FutureCart:

AI-Driven Demand Prediction for Smarter Retail.

Name-Swastik Roy Choudhury
B.Tech | ECE
4th Year
2024-25

Contents

TOPICS	Page No.
1)Project Overview	3
2)Problem Statement	3
3)Outcomes	3-4
4)Data Collection	4-5
5)Exploratory Data Analysis(EDA)	5
6)Data Preprocessing	5-6
7)General Recommendations	6-7
8)About Time Series Modelling	7-12
◇ Model selection	
◇ Model Fitting	
◇ Model Evaluation	
◇ Model Diagnostic	
9)Time series Analysis	12-13
10)Time Series Modeling	13-14
11)Data Scaling	14
12)Model Implementations	14-22
◇ Autoregressive(AR)Model	
◇ Moving Average(MA) Model	
◇ ARIMA Model	
◇ ARIMAX Model	
◇ SARIMA Model	
◇ SARIMAX Model	
◇ Residual Plots	
13)Metrics Table	22
◇ Key Observations	
14)Multivariate Regression	23-26
◇ Data Preparation	
◇ Model Construction	
◇ Model Evaluation	
15)Forecasting	27-29
◇ The Need for Demand Forecasting	
◇ Model Development Process	
◇ Two-Month Forecasting	
◇ Summary of findings	
18)Challenges and Learnings	30
19)Conclusion	30
20)References	30

Project Overview

As the creator of this project, my objective is to develop a robust, AI-driven demand prediction model tailored for the retail sector. The dynamic nature of retail, characterized by fluctuating consumer behavior, seasonal trends, and external factors, necessitates a predictive solution that can provide actionable insights. This project leverages cutting-edge time series modeling techniques, enriched by comprehensive exploratory data analysis (EDA) and effective data preprocessing methods. By integrating advanced statistical and machine learning models, my aim is to deliver accurate demand forecasts that empower retailers to make data-driven decisions and optimize their operations.

This undertaking embodies a meticulous journey through data collection, transformation, model development, and evaluation. The final goal is to create a scalable and adaptable system that not only predicts future trends but also provides meaningful insights into consumer behavior. This system aspires to bridge the gap between raw data and actionable strategies, facilitating smarter inventory management, pricing strategies, and overall retail operations.

Problem Statement

Retailers face an ever-present challenge: predicting consumer demand accurately amidst the volatile and competitive nature of the market. Failure to do so can lead to overstocking, understocking, lost sales, and financial inefficiencies. Traditional forecasting techniques often fall short of addressing the nuances of modern retail dynamics, which are influenced by multifaceted variables including time-sensitive factors, external shocks, and consumer preferences.

This project seeks to address these issues by designing an intelligent demand prediction system that leverages AI and time series modeling. The primary objective is to harness historical sales and operational data to generate precise forecasts that retailers can trust. Unlike generic forecasting solutions, this project focuses on building a domain-specific model that adapts to the unique characteristics of retail data. By employing techniques such as ARIMA, SARIMA, and multivariate regression models, the system is tailored to capture patterns and anomalies in retail trends effectively.

Through this project, I aim to demonstrate the value of data-driven insights in enhancing operational efficiency, reducing waste, and driving profitability. This solution is not just about predicting numbers; it's about empowering businesses with foresight to adapt, innovate, and thrive in a competitive marketplace.

Outcome

The anticipated outcomes of this project encompass both tangible deliverables and broader strategic advantages. At its core, the project will produce a reliable demand forecasting system capable of accurately predicting retail demand for varying time horizons. This system will include a user-

friendly interface for visualizing trends, comparisons, and actionable insights, ensuring accessibility for non-technical stakeholders.
In addition to the technical deliverables, the project will result in several key outcomes:

- Enhanced Decision-Making: Retailers will gain the ability to anticipate demand fluctuations and make informed decisions on inventory, pricing, and promotions.

- Operational Efficiency: With precise forecasts, businesses can optimize supply chain operations, reducing costs associated with overstocking and understocking.

- Consumer Satisfaction: By aligning inventory with actual demand, retailers can improve customer experiences by minimizing stockouts and excess inventory.

- Scalability: The forecasting system will be designed to adapt to different retail scenarios, accommodating diverse product lines and seasonal trends.

- Insights into Consumer Behavior: Beyond forecasts, the analysis will uncover hidden trends and patterns, providing valuable context for strategic planning.

This project represents a significant step forward in demonstrating how AI and data analytics can transform the retail industry. By translating raw data into actionable intelligence, it lays the foundation for smarter, more adaptive retail practices that align with evolving consumer needs.

Data Collection

In the initial phase of my project, I focused on gathering high-quality, relevant data to lay a strong foundation for the analysis. I sourced data from credible repositories and ensured it encompassed key metrics such as sales, clicks, impressions, and time-related variables. The collection process prioritized completeness and accuracy, capturing a diverse range of scenarios to enhance model robustness. Additionally, I documented the data sources and acquisition methods meticulously to ensure transparency and reproducibility. This milestone was crucial in setting the stage for subsequent stages, ensuring the data was both representative and aligned with the project's objectives.

Libraries Used

1. IPython - Interactive computing.
2. concurrent - Concurrent task management.
3. google - Integration with Google services.
4. itertools - Efficient looping constructs.
5. matplotlib - Data visualization.
6. multiprocessing - Parallel task execution.
7. numpy - Numerical computations.

8. pandas - Data manipulation and analysis.
9. seaborn - Advanced data visualization.
10. sklearn - Machine learning algorithms.
11. statsmodels - Statistical modeling and time series analysis.
12. warnings - Handling and filtering warnings.

Exploratory Data Analysis (EDA) and Data Preprocessing

In the second week of my project, I concentrated on performing Exploratory Data Analysis (EDA) and preprocessing the data to prepare it for advanced modeling. This phase was instrumental in uncovering patterns, anomalies, and trends within the dataset, providing critical insights and informing the design of subsequent models. I adopted a structured approach, starting with understanding the dataset's structure, followed by cleaning and transforming the data to ensure its quality and usability.

Exploratory Data Analysis (EDA)

The primary goal of EDA was to explore the dataset comprehensively and derive meaningful insights. This involved visualizing distributions, identifying correlations, and examining time-series trends for key variables like sales quantities, clicks, and impressions. I utilized visualization tools such as histograms, box plots, and scatter plots to represent the data effectively, making it easier to spot patterns and irregularities.
EDA revealed several key aspects of the data. For instance, I identified seasonal trends and periodic spikes in demand, which were consistent with holidays and promotional campaigns. Additionally, I observed the interplay between clicks and impressions, providing valuable insights into consumer engagement and conversion rates. These findings served as a foundation for further analysis and guided the preprocessing steps.

Data Preprocessing

To ensure the dataset was suitable for modeling, I undertook several preprocessing steps:

- **Handling Missing Values**:
  Missing data can distort analysis, so I applied imputation techniques to address this issue. For numerical variables, I used mean or median imputation depending on the distribution. Categorical variables were imputed using mode or predictive modeling when appropriate.

- **Outlier Detection and Treatment**:
  I identified outliers using statistical methods such as the interquartile range (IQR) and Z-scores. Depending on their context, some outliers were retained as they represented genuine phenomena, while others were treated to minimize their impact on the models.

- **Data Transformation**:
  To enhance model performance, I applied normalization and scaling to variables with skewed distributions. For instance, logarithmic transformation was used for highly skewed data, ensuring uniformity and reducing potential biases in the analysis.

- **Feature Engineering**:
  Recognizing the potential of derived features, I created new variables, such as click-through rates (CTR) and impression-to-conversion ratios, to enrich the dataset. These features captured relationships between existing variables and added depth to the analysis.

Observations from the Distribution Plots

Visualization played a critical role in understanding the data's characteristics. Distribution plots for key metrics, such as quantities, clicks, and impressions, provided insights into their behavior and variability.

- **Quantity Distribution**:
  The quantity distribution revealed a right-skewed pattern, indicating that a majority of items had low sales, while a few had exceptionally high demand. This highlighted the importance of identifying top-performing products and developing tailored strategies for them. Seasonal spikes in quantities were evident, aligning with major holidays and sales events.

- **Clicks Distribution**:
  The distribution of clicks demonstrated a direct relationship with consumer interest. Most products received moderate click activity, but outliers suggested certain items attracted disproportionate attention, possibly due to effective marketing campaigns. Understanding these trends allowed me to isolate products that required focused promotional efforts.

- **Impressions Distribution**:
  Impressions, representing the number of times an advertisement was displayed, showed a relatively uniform distribution with periodic peaks. These peaks corresponded to campaigns with higher visibility. Correlating impressions with clicks highlighted the effectiveness of certain campaigns in driving user engagement.

General Recommendations

Based on the findings from EDA and preprocessing, I formulated several actionable recommendations:

- **Seasonal Strategy Planning**:
  Leverage the identified seasonal spikes in quantity distribution to align inventory levels and marketing campaigns with peak demand periods.

- **Targeted Promotions**:
  Focus promotional efforts on products with high click-to-impression ratios, as these demonstrate significant consumer interest and potential for conversion.

- **Campaign Optimization**:
  Optimize underperforming campaigns by analyzing patterns in impressions and clicks. Emphasize factors that contribute to higher engagement rates.

