FUTURECART:AI-DRIVEN DEMAND PREDICTION FOR SMARTER RETAIL

Presented by:
Nitin Manohar Mishra
Chris Joy
Sri Harshini
Swastik Roy Choudhury

INTRODUCTION TO THE PROJECT

In today's dynamic retail landscape, staying ahead of demand is the key to success. Traditional demand forecasting methods fail to capture complex patterns influenced by online engagement and seasonal trends.

This project uses AI-powered time series analysis and multivariate regression to forecast product demand. It combines historical sales data with online activity metrics like clicks and impressions in an attempt to enable companies to better plan their inventories, market more smartly, and improve experiences for their customers.

PROBLEM STATEMENT

In the dynamic landscape of e-commerce, accurate demand forecasting is critical for success. Retail businesses often struggle with managing inventory, optimizing marketing strategies, and meeting customer demands due to unpredictable sales patterns and external influences.

This project focuses on developing an AI-driven demand forecasting model leveraging historical sales data and key external metrics (e.g., Google clicks, Facebook impressions).

The goal is to improve inventory management, enhance marketing efficiency, and support data-driven decision-making to meet customer expectations effectively.

OBJECTIVES

SMARTER DEMAND INSIGHTS

Utilize AI models to accurately forecast demand trends using past sales and digital engagement data.

INSIGHTFUL BUSINESS DECISIONS

Drive confident and informed decision-making through data-backed predictions and analysis.

TARGETED CAMPAIGN PLANNING

Leverage demand patterns to craft impactful marketing strategies during peak periods.

STREAMLINED STOCK MANAGEMENT

Maintain optimal inventory levels by predicting demand fluctuations and avoiding inefficiencies.

METHODOLOGY

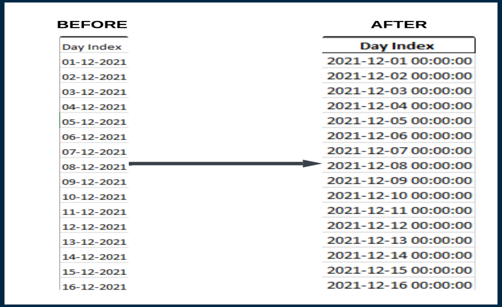
1. Data Collection
2. EDA and Data Processing
3. Timeseries Modelling
4. Multivariate Regression
5. Model Evaluation, Selection & Forecasting

DATA SOURCES AND OVERVIEW

1. Data Description:The dataset used includes historical sales data (target variable: Quantity) and external factors such as Clicks and Impressions (from Google Analytics and social media).
2. Data Structure: Quantity (sales data) has been assigned as Target Variable while Clicks, Impressions have been assigned as Exogenous Variables since these are the external factors that have influence over demand of the product.
3. Data Quality: Challenges included handling missing values and data cleaning.

PART 1 : DATA PRE PROCESSING

STEP 1: Convert the 'Day Index' column to a proper datetime format
Converting the 'Day Index' column to a proper datetime format ensures consistency, accuracy, and compatibility with date-based operations. It allows for easier analysis, time-based aggregations, and smooth data merging or filtering, while preventing errors or mismatches due to incorrect date formats

