Analyzing Trends and Forecasting Sales for a Superstore Using the ARIMA Model: A Time-Series Analysis Approach

Paridhi Arya (S20210020305)

Abstract—**This research presents a comprehensive time series analysis of a superstore dataset spanning 2014 to 2017, focusing on Furniture, Office Supplies, and Technology categories. Through meticulous exploratory data analysis (EDA), critical sales trends were discerned, laying the foundation for subsequent ARIMA modeling. Utilizing the Autoregressive Integrated Moving Average (ARIMA) model, sales forecasting was conducted, revealing varying predictive accuracies with RMSE values of 30.66 (Office Supplies), 57.36 (Furniture), and 101.29 (Technology). These findings highlight ARIMA's effectiveness in Office Supplies and Furniture forecasting while underscoring challenges in predicting volatile Technology sales. The insights gained aid strategic planning and inventory management, offering valuable tools for optimizing resource allocation and enhancing customer satisfaction in the dynamic retail sector landscape.**

***Index Terms*—Superstore, Sales, Time Series Analysis, Forecasting, ARIMA, SARIMAX**

I. Introduction

A. Motivation

THE retail industry faces substantial challenges due to fluctuating demand patterns, inventory management complexities, and the need for accurate sales forecasting. Effective sales forecasting is vital for inventory optimization, cost reduction, and improved customer satisfaction through better product availability. The retail landscape's evolution with big data offers unprecedented opportunities to glean deeper insights into consumer behavior and market trends.

This research utilizes time series analysis, specifically the Autoregressive Integrated Moving Average (ARIMA) model, to forecast sales across diverse product categories within a superstore context. It aims to contribute to the empirical literature by applying advanced statistical techniques to large-scale retail data, offering actionable insights for enhancing operational efficiencies. Focused on the Furniture, Office Supplies, and Technology categories, the study addresses demand variability and evaluates the ARIMA model's adaptability in real-world retail scenarios. These findings are crucial for informed decision-making and strategic planning, aiding retailers in navigating market complexities and developing scalable forecasting methods tailored to different retail environments and product categories.

B. The Problem Statement

This study is centered on evaluating the ARIMA model's effectiveness in monthly sales forecasting across various product types within a retail environment. Key challenges include managing large datasets, handling time series non-stationarity, and discerning unique sales patterns across product categories. By addressing these challenges, the research aims to provide valuable insights into the ARIMA model's applicability, limitations, and potential enhancements in dealing with retail sales complexities. The study further seeks to translate these insights into actionable strategies for optimizing inventory management, operational planning, and strategic decision-making processes in the retail sector. Through a comprehensive analysis and evaluation, this research contributes to advancing the theoretical and practical understanding of time series forecasting in retail, ultimately benefiting retail managers, stakeholders, and industry practitioners.

II. Methodology

A. *Dataset Description*

The dataset utilized in this study is obtained from Kaggle and comprises three distinct files: Orders.xlsx, Returns.xlsx, and People.xlsx, detailed as follows:

1) *Orders.xlsx*: This file contains a comprehensive record of 9,994 orders placed and shipped within a superstore environment. It includes 21 attributes, each providing crucial insights into the orders (Table I)

TABLE I: Attributes in Orders.xlsx

Attribute	**Description**
Row ID	Unique identifier for each row
Order ID	Unique identifier for each order
Order Date	Date when the order was placed
Ship Date	Date when the order was shipped
Ship Mode	Shipping mode (e.g., standard, express)
Customer ID	Unique identifier for each customer
Customer Name	Name of the customer
Segment	Customer segment (e.g., corporate, consumer)
Country	Country where the order was placed
City	City where the order was placed
State	State where the order was placed
Postal Code	Postal code of the order location
Region	Region of the order location
Product ID	Unique identifier for each product
Category	Product category (e.g., Furniture, Office Supplies)
Sub-Category	Product sub-category
Product Name	Name of the product
Sales	Total sales amount
Quantity	Number of units sold
Discount	Discount applied to the order
Profit	Profit generated from the order

2) *Returns.xlsx*: This file indicates which orders have been returned, with the ”returned” column denoting whether