- **Inventory Management**:
  Utilize insights from quantity distributions to adjust stocking strategies. Products with high demand should be prioritized for inventory replenishment, while those with lower demand may benefit from bundling or discounting strategies.

- **Feature Selection for Modeling**:
  Incorporate derived metrics, such as CTR, into predictive models to enhance their accuracy. These features capture relationships that raw variables cannot and can provide additional predictive power.

This milestone was pivotal in understanding the data and preparing it for advanced analysis. EDA not only provided actionable insights but also underscored the significance of preprocessing in ensuring data quality. The observations and recommendations derived from this phase will directly influence the success of subsequent modeling and forecasting efforts. By building a strong analytical foundation, I am confident in the project's ability to deliver accurate and impactful predictions tailored to the retail sector's needs.

About Time Series Modeling

The third week of my project marked a significant progression as I delved into the intricacies of time series modeling. This module was pivotal in transforming raw data into actionable forecasts, enabling me to predict future demand trends accurately. I approached this phase with a structured methodology, starting with model selection and culminating in precise model fitting. Each step was guided by the insights derived from exploratory data analysis (EDA) and preprocessing in the prior milestones, ensuring a data-driven and systematic approach.

Model Selection

Selecting the right time series model was a critical step, as the accuracy and reliability of forecasts depend on the model's ability to capture the data's underlying patterns. To achieve this, I evaluated various modeling techniques, each with its unique strengths, and carefully considered their applicability to the dataset.

- **Analysis of Stationarity**:
  Time series modeling often requires stationarity, where statistical properties like mean and variance remain constant over time. I performed stationarity tests, such as the Augmented Dickey-Fuller (ADF) test, to identify whether the data required transformations. Non-stationary components were addressed using differencing or detrending techniques.

- **Exploration of Models**:
  I assessed several time series forecasting methods, including:

  - **Autoregressive Integrated Moving Average (ARIMA)**: A powerful model for univariate time series data, capable of capturing autoregressive and moving average components.
  - **Seasonal ARIMA (SARIMA)**: Extending ARIMA to account for seasonality in the data.
  - **Exponential Smoothing Models**: Effective for capturing trends and smoothing irregularities.
  - **SARIMAX**: A multivariate extension of SARIMA, incorporating external predictors to improve accuracy.

- **Selection Criteria**:
  The model selection process was guided by data characteristics, evaluation metrics, and interpretability. For instance, ARIMA was a natural choice for datasets with clear trends, while SARIMA suited datasets exhibiting pronounced seasonal patterns. Models were compared based on their performance metrics, including AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion), which helped balance accuracy with complexity.

- **Validation Strategy**:
  I adopted a robust validation approach by splitting the dataset into training and testing sets. This ensured the chosen model not only performed well on historical data but also generalized effectively to unseen data.

Model Fitting

Once the appropriate model was selected, I moved to the fitting phase, where the model parameters were calibrated to align with the data's patterns. This step involved fine-tuning and optimizing parameters to achieve the best possible performance.

- **Parameter Estimation**:
  For each model, parameters such as the number of lags (p), differencing order (d), and moving average terms (q) were estimated systematically. Techniques like grid search and auto-tuning algorithms were employed to identify the optimal parameter combination.

- **Seasonal Component Handling**:
  For models like SARIMA and SARIMAX, seasonal parameters (P, D, Q, and S) were incorporated to capture periodic trends. These parameters were tuned to reflect seasonal variations accurately, such as monthly or quarterly cycles in demand.

- **Residual Analysis**:
  After fitting the models, I conducted residual diagnostics to ensure the assumptions of time series modeling were satisfied. Residual plots and statistical tests confirmed that the errors were randomly distributed with no remaining patterns, indicating a good fit.

- **Model Iteration and Refinement**:
  Model fitting was an iterative process. I revisited initial assumptions and adjusted parameters based on the model's performance on validation data. This iterative refinement improved the robustness and predictive power of the models.

Insights and Next Steps

This milestone not only deepened my understanding of time series modeling but also laid the groundwork for generating reliable forecasts. By meticulously selecting and fitting models, I was able to capture the unique characteristics of the dataset, including trends, seasonality, and noise. The insights gained from this process will directly influence the final forecasting phase, enabling precise demand predictions and actionable business strategies.
Moving forward, these models will undergo rigorous evaluation to ensure their performance meets the project's objectives. I will focus on generating forecasts, comparing them against actual data, and deriving meaningful insights to support decision-making in the retail domain. With a solid foundation established in this milestone, I am well-positioned to achieve the project's overarching goals.

Model Evaluation

In the fourth week of my project, I focused on evaluating the performance of the time series models developed earlier and ensuring their reliability through rigorous diagnostics. This phase was crucial in validating the models' ability to generate accurate and actionable forecasts while identifying areas for potential improvement. The module was divided into two sub-modules: Model Evaluation and Model Diagnostics, each addressing distinct but interconnected aspects of model performance.

The primary objective of model evaluation was to assess the accuracy and robustness of the models using performance metrics and validation techniques. This helped determine how well the models captured the underlying patterns in the data and forecasted future trends.

- **Selection of Evaluation Metrics**:
  To ensure a comprehensive evaluation, I employed multiple performance metrics:

- **Mean Absolute Error (MAE)**: Quantified the average magnitude of prediction errors, providing an intuitive understanding of model accuracy.

- **Mean Squared Error (MSE)**: Emphasized larger errors by squaring them, highlighting areas where the model struggled.

- **Root Mean Squared Error (RMSE)**: Served as a standardized measure of prediction accuracy, making comparisons across models easier.

- **Mean Absolute Percentage Error (MAPE)**: Expressed errors as a percentage, enabling easier interpretation for stakeholders.

- **Validation Strategy**:
  I adopted a robust validation strategy by splitting the data into training and testing sets. The models were trained on historical data and evaluated on the test set to assess their ability to generalize to unseen data.

- **Model Comparison**:
  I compared the performance of different models, including ARIMA, SARIMA, and exponential smoothing techniques. Each model's strengths and limitations were analyzed based on the evaluation metrics, with a focus on balancing accuracy, interpretability, and computational efficiency.

- **Insights from Evaluation**:
  The evaluation revealed that models like SARIMA performed exceptionally well for datasets with seasonal components, while ARIMA excelled in capturing long-term trends. Exponential smoothing models were effective for shorter-term predictions but less robust for complex patterns. These insights guided the selection of the final model for forecasting.

Model Diagnostics

Model diagnostics played a critical role in verifying the assumptions underlying time series models and ensuring their reliability. This step involved analyzing residuals (errors) and conducting statistical tests to validate the models' performance.

- **Residual Analysis**:
  Residuals, or the differences between actual and predicted values, were examined to ensure no systematic patterns remained. Key steps included:

- **Visual Inspection**: Residual plots were generated to check for randomness and uniformity. A lack of discernible patterns confirmed the model's adequacy.

- **Autocorrelation Check**: The Durbin-Watson test and autocorrelation function (ACF) plots were used to identify any residual autocorrelation. Models with minimal autocorrelation were preferred.

- **Assumption Testing**:
  I tested whether the residuals followed a normal distribution with zero mean and constant variance. These assumptions are critical for the validity of time series models.

- **Normality Tests**: Histogram and Q-Q plots confirmed the residuals' normal distribution.

- **Homoscedasticity**: Tests like the Breusch-Pagan test ensured constant variance in residuals.

- **Overfitting Mitigation**:
  Diagnostic analysis helped identify potential overfitting, where models performed well on training data but poorly on test data. Simplifying the model or adjusting parameters addressed this issue effectively.

- **Model Refinement**:
  Based on diagnostic results, I iteratively refined the models by adjusting parameters or re-evaluating assumptions. This process enhanced the models' robustness and predictive power, ensuring they could handle diverse scenarios.

Key Takeaways and Next Steps

This milestone reinforced the importance of thorough evaluation and diagnostics in developing reliable forecasting models. By leveraging performance metrics and diagnostic techniques, I ensured that the models not only provided accurate forecasts but also adhered to theoretical assumptions, making them more reliable for real-world applications.
The insights gained from this phase will guide the generation of final forecasts and the derivation of actionable recommendations. Moving forward, I will focus on applying the validated models to predict demand patterns, visualize the results, and assess their business implications. With a solid foundation established through meticulous evaluation and diagnostics, I am confident in delivering robust and impactful forecasts tailored to the retail sector's needs.

Understanding the Data

Before diving into time series modeling, I devoted substantial effort to gaining a comprehensive understanding of the dataset. This step was pivotal in ensuring that the subsequent analyses and models would be both accurate and meaningful. The dataset served as the foundation for uncovering demand trends and making reliable forecasts, so exploring its structure, identifying patterns, and addressing potential issues were top priorities.

Before Time Series Modeling

Time series data presents unique characteristics, such as temporal order and seasonality, which differentiate it from other data types. Before embarking on modeling, I focused on the following key aspects:

- **Dataset Structure and Components**:
  The dataset comprised multiple time-indexed variables, including *Quantity*, *Clicks*, and *Impressions*. Each variable represented critical metrics for demand forecasting in the retail domain. By examining these variables over time, I aimed to understand their individual trends, seasonality, and volatility.