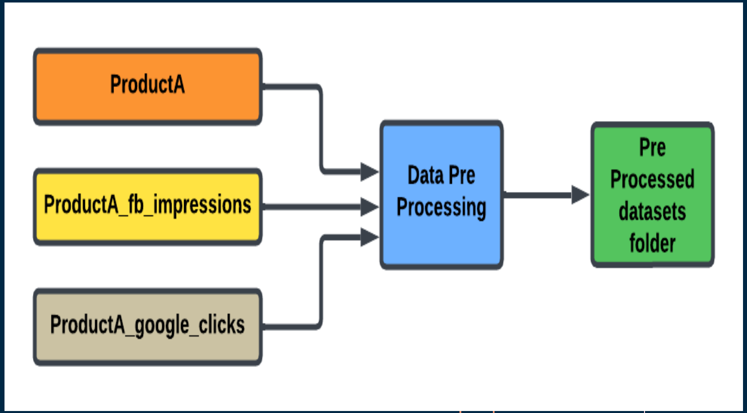


BEFORE	AFTER
Day Index	Day Index
01-12-2021	2021-12-01 00:00:00
02-12-2021	2021-12-02 00:00:00
03-12-2021	2021-12-03 00:00:00
04-12-2021	2021-12-04 00:00:00
05-12-2021	2021-12-05 00:00:00
06-12-2021	2021-12-06 00:00:00
07-12-2021	2021-12-07 00:00:00
08-12-2021	2021-12-08 00:00:00
09-12-2021	2021-12-09 00:00:00
10-12-2021	2021-12-10 00:00:00
11-12-2021	2021-12-11 00:00:00
12-12-2021	2021-12-12 00:00:00
13-12-2021	2021-12-13 00:00:00
14-12-2021	2021-12-14 00:00:00
15-12-2021	2021-12-15 00:00:00
16-12-2021	2021-12-16 00:00:00

STEP 2: Remove duplicate rows
There were no duplicate rows in the dataset, so this step was not necessary.

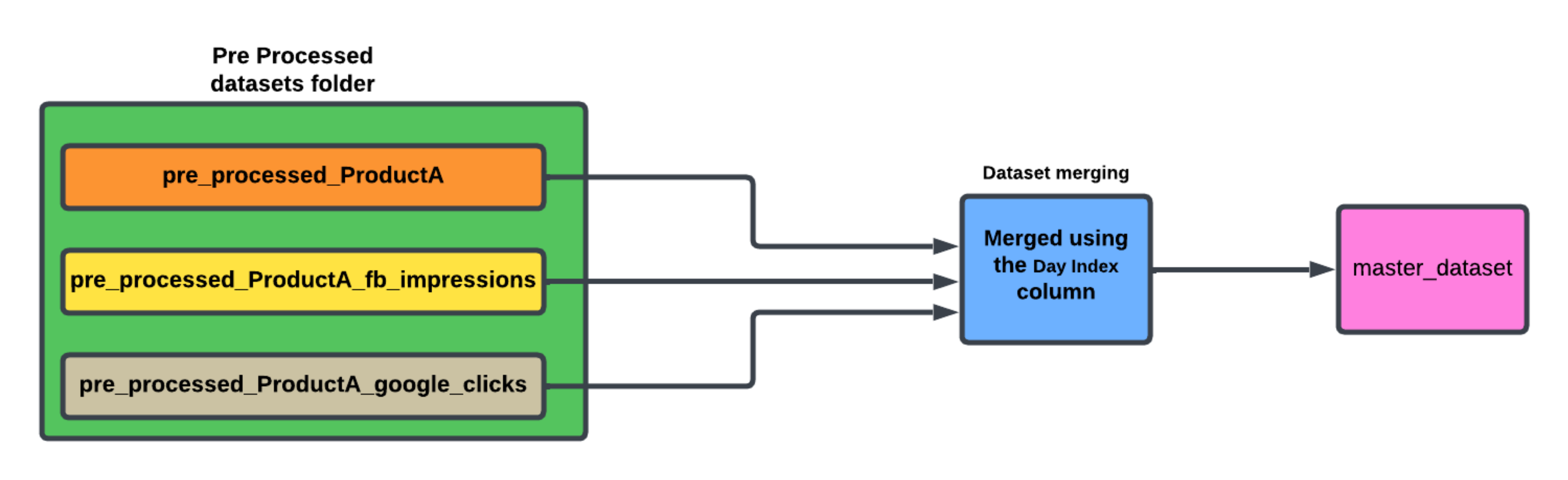
STEP 3: Fill any missing values using forward fill
There were no missing values in the dataset, so this step was not needed.

FINAL STEP :Output Data
Cleaned datasets are saved into the pre_processed_datasets folder, ready for merging.



PART 2 : DATASET MERGING

The process involves loading preprocessed datasets from the pre_processed_datasets folder, merging them based on the common column Day Index, and saving the final merged dataset to the master_dataset folder.

