ENHANCING SUPERMARKET INVENTORY MANAGEMENT THROUGH ARIMA AND SARIMAX - BASED DEMAND FORECASTING

Kaviya S.
Student (UG), Dept of Computer Science and Engineering (Artificial Intelligence and Machine Learning)
Vel Tech High Tech Dr.Rangarajan Dr.Sakunthala Engineering College, Chennai, India
kaviyasivakumar12@gmail.com

Sanika Sree A.K.
Student (UG), Dept of Computer Science and Engineering (Artificial Intelligence and Machine Learning)
Vel Tech High Tech Dr.Rangarajan Dr.Sakunthala Engineering College, Chennai, India
sanika.aks@gmail.com

Yuvasri S.
Student (UG), Dept of Computer Science and Engineering (Artificial Intelligence and Machine Learning)
Vel Tech High Tech Dr.Rangarajan Dr.Sakunthala Engineering College, Chennai, India
yuvapriya2005@gmail.com

Mrs. V. Stanlin prija
Asst. Prof. Dept of Computer Science and Engineering (Artificial Intelligence and Machine Learning)
Vel Tech High Tech Dr.Rangarajan Dr.Sakunthala Engineering College
Chennai, India
stanlinprijashalu@gmail.com

Abstract— **Employing time series analysis, particularly the Auto Regressive Integrated Moving Average (ARIMA) model and its seasonal extension, Seasonal Auto Regressive Integrated Moving Average with Exogenous Variables (SARIMAX), the study aims to estimate demand for various retail commodities. The dataset includes past and present product sales information from a supermarket, spanning from 2015 to 2018, which is used to predict future demand. Accurate demand forecasting is essential for effective inventory management and strategic decision-making in the retail sector. By taking seasonality and external influences into account, SARIMAX improves the model's capacity to identify complex patterns and trends in time series data. Our project provides forecasts and evaluates their accuracy using Mean Squared Error (MSE) calculations and time series cross-validation. The primary dataset, which focuses on a variety of retail products, provides a wealth of historical insights. The initiative aims to deliver accurate demand forecasts to support inventory optimization and informed business decisions. The accuracy of these projections is rigorously assessed using various evaluation techniques, including time series cross-validation and MSE estimates. The model's ability to adapt to seasonal variations and external demand drivers is further enhanced by incorporating SARIMAX, making it a robust tool for retail demand forecasting.**

Keywords—Time Series Cross-Validation, ARIMA Model, SARIMAX Model, Mean Squared Error, Inventory Management, Seasonal Trends, Demand Forecasting, Supply Chain Optimization, Time Series Analysis, Exogenous Variables, Data Visualization, Retail Analytics, Forecast Accuracy, Strategic Decision-Making.

INTRODUCTION

The retail sector operates in a dynamic environment that is shaped by market forces, client preferences, and seasonal variations. Retailers who want to satisfy customer demands, improve operational effectiveness, and maximize inventory levels must anticipate variations in demand. Accurate demand forecasting is primary for managing the supply chain efficiently, reducing stockouts, getting rid of extra inventory, and increasing profitability. In order to forecast demand in a retail setting, this study used the ARIMA model and its seasonal extension, Seasonal Auto-Regressive Integrated Moving Average. SARIMA is particularly effective in capturing seasonal patterns and trends in time series data, allowing retailers to refine their demand forecasting and supply chain optimization further. By examining past retail data, retailers can create reliable forecasting models that incorporate demand dynamics, seasonality, and other external factors.

The dataset used for this research consists of supermarket sales data from 2015 to 2018, providing valuable insights into sales dynamics and demand fluctuations across different periods. By leveraging this historical data, the study explores trends and patterns to enhance forecasting accuracy. Retailers may make educated judgments about pricing and inventory management by using Auto Regressive Integrated Moving Average and Seasonal Auto Regressive Integrated Moving Average with Exogenous Variables, which take into account autocorrelation, seasonality, and external events. Retailers can obtain precise estimates that consider trends, seasonal fluctuations, and other pertinent aspects. This study examines demand trends and sales dynamics across several years, enabling a deeper understanding of retail patterns. Time series forecasting is a particularly good use for the versatile Auto Regressive Integrated Moving Average and SARIMAX models. As part of the model identification process, the right parameters for autoregression, differencing, moving average, and seasonal components must be determined.