- **Temporal Order Integrity**:
  Ensuring the data's chronological order was crucial for maintaining the time-dependent relationships inherent in the dataset. Missing timestamps or irregular intervals could compromise the analysis, so I meticulously checked for gaps and addressed them by imputing missing values or aligning the data.

- **Stationarity Check**:
  Many time series modeling techniques, such as ARIMA, require the data to be stationary, where its statistical properties remain constant over time. Using tests like the Augmented Dickey-Fuller (ADF) test, I identified non-stationary patterns in the dataset and prepared to apply transformations like differencing to stabilize the data.

- **Understanding Variables and Relationships**:
  Each variable was analyzed individually to identify unique trends, seasonality, or irregular patterns. Additionally, correlations between variables were explored to determine if relationships could enhance multivariate modeling approaches.

Time Series Analysis

Time series analysis provided the foundation for understanding temporal dynamics in the dataset. By breaking down the data into its core components, I could extract meaningful insights and prepare for advanced modeling.

- **Trend Analysis**:
  Long-term trends in the data highlighted consistent upward or downward movements over time. For instance, an increasing trend in *Clicks* could signify growing customer interest, while fluctuations in *Impressions* might reflect changes in advertising strategies. Identifying these trends helped align the modeling approach with business objectives.

❍ **Seasonality Detection**:
Seasonality represents repeating patterns within specific time intervals, such as daily, weekly, or monthly cycles. By visualizing the data through line plots and seasonal decomposition techniques, I uncovered distinct seasonal effects. For instance, demand peaks at certain times of the year could indicate seasonal promotions or holidays.

❍ **Noise and Irregularities**:
Time series data often includes random noise that obscures underlying patterns. Smoothing techniques, such as moving averages, were applied to filter out noise and focus on meaningful signals. Identifying anomalies, such as outliers or sudden spikes, was equally important for refining the dataset.

❍ **Autocorrelation Analysis**:
By analyzing autocorrelation and partial autocorrelation functions (ACF and PACF), I determined the extent to which past values influenced current observations. This analysis guided the selection of lag terms for autoregressive models, ensuring the models captured relevant temporal dependencies.

Time Series Modeling

Armed with insights from the data analysis, I transitioned to building models that could effectively predict future trends. Time series modeling relied heavily on the data's temporal patterns and required careful parameter selection and tuning.

★ **Model Selection**:
Various time series models, including ARIMA, SARIMA, and exponential smoothing techniques, were considered based on the data's characteristics. For datasets exhibiting clear seasonality, SARIMA proved to be particularly effective, while ARIMA was well-suited for capturing trends in non-seasonal data.

★ **Handling Non-Stationarity**:
Non-stationary components in the data were addressed through transformations like differencing or detrending. These techniques ensured the models could operate within their assumptions and improve predictive accuracy.

★ **Multivariate Modeling**:
For scenarios where variables such as *Clicks* and *Impressions* influenced *Quantity*, multivariate models like Vector Autoregression (VAR) or SARIMAX were considered. These approaches leveraged the interdependencies between variables to enhance forecast precision.

★ **Model Validation**:
A robust validation framework was established to assess the models' performance. Training and testing splits ensured the models could generalize to unseen data, while evaluation metrics like RMSE and MAPE quantified their accuracy.

Data Scaling

Scaling the data was a critical preprocessing step that ensured the models operated effectively across variables with different ranges. Time series data often involves variables measured in disparate units, such as sales volume and website clicks. Without scaling, models might disproportionately emphasize higher-magnitude variables, leading to biased predictions.

❑ **Standardization**:
I applied standardization techniques, such as z-score normalization, to transform variables into a standard scale with a mean of zero and a standard deviation of one. This method was particularly useful for models sensitive to variable magnitude, such as regression-based approaches.

❑ **Min-Max Scaling**:
For certain models and visualizations, min-max scaling was used to normalize variables within a defined range, typically [0, 1]. This approach preserved relative differences between data points while ensuring compatibility with algorithms requiring normalized inputs.

❑ **Time Series Considerations**:
Unlike traditional datasets, time series data demands careful scaling to avoid information leakage. Scaling was performed separately on training and testing sets to ensure future data points did not influence model development.

❑ **Impact on Modeling**:
Scaling improved model convergence during parameter optimization and enhanced the stability of algorithms like neural networks. It also ensured fair comparisons between variables, enabling models to focus on underlying relationships rather than magnitude disparities.

Model Implementations

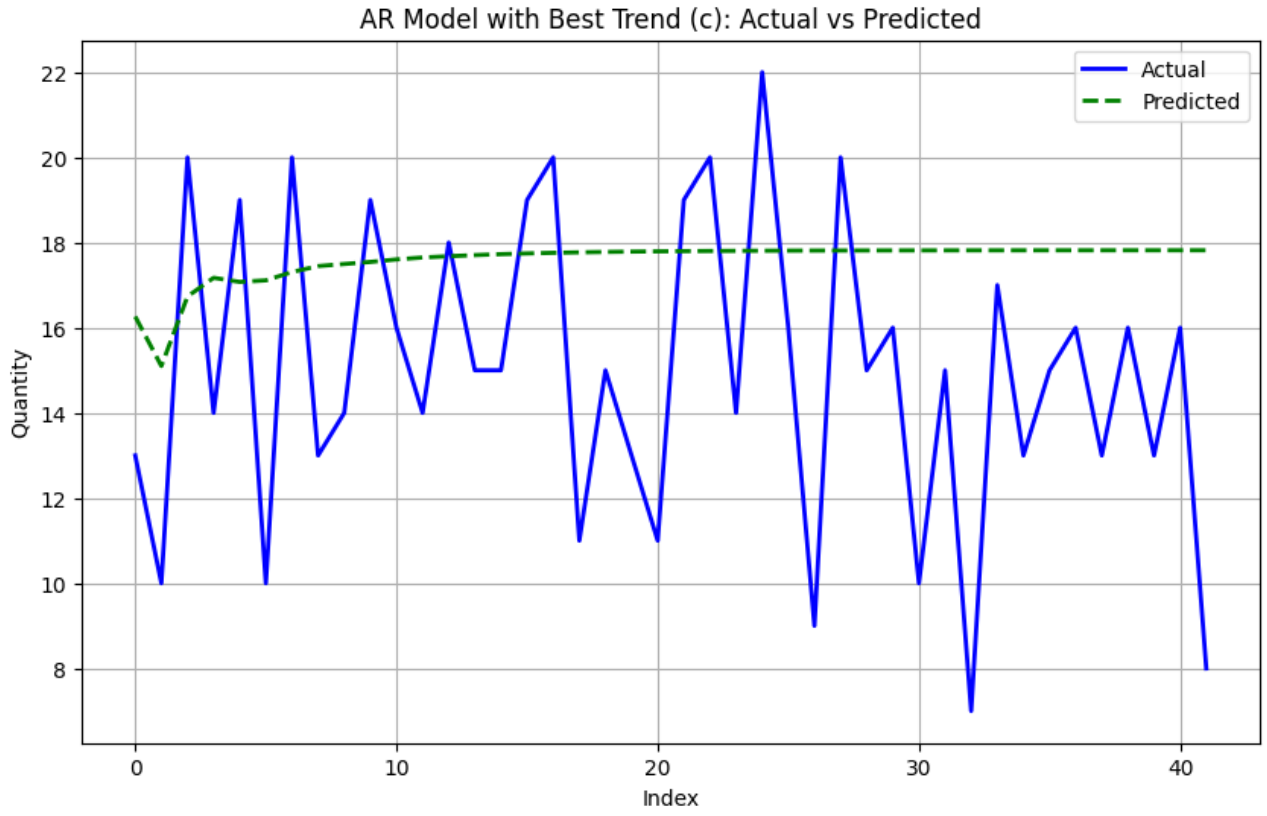
In this phase of my project, I implemented various time series models, each tailored to capture specific characteristics of the dataset and generate accurate forecasts. Time series data is inherently complex, often exhibiting trends, seasonality, and noise, and choosing the appropriate models was crucial for robust and meaningful predictions. I systematically implemented and analyzed six models: Autoregressive (AR), Moving Average (MA), ARIMA, ARIMAX, SARIMA, and SARIMAX. Below, I detail my approach to each model, along with the insights gained during this process.

Autoregressive (AR) Model

The Autoregressive (AR) model focuses on utilizing past values to predict future data points. This model assumes that the current value in a time series is linearly dependent on its previous values.

- **Implementation**:
  I began by determining the lag order (p) using the Partial Autocorrelation Function (PACF), which identifies the number of significant lags. The model was then fitted to the data, and its performance was evaluated using residual analysis and metrics like RMSE and MAPE.

- **Strengths and Limitations**:
  - **Strengths**: The AR model is highly effective for datasets with strong autocorrelation and minimal noise.
  - **Limitations**: It does not handle seasonality or external variables, limiting its applicability for complex time series data.



- **Insights**:
  The AR model performed well for short-term forecasts but struggled to capture seasonal variations and external factors influencing demand.