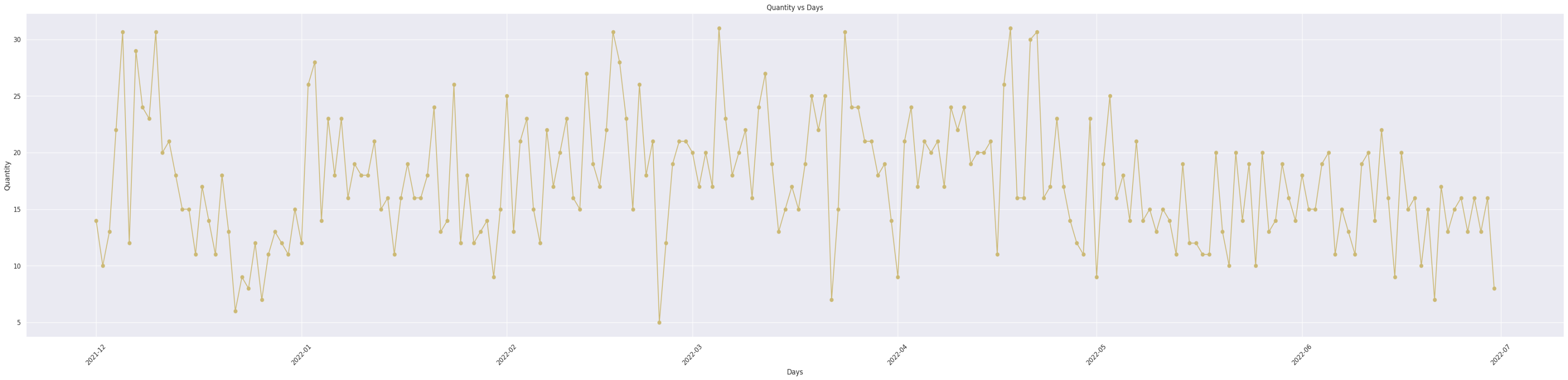


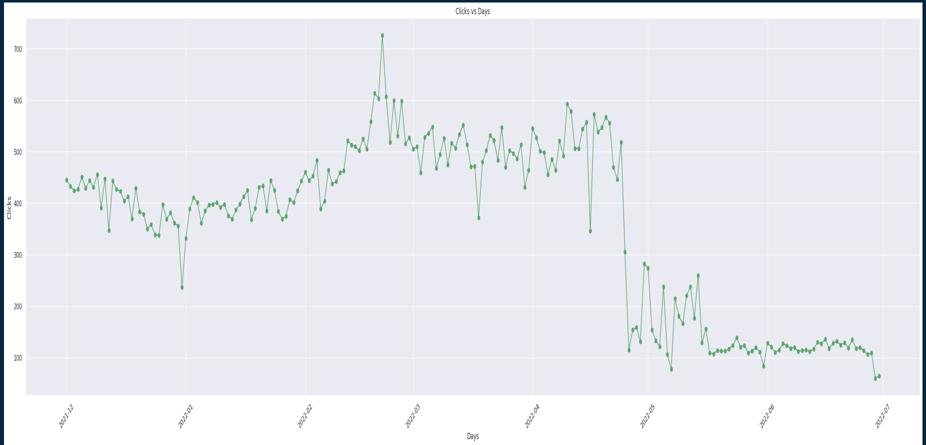
PART 3: EXPLORATORY DATA ANALYSIS

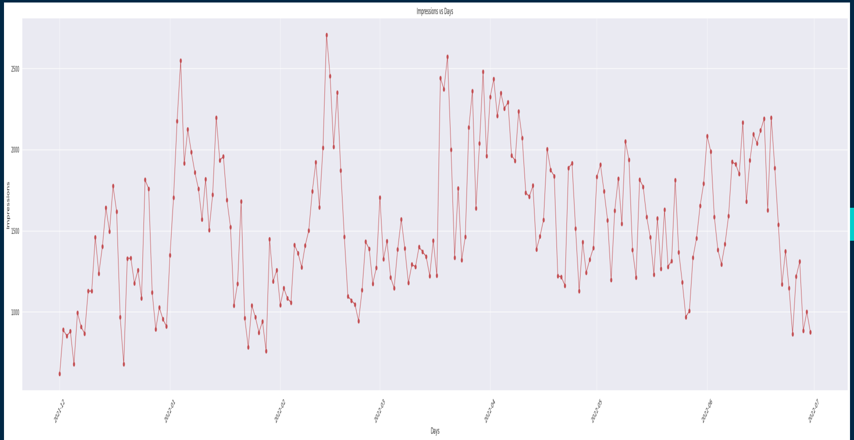
Why we need this code: The code is essential for analyzing the Master Dataset, as it transforms raw data into actionable insights. It helps uncover trends, correlations, and key metrics crucial for making informed decisions in the project.