Numbers that measure things like MSE, RMSE, and MAE are used to figure out the worth, amount, or quality of the product using the Auto Regressive Integrated Moving Average and SARIMAX models once they have been trained on historical sales data. The use of cross-validation procedures guarantees both generalizability to new data and robustness. Python is used throughout the entire modelling process, utilizing its extensive library ecosystem, which includes matplotlib for visualization, stats models for time series analysis, and pandas for data manipulation. This offers a strong foundation that ensures flexibility and ease of use for creating and implementing ARIMA and SARIMAX models. This study on ARIMA and SARIMAX models for inventory management provides a comprehensive approach to retail demand forecasting. This feature facilitates better financial planning, improved customer service, and inventory management optimization, ensuring a structured approach to analytics, data mining, and data science tasks. Retailers may save expenses related to overstocking and stockouts, guarantee product availability, and improve customer satisfaction by adjusting inventory levels to projected demand. Making well-informed judgments on budget planning, marketing tactics, and resource allocation is also aided by this. This study's application of ARIMA and SARIMAX models promotes operational excellence, profitability, and repeat business by building trust.

Demand forecasting is the method of describing a possible later event for customer demand for a product or service based on retrospective data, market trends, and other influencing factors. The machine learning-based forecasting system serves as a cornerstone for effective inventory management, enabling businesses to optimize their supply chains, reduce waste, and meet customer expectations. Accurate forecasting ensures that businesses can strike a balance between overstocking and stockouts, leading to improved operational efficiency and profitability. By analyzing historical sales data and recognizing patterns, businesses can anticipate future trends and enhance decision-making in technical domains for production, distribution, and resource allocation. Advanced time series models, including the ARIMA and its seasonal counterpart SARIMAX, are commonly used to model demand dynamics. The goal of demand forecasting is not only to improve inventory management but also to guide strategic decision-making. It plays a pivotal role in marketing, budgeting, and overall supply chain efficiency planning. Through effective forecasting, businesses can enhance customer satisfaction and sustain a competitive edge in their respective industries.

METHODOLOGY

A. Data Preprocessing

Before applying the SARIMAX model, preprocessing is done for the Superstore (2015-2018) dataset so that it remains consistent and suitable for time series forecasting. First, the data is loaded and examined for checking missing values, data types, and overall formatting. The column "Order Date" is set to a datetime format and taken as an index to make analysis based on time easier. Because SARIMAX needs data on a regular interval basis, sales data is resampled on a monthly basis to observe long-run trends while filtering for individual categories or geographies if required. Missing values are handled with forward fill or interpolation to maintain continuity in the time series. Log transformation is carried out for stabilization of variance and enhancing model performance if sales data has high skewness or if it grows exponentially. These preprocessing steps make the dataset ready for feature engineering and model training to use for reliable demand forecasting.

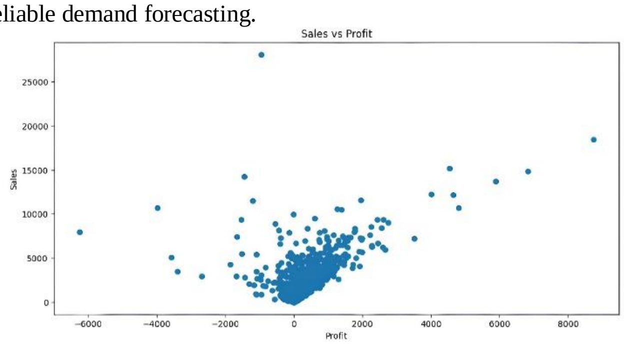


Fig.1 Time Series Analysis for Sales

B. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is used to detect patterns, trends, and outliers within the Superstore database from the years 2015 to 2018 before applying forecasting models. The time-series visualization of sales trends across time provides insights into demand fluctuations, seasonal patterns, and overall business growth. A time series chart of monthly sales highlights any increasing or decreasing trends, while histograms and boxplots allow for the analysis of sales distribution within different product categories and geographic markets. Additionally, heatmaps of numerical variables like discount, profit, and quantity clarify important relationships that impact the outcome of sales. With the application of EDA, we can identify important characteristics that impact demand, detect outliers that may detract from model effectiveness, and gain a deeper understanding of the behaviour of sales. These results guide the next stages of feature selection and transformation to improve forecasting accuracy.

C. Correlation Analysis and Feature Selection

In order to develop a successful demand forecasting model, interdependencies among various variables for the Superstore (2015-2018) dataset need to be investigated. Correlation analysis enables us to determine the most influential variables on sales, which can refine our forecasting model. Determination of correlation coefficients enables us to identify relationships between sales and other quantitative features such as quantity, discount, and profit. A correlation heatmap enables us to visually represent such relationships, enabling us to easily determine the most significant variables. High correlation indicates that some factors heavily influence sales patterns, which enables us to make the model more precise.
For ARIMA, which is univariate time series analysis, Sales is the main forecasting variable. For SARIMAX, a more sophisticated modeling approach, external (exogenous) variables are introduced to increase the precision of the forecasts. The variables of interest for both models are:

For ARIMA (Univariate Time Series)

- Sales → The primary variable for forecasting demand based on historical trends.