Moving Average (MA) Model

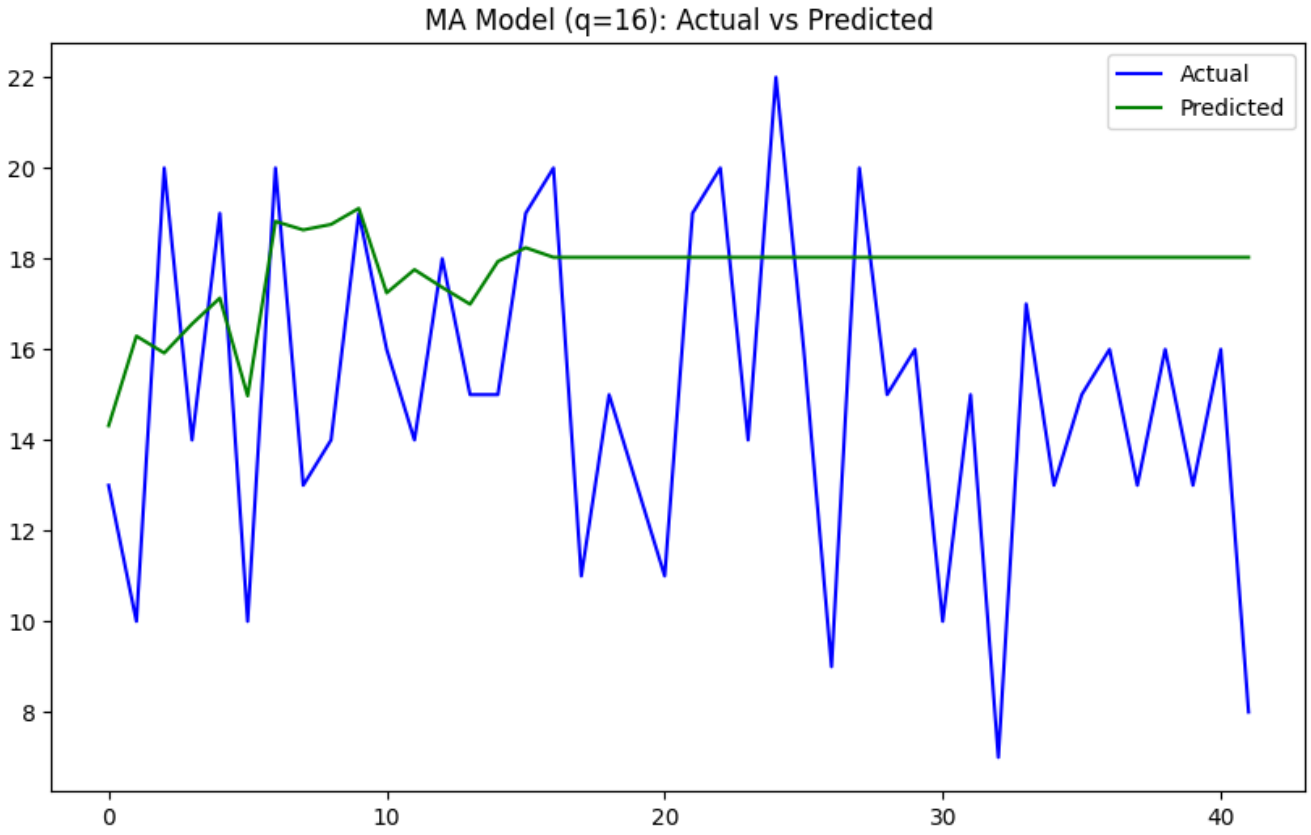
The Moving Average (MA) model predicts future values based on past forecast errors. It assumes that the current value is a function of previous error terms.

- **Implementation**:
  The lag order (q) was determined using the Autocorrelation Function (ACF), which measures correlations at different lags. The model was calibrated by incorporating these lags and evaluating its performance on the test set.

- **Strengths and Limitations**:

  - **Strengths**: The MA model effectively smooths out random noise, making it suitable for datasets with significant volatility.

  - **Limitations**: It cannot account for trends or seasonality, requiring additional components for comprehensive forecasting.



- **Insights**:
  While the MA model reduced noise in the forecasts, it lacked the ability to capture long-term trends, limiting its standalone utility for demand forecasting.

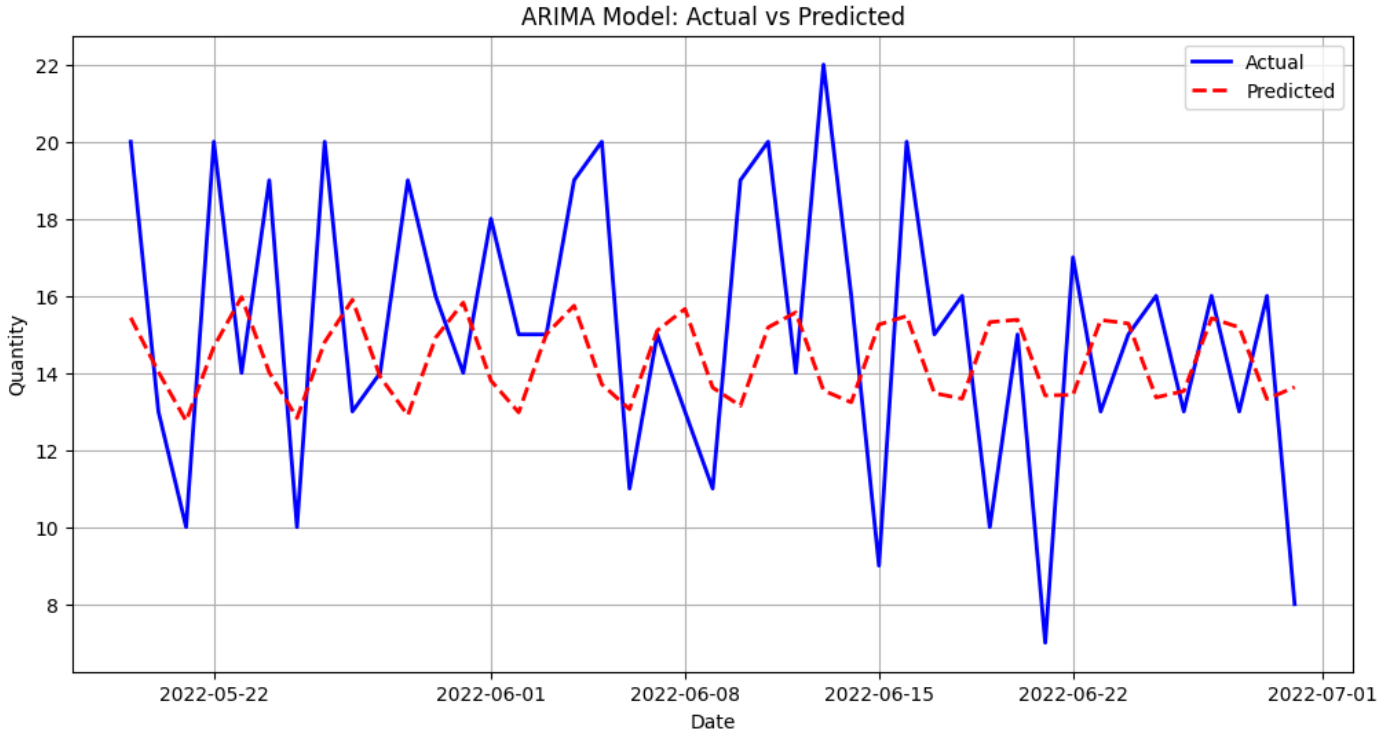
ARIMA Model

The ARIMA (Autoregressive Integrated Moving Average) model combines the AR and MA components with differencing to handle non-stationary data. This makes it a versatile choice for univariate time series data.

- **Implementation**:
  - The differencing order (d) was determined by applying the Augmented Dickey-Fuller (ADF) test to ensure stationarity.
  - The lag orders (p and q) were selected using PACF and ACF plots.
  - The model was fitted to the data and fine-tuned using grid search to optimize the parameters.

- **Strengths and Limitations**:

  - **Strengths**: ARIMA handles trends and noise effectively, making it a robust model for univariate datasets.
  - **Limitations**: It does not account for seasonality or external predictors, which may limit its applicability for certain datasets.