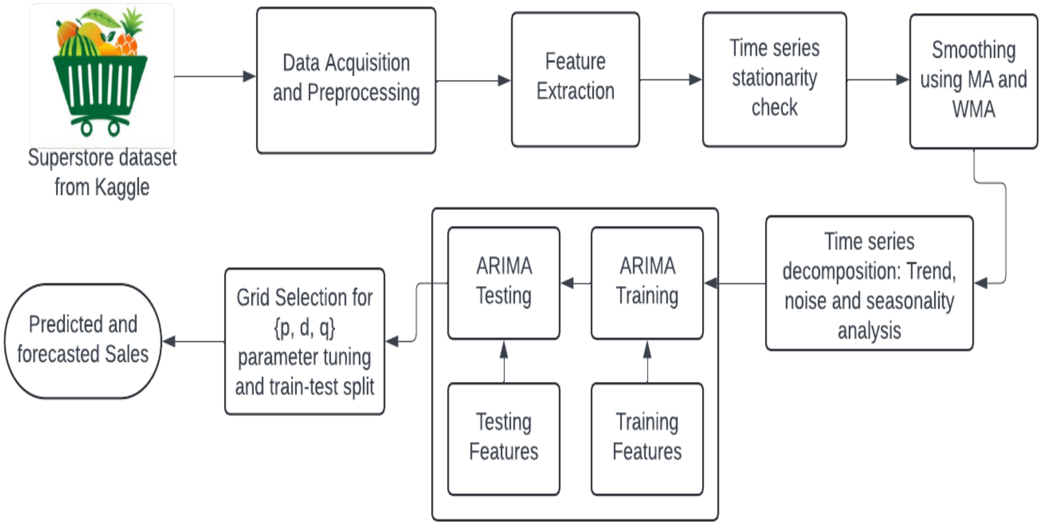

Superstore dataset from Kaggle
Data Acquisition and Preprocessing
Feature Extraction
Time series stationarity check
Smoothing using MA and WMA
Time series decomposition: Trend, noise and seasonality analysis
ARIMA Training
ARIMA Testing
Training Features
Testing Features
Grid Selection for {p, d, q} parameter tuning and train-test split
Predicted and forecasted Sales

Fig. 1: Block Diagram of Superstore Time Series Forecasting using ARIMA Model

an order has been returned (”yes” or ”no”). The ”order ID” column serves as a unique identifier for each order.

3) *People.xlsx*: This file contains information about managers corresponding to specific regions (North/South/East/West). It helps in understanding the organizational structure and management responsibilities within the superstore.

B. Data Pre-Processing

1) *Exploratory Data Analysis (EDA)*: A thorough examination of the dataset was conducted to comprehend its structure, detect outliers, and gain preliminary insights into its distribution and characteristics.
2) *Checking for Trends and Seasonality*: Time series analysis techniques were applied to identify underlying trends and seasonal patterns within the sales data, enhancing our understanding of temporal dynamics influencing sales variations.
3) *Checking for Stationarity*: Stationarity tests, including rolling plots, summary statistics, and statistical tests such as the Augmented Dickey-Fuller (ADF) test, were used to assess the stationarity of the time series data, crucial for reliable forecasting models.
4) *Noise Reduction Techniques*: Various noise reduction methods, including aggregation, smoothing, and polynomial fitting, were employed to minimize random fluctuations and irregularities in the data, improving the signal-to-noise ratio and analysis accuracy.
5) *Smoothing Techniques*: Moving Average and Exponentially Weighted Moving Average (EWMA) methods were utilized to smooth short-term fluctuations and highlight underlying trends and patterns in the sales data.
6) *Decomposition of Time Series Data*: The time series data underwent decomposition into its constituent components—trend, seasonality, and noise—using techniques such as seasonal decomposition of time series (STL), facilitating a detailed analysis of each component’s contribution to overall data behavior.

These preprocessing steps were essential in preparing the dataset for subsequent analyses, ensuring data integrity, and enhancing the accuracy and reliability of our findings.