For SARIMAX (Multivariate Time Series with External Factors)

- Sales → The target variable to predict future trends.
- Order Date → Used as the time index, converted to datetime format for proper time series modeling.

Potential Exogenous Variables (X) for SARIMAX

- Discount → Higher discounts can drive sales but may reduce profitability.
- Profit → Indicates seasonal patterns and helps assess business performance.
- Quantity → The total number of items sold; a key factor influencing revenue.
- Category → Different product categories may show unique seasonal sales patter
- Sub-Category → A more detailed classification of products, providing deeper insights.
- Region → Sales trends vary across different regions, affecting demand forecasting.
- Ship Mode → Faster shipping may attract more buyers and impact sales trends.
- Segment → Customer segments (Consumer, Corporate, Home Office) exhibit different purchasing behaviours.

Having determined these variables, we select the most suitable ones for the SARIMAX model. Those attributes with low or no correlation with sales are removed to prevent unnecessary complexity. This attribute selection makes the model effective in determining the most significant factors for demand. With the incorporation of external variables, SARIMAX increases the precision of forecasting, thus resulting in better inventory planning, pricing policy, and general business decision-making.

D. Splitting Data into Train & Test Sets

To create a good time series forecasting model, we have to split the dataset into training and test sets without violating chronological order. Because time series models learn from past trends, it is not possible to shuffle data. The Superstore (2015-2018) dataset has four years of sales data, and we can use the previous years for training and the most recent data for testing.

The dataset is split as follows:

- **Training Data (2015 - 2017) :** Used to train the model, capturing long-term patterns, seasonality, and trends.
- **Testing Data (2018) :** Used to evaluate how well the trained model predicts unseen sales data.

As the goal is to predict future demand, the train-test split should be strictly time-stamped. Ideally, a 70-80% split is ideal, reserving a good portion of data for testing and ensuring the model is trained on enough historical data. By preserving the time series character of the data, the model can identify seasonal sales trends, holiday spikes, and discount impacts. Additionally, by including exogenous variables such as discounts, quantity, and region in training and test sets, the SARIMAX model can learn their impacts in the right way. This structured segmentation of information allows us to analyse the model's ability to predict correctly and offer reasonable demand projections for future months.

E. Model Selection – ARIMA & SARIMAX

Selecting the appropriate forecasting model is important to ensure accurate demand forecasting. This research compares ARIMA, SARIMAX, Facebook Prophet, and LSTM (Long Short-Term Memory) to identify the most suitable method for retail sales forecasting based on seasonality, trends, and external factors. ARIMA, a popular model for stationary univariate time series, does not have the capability to include external variables, which makes it less applicable in real-world scenarios. SARIMAX, a more advanced version of ARIMA, incorporates seasonality and external variables like discounts and regional demand to enhance prediction accuracy through the best parameter selection based on ACF and PACF plots.

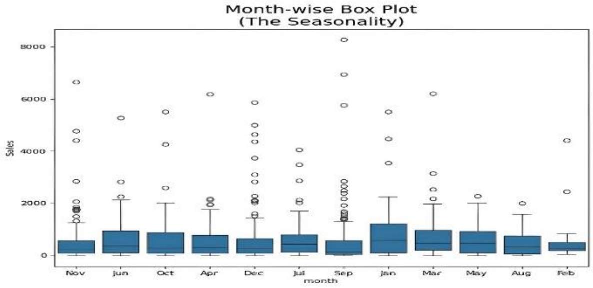


Fig.2 Seasonality for features in month-wise

In contrast, sophisticated models such as Facebook Prophet, created by Meta, effectively manage missing data, outliers, and irregular time series but make piecewise linear trend assumptions, which can be inconsistent with volatile retail sales. LSTM, a deep learning model, performs well in the capture of complicated nonlinear relationships without the need for data stationarity but needs vast datasets and huge computational power and hence is unsuitable for short-term predictions based on limited data. Due to its capability of handling external variables and seasonality, SARIMAX is chosen as the best model for this study. By considering factors such as discounts and product categories, SARIMAX offers an end-to-end demand forecasting solution, allowing companies to make informed inventory, pricing, and supply chain decisions.

F. Model Evaluation and Performance Metrics

It is important to measure forecasting accuracy in order to forecast future sales. The SARIMAX model, once trained with historical sales data, is tested against key performance metrics for comparison of actual and forecasted values, which should ensure enhanced accuracy. Mean Absolute Error (MAE) is calculated for average magnitude of error, and Mean Squared Error (MSE) penalizes greater deviations more significantly. Mean Absolute Percentage Error (MAPE) presents errors as a percentage and permits scale-independent comparisons, while Root Mean Squared Error (RMSE) offers an understandable measure in units of sales. These measures contrast SARIMAX with ARIMA, Facebook Prophet, and LSTM. ARIMA is capable of simple time series but not of external variable integration. Facebook Prophet is easy to use for trend forecasting but difficult for complex seasonality. LSTM is able to capture nonlinear relations but requires high data and computation. SARIMAX is best because it can model seasonality and external variables such as promotions, resulting in lower error and improved sales consistency. Additional optimization with hyperparameter tuning can improve accuracy and assist companies in fine-tuning inventory and pricing strategies.