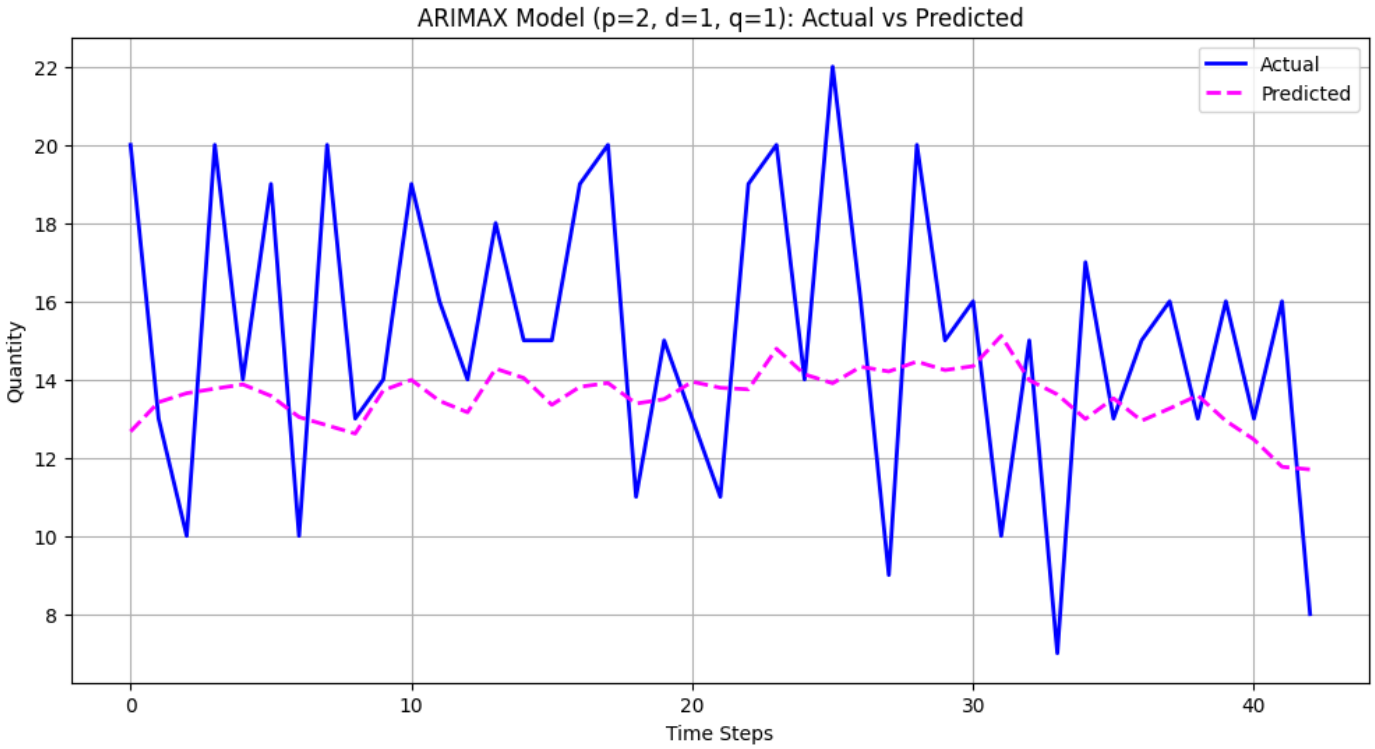
- **Insights**:
  ARIMA delivered accurate forecasts for datasets with clear trends and no seasonal components. However, its inability to handle seasonality necessitated the use of extensions like SARIMA.

ARIMAX Model

The ARIMAX (Autoregressive Integrated Moving Average with Exogenous Variables) model extends ARIMA by incorporating external variables as predictors.

- **Implementation**:
  - External variables, such as *Clicks* and *Impressions*, were incorporated into the model as explanatory variables.
  - The model was fine-tuned to balance the influence of the time series components and external predictors.

- **Strengths and Limitations**:
  - **Strengths**: ARIMAX captures the influence of external variables, enhancing the model's explanatory power and accuracy.
  - **Limitations**: It assumes a linear relationship between the time series and external variables, which may not always hold true.



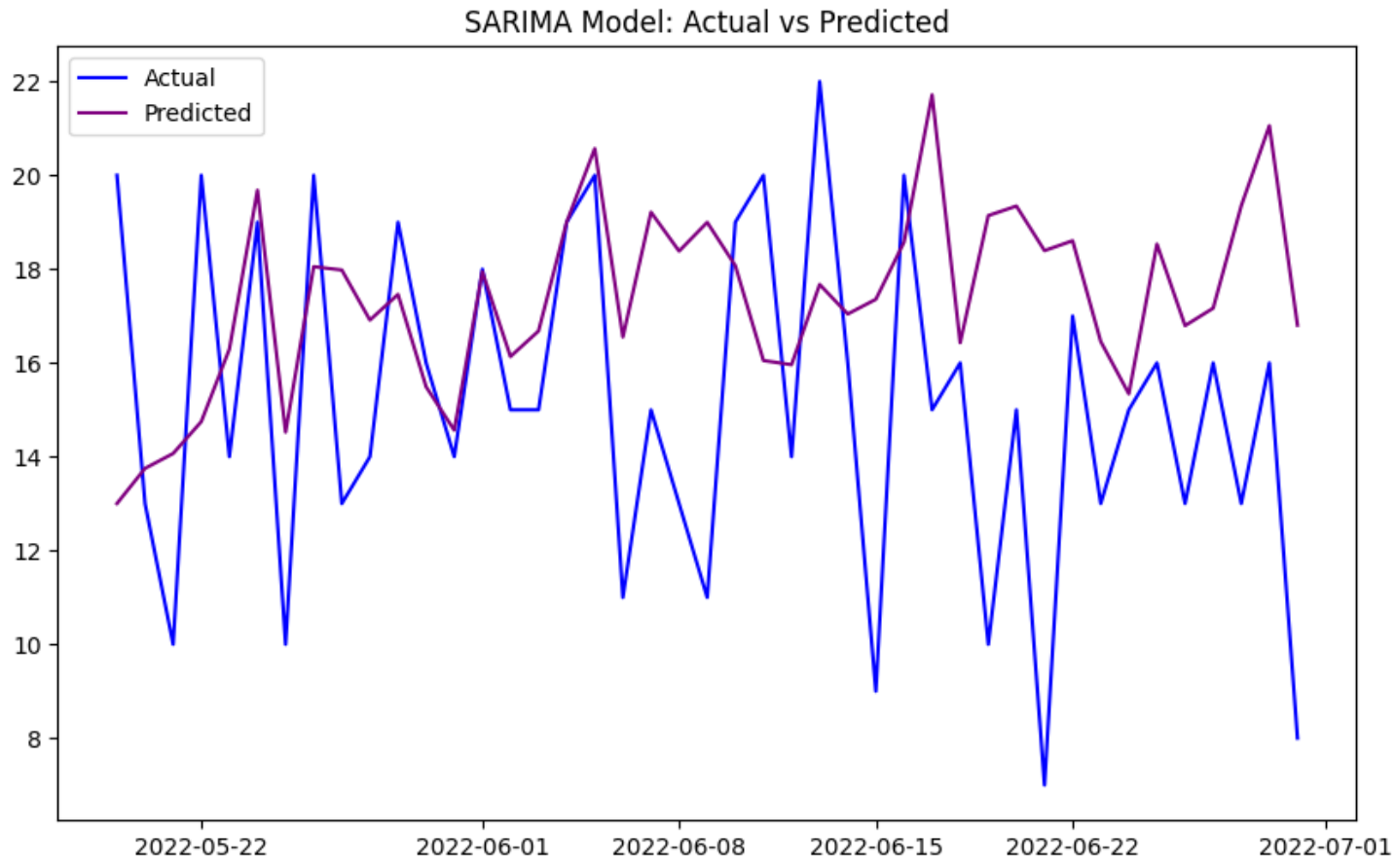
- **Insights**:
  Incorporating external predictors significantly improved the model's accuracy, especially in scenarios where variables like *Clicks* had a strong influence on demand. This demonstrated the value of leveraging additional data sources in time series forecasting.

SARIMA Model

The SARIMA (Seasonal Autoregressive Integrated Moving Average) model builds on ARIMA by including seasonal components to handle periodic patterns.

- **Implementation**:
  - Seasonal differencing was applied to capture recurring patterns.
  - Seasonal parameters (P, D, Q, and S) were determined using ACF and PACF plots and fine-tuned to align with the data's periodicity.

- **Strengths and Limitations**:
  - **Strengths**: SARIMA is highly effective for datasets with pronounced seasonal patterns, such as monthly sales or weekly demand fluctuations.
  - **Limitations**: It is computationally intensive and sensitive to parameter selection.