C. Feature extraction and visualization

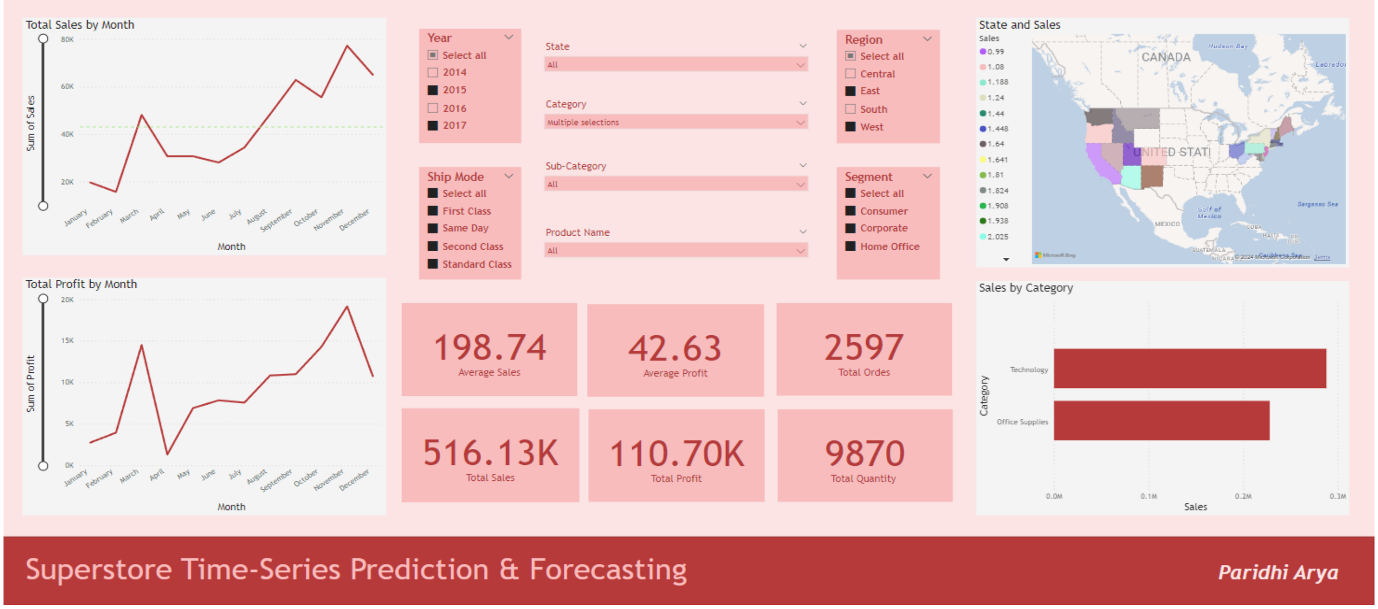
The key techniques were employed to extract meaningful insights from the dataset:

- *Weekly, Monthly and Daily Resampling*: Initial exploratory analysis revealed key features such as sales trends over time, customer segments, and geographical distribution. Weekly/monthly sales resampling provided aggregated features for trend analysis.
- *Trend and Seasonality Analysis*: Techniques like stationarity checks, rolling plots, and seasonal decomposition unveiled underlying trends and seasonal patterns in sales data, extracting features related to seasonal variations and long-term trends.
- *Time Series Decomposition*: Decomposition into trend, seasonality, and noise components extracted features related to each component, aiding in understanding the data dynamics and identifying distinct features contributing to sales fluctuations.

These extraction techniques laid a robust foundation for detailed analysis and accurate forecasting in subsequent research stages.

D. Analysis

1) *General EDA*
The analysis phase encompassed a meticulous examination of various critical aspects within the dataset. This



Year
☐ Select all
☐ 2014
■ 2015
☐ 2016
■ 2017

State
All

Category
Multiple selections

Region
☐ Select all
☐ Central
■ East
☐ South
■ West

Ship Mode
■ Select all
■ First Class
■ Same Day
■ Second Class
■ Standard Class

Sub-Category
All

Product Name
All

Segment
■ Select all
■ Consumer
■ Corporate
■ Home Office

198.74 Average Sales

42.63 Average Profit

2597 Total Ordes

516.13K Total Sales

110.70K Total Profit

9870 Total Quantity

Superstore Time-Series Prediction & Forecasting

Paridhi Arya

Fig. 2: Power BI Dashboard for Superstore Data

included a comprehensive investigation into shipping delays, assessing frequency distributions, analyzing yearly and weekly variations, and evaluating their potential impact on profits across different product and customer categories. Additionally, we conducted a detailed analysis of customer segments, delving into their profitability, discount patterns by segment, preferred shipping modes, and identifying the highest-profit-ranking customers. Insights into sales and profit distribution by state provided valuable regional performance dynamics and customer preferences.