Purpose of this code:

- Data Cleaning: The code addresses missing values and infinities, ensuring that calculations and visualizations are robust and reliable.
- Visualizations: A series of plots and charts are generated to highlight relationships, growth patterns, and insights in the data, making it easier to identify trends over time.
- Feature Engineering : It adds additional columns to the dataset to get out more information from the data we have .This helps in timeseries modelling, regression and finally in forecasting to understand seasonality of the data.
- Key Takeaways and conclusion: This code allows us to analyze and visualize key performance metrics, track trends over time, and identify factors influencing sales. It is a crucial step in understanding data, making it more actionable, and supporting better decision-making for demand prediction and business planning.


Quantity vs Days
Quantity
Days
2021-12
2022-01
2022-02
2022-03
2022-04
2022-05
2022-06
2022-07


Clicks vs Days
Clicks
Days


Impressions vs Days
Impressions
Days

PART 4: TIME SERIES ANALYSIS

1. Auto Regressive:

- Predicts future values based on a linear combination of its past values.
- Captures temporal dependencies in the dataset, focusing on lagged relationships.

2. Moving Average:

- Models the error terms (residuals) from past predictions rather than the actual values.
- Useful for smoothing short-term fluctuations and identifying trends.

3. ARIMA(Autoregressive Integrated Moving Average):

- Combines AR and MA techniques, adding a differencing step to handle non-stationary data.
- Ideal for datasets with trends but no significant seasonality.

4. SARIMA (Seasonal ARIMA):

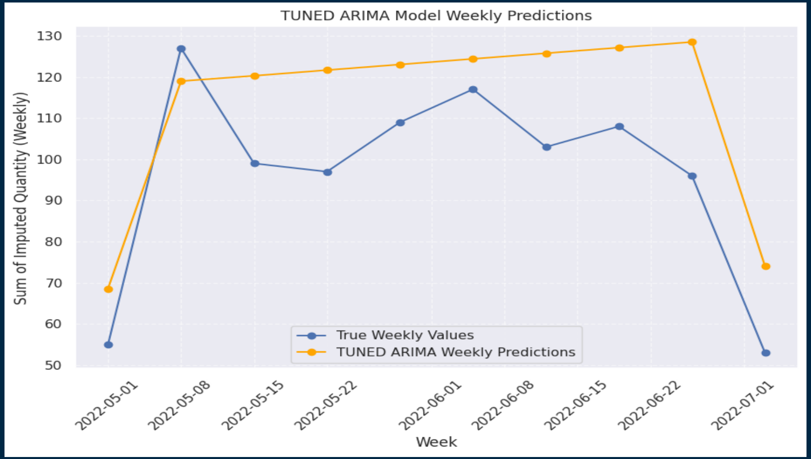
- Extends ARIMA by incorporating seasonal components to address repeating patterns over fixed intervals.
- Suitable for data with clear seasonal variations, such as monthly or quarterly sales.

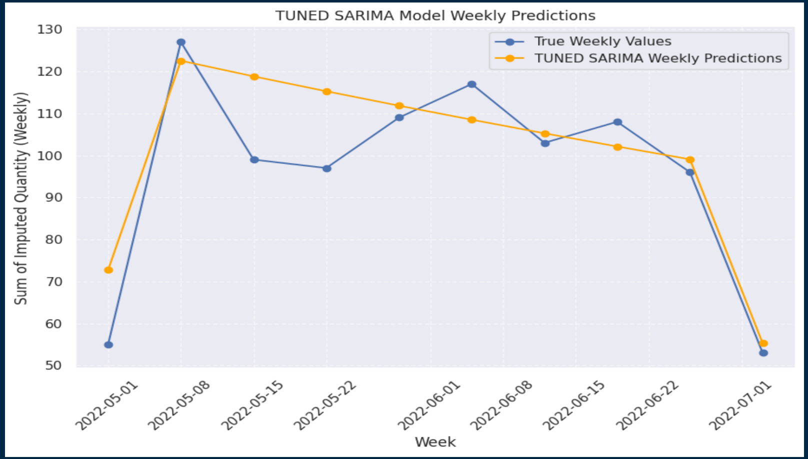
5. ARIMAX (ARIMA with Exogenous Variables):