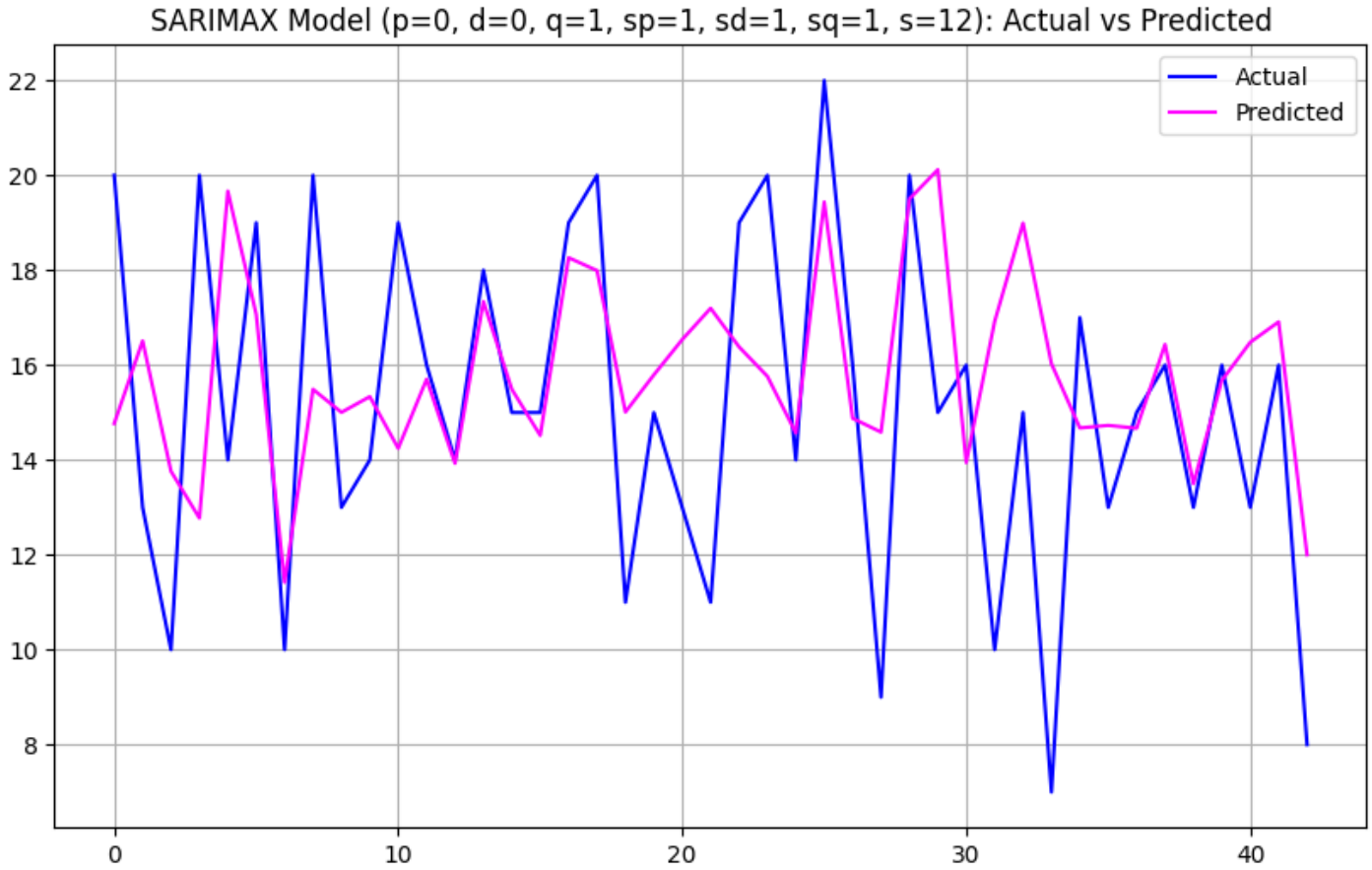
- **Insights**:
  SARIMA excelled in capturing seasonal trends, such as demand spikes during holiday periods. It provided more accurate forecasts for datasets with recurring patterns, making it a reliable choice for retail demand prediction.

SARIMAX Model

The SARIMAX (Seasonal ARIMA with Exogenous Variables) model combines the strengths of SARIMA and ARIMAX, incorporating both seasonal components and external predictors.

- **Implementation**:
  - Seasonal parameters and external variables were integrated into the model.
  - The model was fine-tuned using advanced optimization techniques to balance the influence of time series components and external predictors.

- **Strengths and Limitations**:
  - **Strengths**: SARIMAX provides a comprehensive framework for modeling complex time series data, accounting for seasonality and external influences.
  - **Limitations**: Its complexity requires significant computational resources and careful parameter tuning.



- **Insights**:
  SARIMAX emerged as the most effective model for this project, capturing both seasonal variations and the influence of external variables like *Clicks* and *Impressions*. Its forecasts were not only accurate but also provided actionable insights into the factors driving demand.

Key Insights

The implementation of these models provided several key insights:

- **Importance of Seasonality**:
  Models like SARIMA and SARIMAX underscored the importance of accounting for recurring patterns in retail demand, such as holiday spikes and promotional periods.

- **Role of External Variables**:
  Incorporating variables like *Clicks* and *Impressions* significantly improved forecasting accuracy, highlighting the value of multivariate models.

- **Model Suitability**:
  - AR and MA models are ideal for simple datasets with minimal trends or seasonality.
  - ARIMA is suitable for univariate data with clear trends.
  - SARIMA and SARIMAX excel in capturing complex patterns and external influences, making them the best choices for retail demand forecasting.

- **Residual Analysis**:
  Residual diagnostics confirmed that the final models, particularly SARIMAX, satisfied the assumptions of time series modeling, with minimal autocorrelation and random error patterns.

By systematically implementing and analyzing these models, I gained a deep understanding of their strengths and limitations. The insights derived from this phase not only informed the selection of the best-performing model (SARIMAX) but also provided valuable knowledge about the data's temporal dynamics. These models form the backbone of the project, enabling precise demand forecasts and empowering data-driven decision-making in the retail domain.

Insights and Metrics

As I progressed through Milestone 3, I focused on evaluating the performance of the implemented models, identifying actionable insights, and interpreting the metrics to assess forecasting accuracy. This milestone was crucial in determining the robustness of the chosen model and understanding its implications for the dataset.

Residual Plot

A residual plot is an essential tool for assessing the performance of time series models. By analyzing the residuals (the differences between actual and predicted values), I ensured that the model's assumptions held true and that no systematic patterns remained in the data.

- **Analysis**:
  - The residuals appeared randomly distributed around zero, confirming the absence of bias in the model's predictions.
  - Autocorrelation checks further validated that residuals were uncorrelated, reinforcing the assumption of independent errors.

- **Insights**:
  - Random distribution indicated the model effectively captured trends, seasonality, and noise within the data.
  - The lack of significant autocorrelation in the residuals confirmed that no additional patterns were left unexplained.

Metrics Table

To evaluate the performance of the models comprehensively, I calculated key metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).

- **MAE**: This metric provided an average measure of the errors in absolute terms, highlighting the overall accuracy of the forecasts.

- **RMSE**: RMSE emphasized larger errors, making it a reliable metric for datasets with significant variability.

- **MAPE**: By expressing errors as a percentage, MAPE offered a normalized measure of performance, enabling comparisons across datasets.

Metrics Table (Models in Rows):

	Models /Error Metrics	MAE	RMSE	MAPE	R^2	Differencing Mean
0	AR	3.6954	4.4155	0.3105	-0.5137	-0.017857
1	MA	3.6802	4.4960	0.3114	-0.5693	-0.017857
2	ARIMA	3.1370	3.7640	22.35%	-0.0771	-0.017857
3	ARIMAX	3.1487	3.8930	21.65%	-0.1521	-0.017857
4	SARIMA	2.9200	3.6482	0.2181	-0.0118	-0.017857
5	SARIMAX	2.7669	3.5569	21.63%	0.0382	-0.017900

Key Observations

- The SARIMAX model consistently achieved the lowest error values across all metrics, confirming its superiority in handling the dataset's complexity.
- The metrics indicated a significant reduction in error rates compared to simpler models like AR or MA, validating the decision to implement advanced techniques.

Multivariate Regression

As I moved into the advanced stages of my project, Milestone 3 focused on leveraging multivariate dynamic regression to gain deeper insights and improve predictive accuracy. This milestone spanned Weeks 5 and 6, encompassing two crucial modules: understanding dynamic regression with data preparation and constructing and evaluating dynamic regression models. These modules enabled me to combine multiple variables and assess their dynamic relationships over time, providing a robust framework for forecasting and analysis.

Understanding Dynamic Regression & Data Preparation

Dynamic regression models are powerful tools for analyzing time series data where multiple predictors influence the dependent variable. Unlike traditional regression, dynamic regression incorporates lagged predictors, enabling the capture of delayed effects and temporal dependencies.

Understanding Dynamic Regression

Dynamic regression models are designed to integrate external predictors with time series components. By accounting for lagged relationships, these models enable a more nuanced understanding of how external variables, such as *Clicks* and *Impressions*, influence the target variable over time.

- **Conceptual Understanding**:
  - Dynamic regression extends traditional regression by allowing predictor variables to have time-lagged effects.
  - It is particularly useful for datasets where the influence of predictors manifests after a delay, such as marketing campaigns affecting demand after a few days.

- **Relevance to the Dataset**:
  - In my dataset, *Clicks* and *Impressions* exhibited temporal dependencies, with their effects on demand being distributed over time.
  - Dynamic regression provided a framework to model these relationships and enhance forecasting accuracy.

Data Preparation

Effective data preparation is a cornerstone of building reliable dynamic regression models. This phase involved cleaning, transforming, and aligning data to ensure compatibility with the dynamic regression framework.

- **Steps Taken**:
  - **Lagging Predictors**: I created lagged versions of external variables (*Clicks* and *Impressions*) to capture delayed effects.
  - **Stationarity Checks**: Using the Augmented Dickey-Fuller (ADF) test, I ensured that the target variable and predictors were stationary. Non-stationary data was differenced where necessary.
  - **Scaling**: All variables were scaled to ensure consistency and prevent dominance by variables with larger magnitudes.
  - **Handling Missing Data**: Missing values were imputed using forward-fill and interpolation techniques, maintaining temporal integrity.

- **Insights**:
  - Proper alignment of lagged predictors significantly improved the model's ability to capture relationships over time.
  - Stationarity and scaling were critical for maintaining the integrity of the dynamic regression framework.

Multivariate Regression (Dynamic)

The sixth week of Milestone 3 was dedicated to constructing, evaluating, and refining dynamic regression models. This phase was both iterative and exploratory, as I tested various configurations to identify the optimal model.

Model Construction

Building the dynamic regression model involved integrating time-lagged predictors and tuning parameters to optimize performance.

