | 3.11 |
| MAPE | 20.65% | 23.14% |
| R2 | -0.75 | 0.06 |
| Adjusted R2 | -2.19 | -0.65 |

# COMPARISON TABLE

## TEST SET

| Error Metrics | AR | MA | ARIMA | SARIMA | ARIMAX | SARIMAX | Multivariate Linear Regression |
|---|---|---|---|---|---|---|---|
| RMSE | 4.19 | 4.17 | 4.1 | 5.05 | 4.5 | 4.43 | 4.39 |
| MAE | 3.49 | 3.48 | 3.44 | 4.28 | 3.56 | 3.51 | 3.48 |
| MAPE | 26.94% | 26.84% | 26.41% | 25.21% | 21.00% | 20.64% | 20.65% |
| R2 | -0.6 | -0.58 | -0.52 | -1.32 | -0.84 | -0.78 | -0.75 |
| Adjusted R2 | 6.33 | 6.28 | 6.08 | 8.74 | -2.35 | -2.24 | -2.19 |

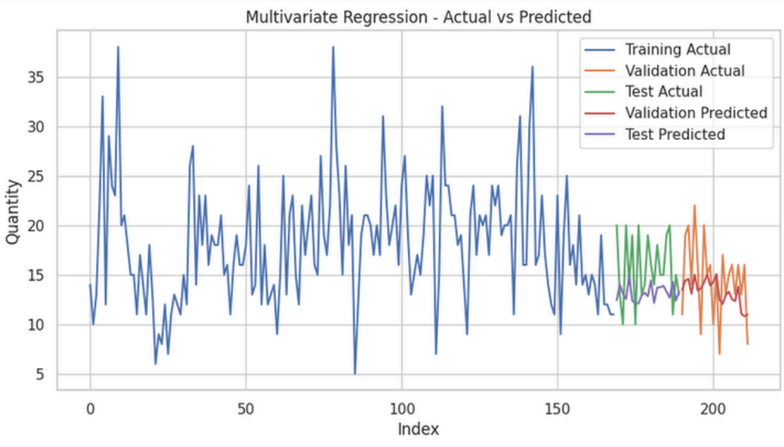
## VALIDATION SET

| Error Metrics | AR | MA | ARIMA | SARIMA | ARIMAX | SARIMAX | Multivariate Linear Regression |
|---|---|---|---|---|---|---|---|
| RMSE | 5.15 | 5.15 | 4.95 | 4.28 | 3.79 | 3.77 | 3.72 |
| MAE | 4.28 | 4.26 | 4.13 | 3.62 | 3.19 | 3.19 | 3.11 |
| MAPE | 38.61% | 38.51% | 36.47% | 25.15% | 23.35% | 23.38% | 23.14% |
| R2 | -0.81 | -0.8 | -0.67 | -0.25 | 0.02 | 0.03 | 0.06 |
| Adjusted R2 | 8.58 | 8.58 | 8 | 6.24 | -0.71 | -0.69 | -0.65 |

- We have negative R2 values, because of the reason :
  - We have a very small amount of data of 7 months.
  - There are only two independent variables (Clicks and Impressions) which are not able to capture the variance of the target variable.

# CONCLUSION

- Based on the test and validation results, the **Multivariate Linear Regression** model appears to be the best choice for demand forecasting. The model demonstrates the MAE (3.11) and RMSE (3.72) on the validation set, as well as the lowest MAPE (23.14%).
- The group of features used in the model : 'Clicks', 'Impressions', 'Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday'.




| Error Metrics | TEST SET | VALIDATION SET |
|---|---|---|
| RMSE | 4.39 | 3.72 |
| MAE | 3.48 | 3.11 |
| MAPE | 20.65% | 23.14% |
| R2 | -0.75 | 0.06 |
| Adjusted R2 | -2.19 | -0.65 |

The chosen model provides a scalable solution by efficiently handling large datasets, adapting to increasing data volumes, and maintaining prediction accuracy, making it suitable for dynamic and growing E-commerce environments.