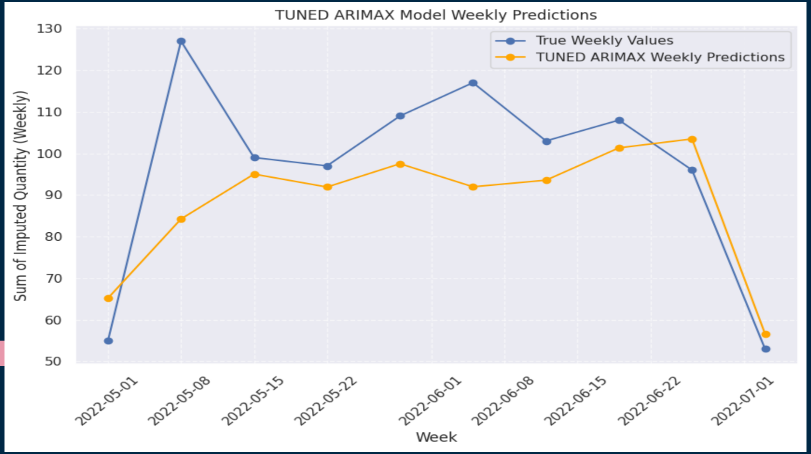
- Enhances ARIMA by including external predictors (e.g., clicks, impressions) to improve accuracy.
- Allows the model to account for external factors influencing the target variable.

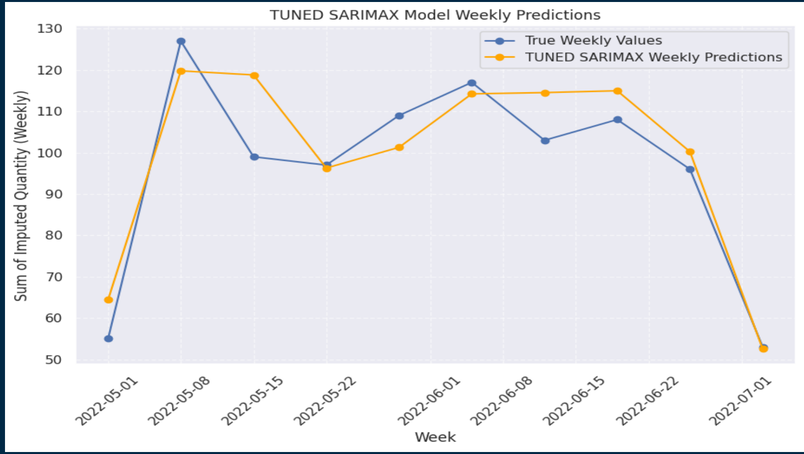
3. SARIMAX (Seasonal ARIMA with Exogenous Variables):

- Combines the seasonal capabilities of SARIMA with the flexibility of external inputs.
- Provides a holistic approach for forecasting when external variables and seasonality are critical.


TUNED ARIMA Model Weekly Predictions
True Weekly Values
TUNED ARIMA Weekly Predictions
Sum of Imputed Quantity (Weekly)
Week
50 60 70 80 90 100 110 120 130
2022-05-01 2022-05-08 2022-05-15 2022-05-22 2022-06-01 2022-06-08 2022-06-15 2022-06-22 2022-07-01


TUNED SARIMA Model Weekly Predictions
True Weekly Values
TUNED SARIMA Weekly Predictions
Sum of Imputed Quantity (Weekly)
Week
50 60 70 80 90 100 110 120 130
2022-05-01 2022-05-08 2022-05-15 2022-05-22 2022-06-01 2022-06-08 2022-06-15 2022-06-22 2022-07-01


TUNED ARIMAX Model Weekly Predictions
True Weekly Values
TUNED ARIMAX Weekly Predictions
Sum of Imputed Quantity (Weekly)
Week
50 60 70 80 90 100 110 120 130
2022-05-01 2022-05-08 2022-05-15 2022-05-22 2022-06-01 2022-06-08 2022-06-15 2022-06-22 2022-07-01