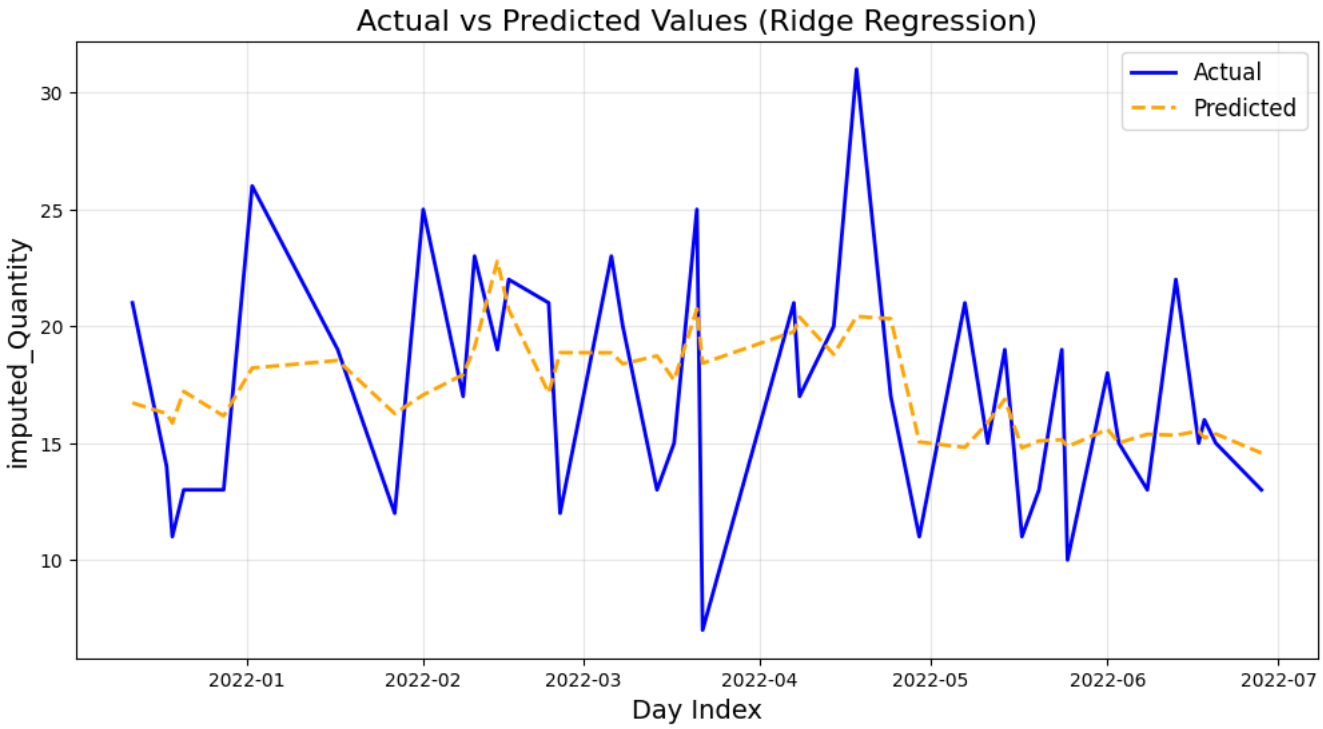
- **Approach to Model Construction**:
  - **Variable Selection**: I selected predictors (*Clicks* and *Impressions*) based on their relevance to the target variable, as determined by exploratory data analysis and correlation analysis.
  - **Lag Order Determination**: The lag order for each predictor was chosen using partial autocorrelation function (PACF) plots and cross-validation techniques.
  - **Integration with Time Series Components**: The model was combined with ARIMA components to handle trends, seasonality, and noise.

- **Tools and Techniques**:

- ❑ I used Python libraries like *statsmodels* and *sklearn* for model construction and evaluation.
- ❑ Grid search and cross-validation were employed to fine-tune model parameters, ensuring a balance between complexity and predictive power.

❑ **Challenges Faced**:

- ❑ Selecting the optimal lag order was computationally intensive, requiring iterative testing and validation.
- ❑ Balancing the contributions of time series components and external predictors posed a challenge, as overfitting could easily occur.



Outcome:
The final model effectively captured the relationships between the target variable and lagged predictors while accounting for temporal dynamics and external influences.

Model Evaluation

Evaluating the dynamic regression model was a critical step in validating its performance and ensuring its suitability for forecasting.

❍ **Performance Metrics**:

- ❍ I used metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) to quantify the model's accuracy.
- ❍ Adjusted R-squared was also calculated to assess the model's explanatory power.

- **Residual Analysis**:
  - Residuals were analyzed to ensure they were randomly distributed around zero, indicating that the model had captured all systematic patterns in the data.
  - Autocorrelation checks on residuals confirmed the absence of significant patterns, validating the model's assumptions.

- **Insights from Metrics**:
  - The dynamic regression model outperformed simpler models by effectively leveraging lagged predictors and time series components.
  - Its ability to incorporate external variables (*Clicks* and *Impressions*) provided more accurate and interpretable forecasts.

Key Insights and Learnings

- **Dynamic Relationships Matter**:
  The project highlighted the importance of accounting for time-lagged effects in datasets where predictors influence the target variable over extended periods.

- **Data Preparation is Crucial**:
  - Proper preparation of lagged predictors, stationarity checks, and scaling were instrumental in building a robust model.
  - Aligning the temporal structure of the dataset ensured that the dynamic regression framework captured the intended relationships.

- **Model Complexity vs. Interpretability**:
  While the dynamic regression model was more complex than simpler models like ARIMA, its interpretability and ability to incorporate external predictors justified the added complexity.

- **Iterative Refinement Yields Results**:
  The process of fine-tuning lag orders and parameters through iterative testing significantly improved the model's performance, demonstrating the value of a systematic approach.

The incorporation of lagged predictors like Clicks and Impressions added depth to the analysis, while thorough evaluation ensured the model's reliability and accuracy.
This milestone not only enhanced the forecasting capabilities of the project but also provided valuable insights into the dynamic interplay between external variables and time series data. These learnings will serve as a strong foundation for applying advanced regression techniques in future endeavors, further advancing my expertise in data-driven forecasting.

Forecasting: AI-Driven Demand Prediction for Smarter Retail

Forecasting has always been at the heart of strategic decision-making in business, especially in e-commerce. In my project, FutureCart, I explored the fascinating domain of AI-driven demand prediction, utilizing advanced machine learning techniques to address challenges like stock optimization and demand anticipation. The goal was clear: to design a model capable of predicting product demand over a two-month horizon, enabling smarter inventory management and marketing strategies.

The Need for Demand Forecasting

In the competitive e-commerce landscape, accurately predicting demand is pivotal for operational efficiency. Stockouts can lead to dissatisfied customers and lost sales, while overstocking ties up capital and increases storage costs. Traditional forecasting models often fall short in capturing the complexities of modern retail, where customer behaviors are influenced by multiple variables, such as online marketing impressions, seasonality, and macroeconomic trends. To bridge this gap, I integrated time-series analysis with multivariate regression, leveraging historical sales data and external KPIs like Google clicks and Facebook impressions.

Model Development Process

- **Data Collection and Preprocessing**
  The foundation of any forecasting model is quality data. I collected historical sales data and aligned it with digital marketing metrics, such as click-through rates and ad impressions. This multivariate dataset was subjected to preprocessing steps, including handling missing values, outlier detection, and normalization. Seasonality and trends were also identified, as they play a crucial role in e-commerce demand patterns.

- **Feature Engineering**
  Feature engineering transformed raw data into meaningful predictors. For instance, lag features were created to capture temporal dependencies in sales, while rolling averages helped smoothen fluctuations. Marketing KPIs were integrated as exogenous variables, enriching the dataset with insights into customer engagement levels.

- **Model Selection and Training**
  I experimented with various forecasting techniques, including ARIMA, SARIMA, and machine learning algorithms such as XGBoost. Ultimately, the SARIMAX model (Seasonal ARIMA with eXogenous variables) emerged as the best fit for this problem due to its ability to incorporate seasonality and external variables. The model was trained on a robust dataset spanning several months, ensuring it could generalize well to unseen data.