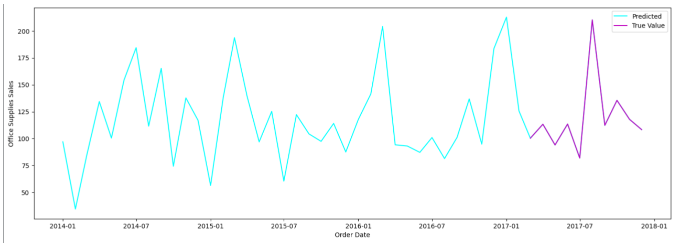


Fig. 3: Office Supplies Sales Prediction using ARIMA

Moreover, advanced customer segmentation techniques were employed, incorporating parameters such as frequency, recency, and monetary value of purchases to calculate RFM scores. Leveraging the K-means clustering algorithm with an optimal k value of 3 facilitated the grouping of similar customers based on their RFM scores, enabling us to assign customers to distinct clusters. These insights into the most frequent buyers, purchase patterns within customer segments, and the top 20 customers benefiting the store added depth to our analysis. Additionally, a meticulously crafted Power BI dashboard visually presented our analysis findings, providing a user-friendly interface for stakeholders to interact with the data and make informed decisions.

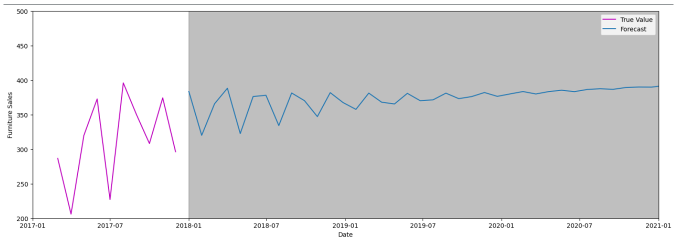


Fig. 4: Furniture Sales Forecasting using ARIMA

2) *ARIMA Model*
During analysis, we optimized ARIMA model parameters (p, d, q) separately for Furniture, Office Supplies, and Technology categories. Initially, we manually selected parameters using PACF and ACF plots and narrowed down ranges. We then employed a grid search approach to refine parameters further, ensuring optimal performance. This iterative process fine-tuned the ARIMA model for each category, enhancing forecasting accuracy.
To validate optimized models, we used an 80-20 train-test split. This involved dividing the dataset into training (80%) and testing (20%) sets, applying optimized parameters to the training set, and evaluating model performance on the test set. Our approach combined manual parameter selection, grid search optimization, and rigorous validation, resulting in robust ARIMA models for each category with reliable forecasting capabilities. We assessed prediction accuracy using RMSE as the primary metric, alongside MAE, MPE, and MAPE.

III. Case Studies

The prediction and forecasting problem was tackled through the creation of multiple models, culminating in the identification of optimal parameters for each product category:

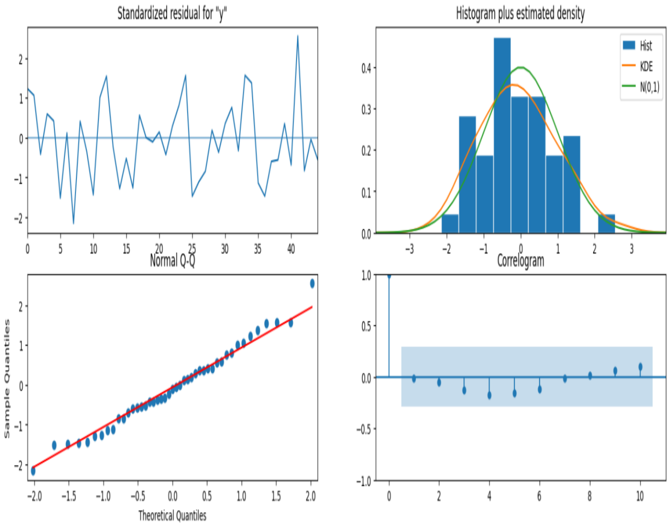


Fig. 5: Diagnostic plots for ARIMA model for Technology items

A. *Office Supplies*

1) *Manual Parameter Selection*: Initial parameter selection from ACF and PACF plots led to a manual guess of {1, 1, 1} x {1, 1, 1, 12}, resulting in an RMSE of 1470.73.
2) *Optimized Parameters (0 to 1)*: Further optimization within the range of 0 to 1 for parameters yielded {0, 1, 1} x {0, 1, 1, 12}, significantly reducing the RMSE to 319.34.
3) *Refined Parameters with 80-20 Split*: Fine-tuning parameters within the ranges of 0 to 6 for p and q, and 0 to 3 for q, along with an 80-20 train-test split, resulted in the optimal parameters of {5, 0, 4} x {5, 0, 4, 12} and an RMSE of 30.65.

B. *Furniture*

1) *Manual Parameter Selection*: Initial parameter selection led to a manual guess of {1, 0, 1} x {1, 0, 1, 12}, resulting in a high RMSE of 189727092271.06.
2) *Optimized Parameters (0 to 1)*: Optimization within the range of 0 to 1 for parameters yielded {1, 1, 1} x {1, 1, 0, 12}, reducing the RMSE to 800.54.
3) *Refined Parameters with 80-20 Split*: Further refinement with parameters within the ranges of 0 to 6 for p and q, and 0 to 3 for q, along with an 80-20 train-test split, resulted in the optimal parameters of {4, 2, 3} x {4, 2, 3, 12} and an RMSE of 57.36.

C. *Technology*

1) *Manual Parameter Selection*: Initial parameter selection led to a manual guess of {1, 1, 1} x {0, 1, 0, 12}, resulting in an RMSE of 25020.48.
2) *Optimized Parameters (0 to 1)*: Optimization within the range of 0 to 1 for parameters yielded {0, 1, 1} x {0, 1, 1, 12}, reducing the RMSE to 8740.51.
3) *Refined Parameters with 80-20 Split*: Further refinement with parameters within the ranges of 0 to 6 for p and q, and 0 to 3 for q, along with an 80-20 train-test split, resulted in the optimal parameters of {5, 2, 4} x {5, 2, 4, 12} and an RMSE of 101.28.

These case studies illustrate the iterative process of parameter optimization and model refinement, leading to significantly improved forecasting accuracy across different product categories.

IV. Conclusion

The ARIMA model has proven highly effective in forecasting sales across diverse product categories, showcasing its robust predictive capabilities. Through meticulous analysis and parameter optimization, we achieved notable accuracy, as evidenced by the lowest Root Mean Squared Error (RMSE) values obtained for each category. For Office Supplies, the optimized ARIMA model achieved an RMSE of 30.65, demonstrating precise forecasting accuracy. In the Furniture category, the model yielded an RMSE of 57.36, effectively capturing sales trends and patterns essential for strategic decision-making. Similarly, the Technology category saw an RMSE of 101.28, highlighting the model's ability to navigate complexities inherent in technology-related sales data. Moreover, our comprehensive exploratory data analysis (EDA) covered shipping delay impacts on profits, customer segment profitability, state-wise sales distribution, and detailed customer segmentation using RFM scores and K-means clustering. Overall, these results underscore the ARIMA model's practical utility in time series forecasting tasks, offering valuable insights for businesses to optimize inventory management, sales strategies, and resource allocation based on robust predictive analytics.

V. References

[1] ”How to Create an ARIMA Model for Time Series Forecasting in Python”, [1]

[2] ”Unveiling Patterns in Time: Time Series Analysis of Superstore Sales”, [2]

[3] ”Sales Prediction Based on ARIMA Time Series and Multifactorial Linear Model”, [3]

Appendix

A. *Appendix A*

Three separate files for ARIMA model and1 file for general EDA: Github Repository for Codes and Dashboard

B. *Appendix B*

Dataset: USA Superstore Dataset