TUNED SARIMAX Model Weekly Predictions
True Weekly Values
TUNED SARIMAX Weekly Predictions
Sum of Imputed Quantity (Weekly)
Week
50 60 70 80 90 100 110 120 130
2022-05-01 2022-05-08 2022-05-15 2022-05-22 2022-06-01 2022-06-08 2022-06-15 2022-06-22 2022-07-01

EVALUATION METRICES

Metrics Before Tuning

	Model	MAE	RMSE	MSE	MAPE
0	AR	4.286008	5.076281	25.768626	35.395913
1	MA	4.299021	5.094176	25.950630	35.521704
2	ARIMA	5.219194	6.178560	38.174599	43.353707
3	SARIMA	7.220430	8.309928	69.054906	57.901700
4	ARIMAX	6.909809	7.888308	62.225402	43.096618
5	SARIMAX	12.855971	14.161707	200.553951	98.471875

Evaluation Metrics After Hyperparameter Tuning

	Model	MAE	RMSE	MSE	MAPE
0	AR	4.286008	5.076281	25.768626	35.395913
1	MA	4.299021	5.094176	25.950630	35.521704
2	ARIMA	5.219194	6.178560	38.174599	43.353707
3	SARIMA	7.220430	8.309928	69.054906	57.901700
4	ARIMAX	3.600522	4.424602	19.577106	25.414755
5	SARIMAX	3.482396	4.315564	18.624092	25.192728

INSIGHTS AND CONCLUSION

After evaluating all the models, the SARIMAX model emerged as the most suitable choice for this project:

Seasonality Capture:It effectively captures seasonality, essential for predicting recurring demand patterns.

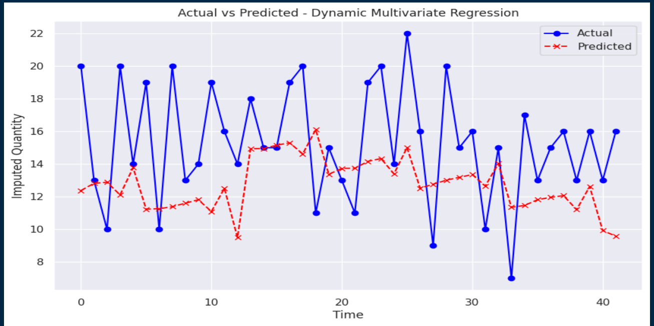
Incorporation of Exogenous Variables: It integrates key external factors like Clicks and Impressions, which influence the target variable (Quantity).

Superior Accuracy: Compared to simpler models (AR, MA), SARIMAX consistently showed higher accuracy and a more comprehensive data understanding.

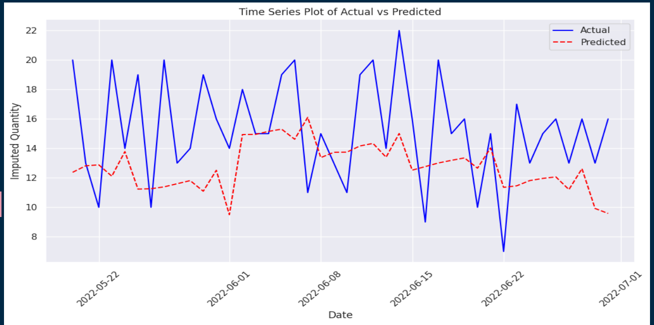
Conclusion: SARIMAX meets the project's forecasting needs, making it the recommended approach for reliable time series analysis and demand prediction.

Final Step: Splitting Data and Running Multivariate Regression with Visualization

Hyperparameter Tuning: The multivariate regression model's hyperparameters were fine-tuned using **GridSearchCV** to optimize parameters such as **fit_intercept** and **copy_X.** After testing multiple configurations, the best model was identified with parameters **{'copy_X': True, 'fit_intercept': True}.** The evaluation metrics showed an **RMSE of 30.13** and an **R-squared value of 0.17**.

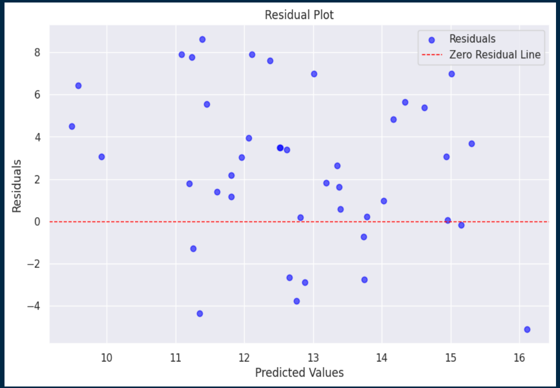


The multivariate regression model was used to predict daily values based on the tuned parameters. Key Insight: The model captures daily fluctuations but struggles with significant spikes, leading to slight discrepancies between predicted and actual values.