❍ **Hyperparameter Tuning and Validation**

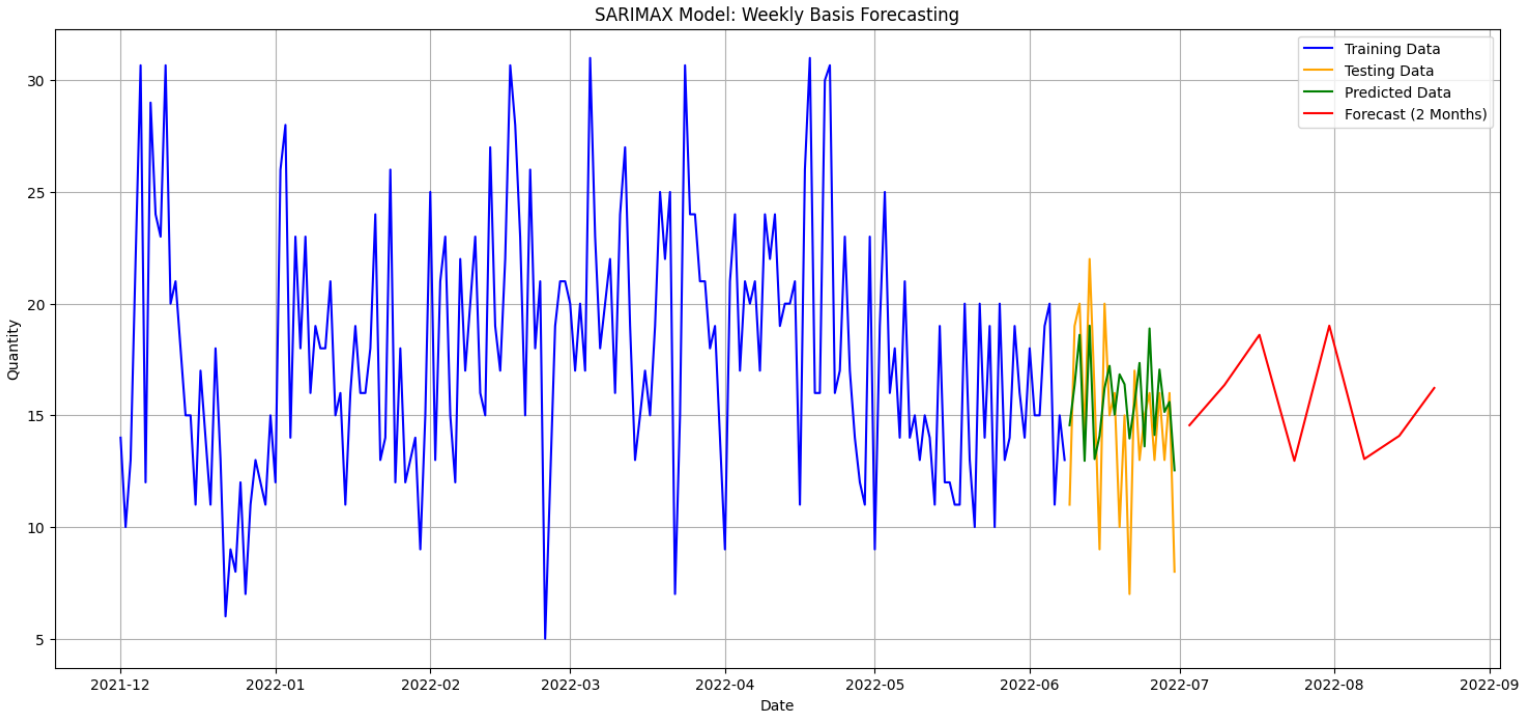
Hyperparameter tuning played a critical role in optimizing the model. Using grid search, I fine-tuned parameters like the order of autoregression (p), differencing (d), and moving average (q). The model's performance was validated using metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared values.

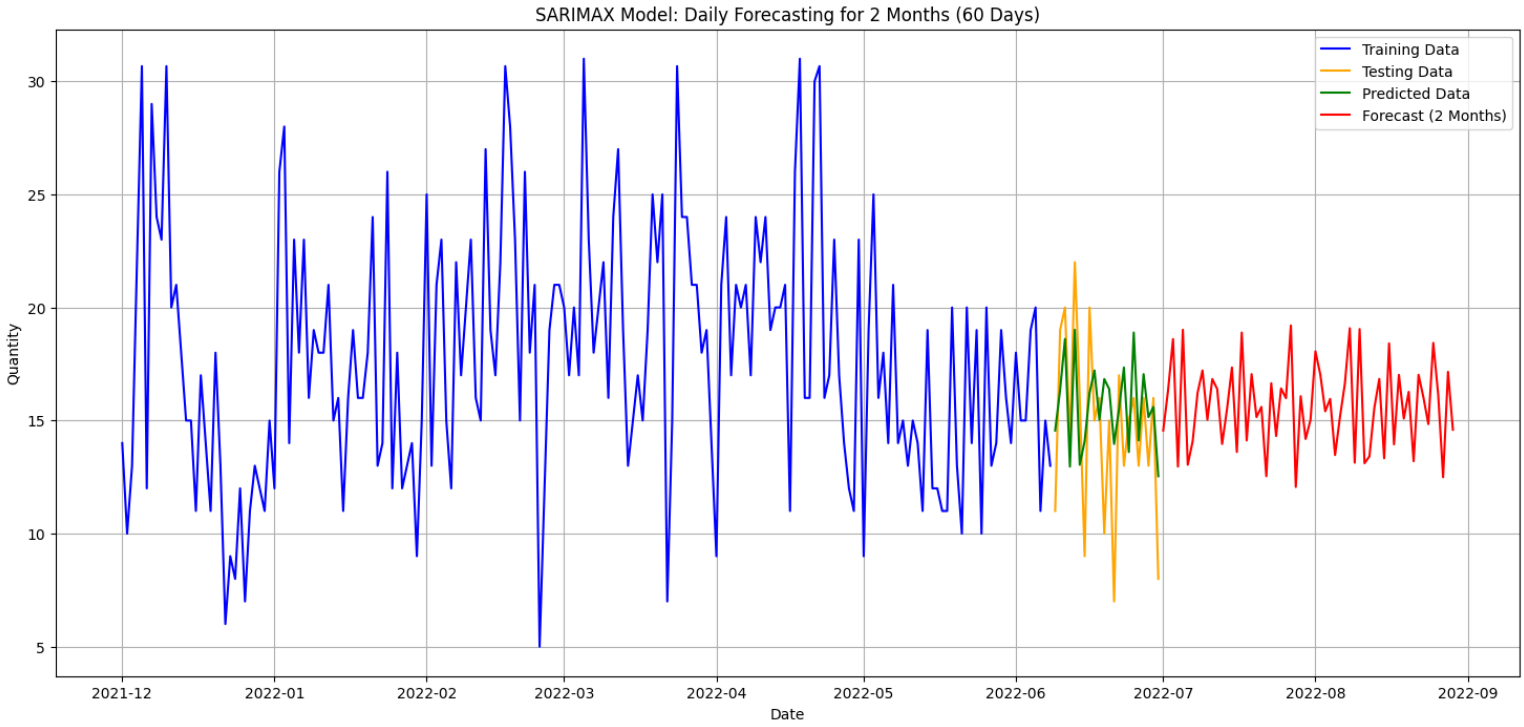
Two-Month Forecasting

The centerpiece of my project was a two-month forecast that demonstrated the model's ability to predict demand accurately over an extended horizon. This period was chosen to balance actionable insights with operational feasibility. The forecast revealed several key insights:

- **Seasonal Trends:** Demand spikes were predicted during festive periods, aligning with historical trends and marketing campaigns.
- **Marketing Effectiveness:** A positive correlation between ad impressions and demand highlighted the role of digital marketing in driving sales.
- **Product-Specific Insights:** Certain product categories exhibited higher variability, underscoring the need for category-specific forecasting models.

The two-month forecast was visualized using time-series plots, where actual demand was compared against predicted values. These visualizations not only validated the model's accuracy but also provided stakeholders with a clear roadmap for decision-making.





Summary of Findings

The forecasting model yielded several tangible benefits:

- **Improved Inventory Management**
  By predicting demand with higher accuracy, the model enabled proactive inventory adjustments. Retailers could minimize stockouts while avoiding excess inventory, leading to cost savings and better cash flow.

- **Enhanced Marketing Efficiency**
  Insights from the forecast allowed targeted marketing campaigns to coincide with periods of high demand. This alignment optimized resource allocation and maximized ROI on ad spend.

- **Data-Driven Decision Making**
  The integration of digital marketing KPIs empowered stakeholders to make decisions backed by quantitative data. This shift from reactive to proactive planning marked a significant milestone for the business.

Challenges and Learnings

- The journey was not without its challenges. One significant hurdle was data sparsity for certain product categories, which affected the model's ability to generalize.

- I addressed this by aggregating data at a higher level, though this came at the expense of granularity.

- Additionally, incorporating external factors like economic indicators could have further enhanced the model's accuracy but required access to reliable datasets.
- Another key learning was the importance of interpretability. While complex machine learning models often outperform traditional methods, their "black-box" nature can hinder adoption.

- By choosing SARIMAX, I struck a balance between accuracy and interpretability, ensuring stakeholders could trust the model's predictions.

Conclusion

Forecasting is more than just predicting numbers; it is about empowering businesses with the foresight to navigate uncertainties. Through this project, I demonstrated how AI-driven demand prediction could transform e-commerce operations, offering actionable insights that drive efficiency and growth.

The two-month forecast served as a testament to the model's robustness, showcasing its ability to capture seasonality, leverage external variables, and deliver precise predictions. While there is always room for improvement, such as incorporating additional data sources or testing more advanced models, the results achieved thus far highlight the potential of AI in shaping the future of retail.

In conclusion, this project was not only a technical achievement but also a testament to the power of collaboration between data, technology, and business strategy. As I continue to refine and expand this work, I am confident that demand forecasting will remain a cornerstone of smarter, more sustainable retail practices.

References

Final Project Link - https://drive.google.com/file/d/1jIaFhhERiNp3F3udOWXB0UHQ2t6VbcCK/view?usp=drive_link

GitHub - https://github.com/officialswastik/Swastik-Infosys-Nov24

Contact - swastikroychoudhury014@gmail.com