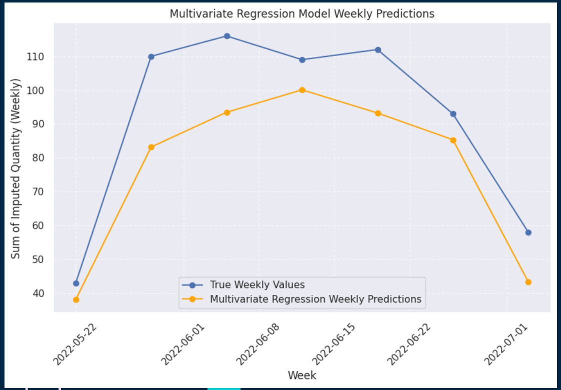


Residuals were calculated by subtracting the predicted values from the actual values and plotted to assess the model's accuracy. The analysis revealed minimal systematic errors and no significant biases, highlighting the robustness of the model.

Final Step: Splitting Data and Running Multivariate Regression with Visualization



The residuals plot was used to check for randomness between actual and predicted values. Key Insight: Residuals show minor biases, but no significant outliers, suggesting that the model largely captures patterns.

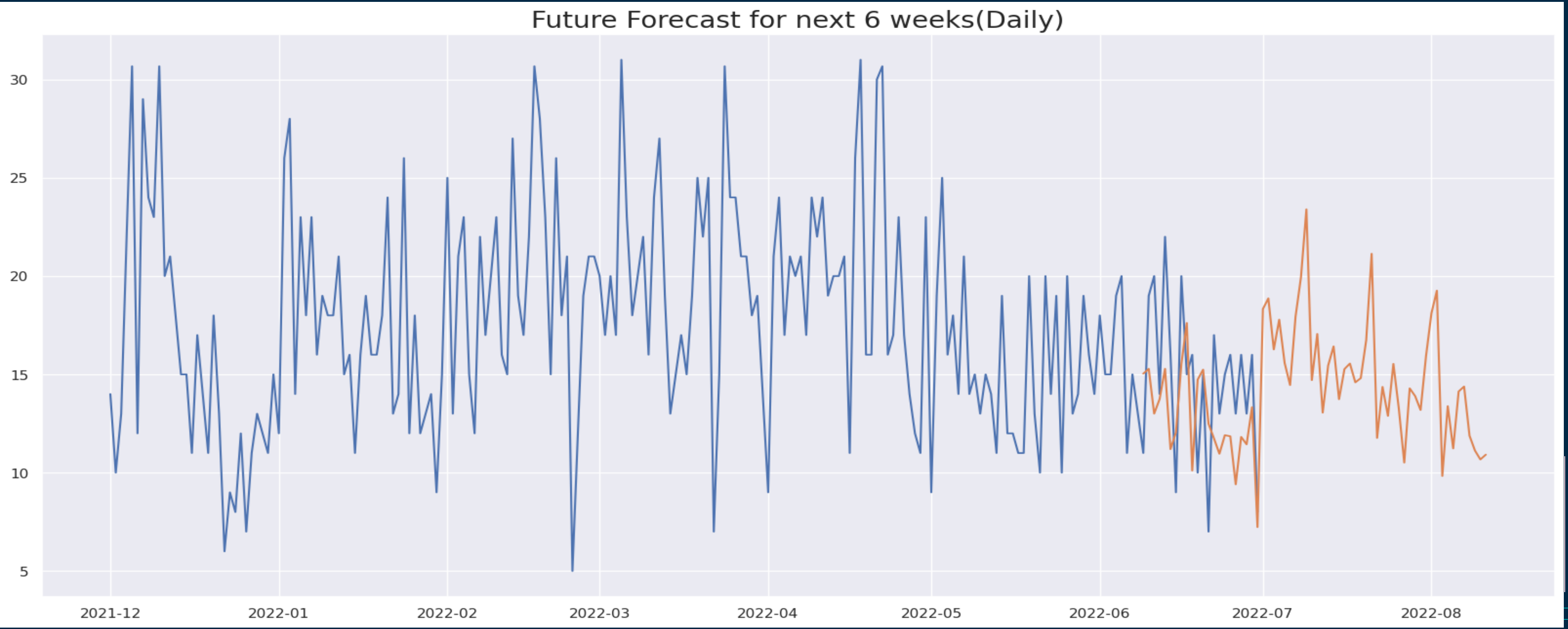


The model was used to generate predictions for the next six months based on simulated future values of exogenous variables (Clicks and Impressions).

Key Insight: The model forecasts a significant increase in the predicted quantity during the initial months, followed by a stabilization. While the trend generally follows expected patterns, further refinement is needed for more precise forecasting.

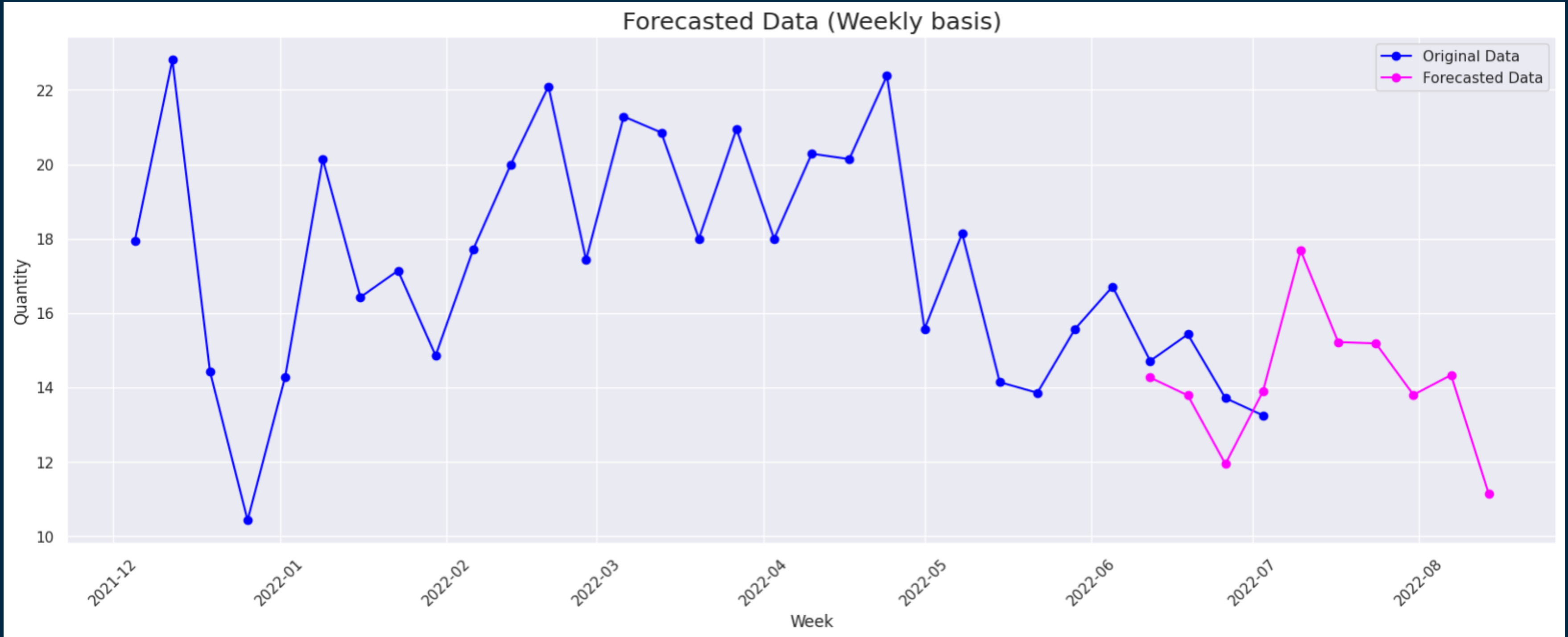
FORECASTING(SARIMAX MODEL)

- DAILY REPRESENTATION



FORECASTING(SARIMAX MODEL)

- WEEKLY REPRESENTATION



THANK YOU