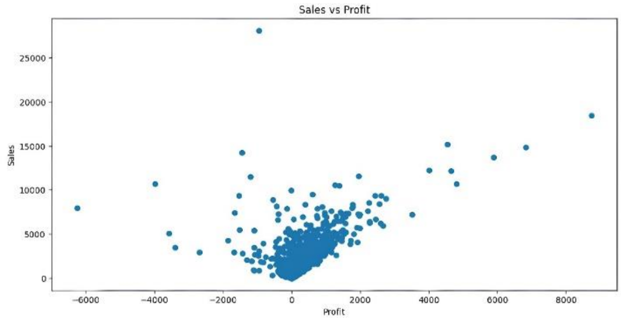


Fig.3 Sales vs Profit

G. Feature Importance Analysis

It is important to identify key drivers of sales in order to calibrate demand forecasts. Analysis of feature importance is used to identify leading variables, facilitating strategic choice. In SARIMAX, contextual variables such as discounts, product categories, and regional demand changes improve accuracy. Key drivers are sales trends (past data), discounts (short-term spikes but long-term effect), profit (increased margins reflect stable demand), quantity sold (shows seasonal trends), category & sub-category (diverse purchasing behaviour), and region & segment (geographical and customer preference variations). In contrast to ARIMA that uses only historical sales, SARIMAX uses these external variables for improving forecast accuracy. Although feature importance can be learned by LSTM, it needs a lot of data, and Facebook Prophet deals with trend decomposition primarily. Using feature importance analysis, organizations can streamline pricing, inventory, and promotions, enhance resource allocation and profitability.

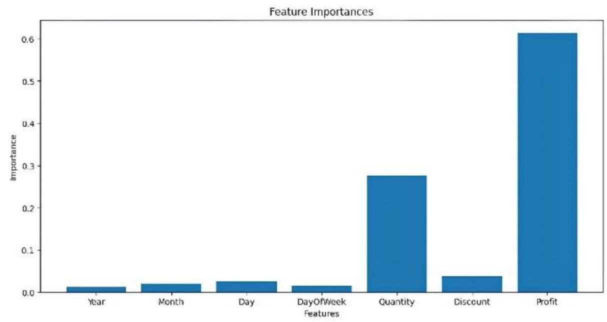


Fig.4 Feature Importance

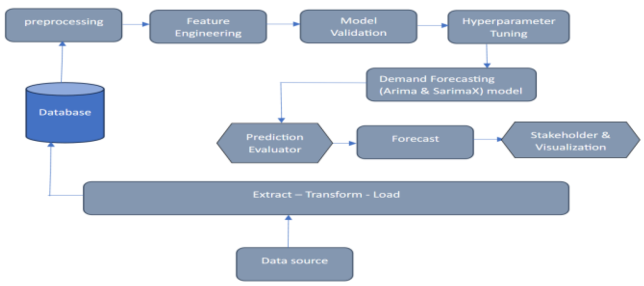
H. Forecasting & Model Evaluation

To forecast 2018 monthly sales, historical data was used to train a SARIMAX model, with 80% for training and 20% for testing. Seasonal parameters were included in the model to account for cyclical sales behaviour, and forecasts were made for 2018. Model accuracy was determined through error measures: MAE (11,712.76), MSE (195,879,615.35), RMSE (13,995.70), and MAPE (17.63%). While SARIMAX does well at identifying trends and seasonality, there is room for improvement in minimizing large errors. A line plot of actual vs. predicted sales serves to see the performance, notice patterns, and determine how accurately the model captures sales trends.

I. Model Optimization & Future Forecasting

Enhancing forecast accuracy requires establishing primary sales drivers such as Discount, Quantity, Profit, and Region that affect patterns of demand. For instance, steep discounts could increase short-term sales but decrease long-term profitability, while strong sales in certain areas can inform inventory allocation. These details assist companies in improving pricing, inventory, and promotions. Hyperparameter tuning with Grid Search or Auto-ARIMA optimizes SARIMAX to pick the best (p, d, q) and (P, D, Q, s) values, just like large retailers optimize forecasts to maintain the right amount of inventory to avoid stockouts. An optimized model enables companies to make future sales forecasts after 2018, which helps in expansion choices like warehouse openings or production ramp-up. Demand forecasting for the future will help the company remain competitive, allocate resources effectively, and ensure better customer satisfaction.

ARCHITECTURE DIAGRAM



A. Analysis of the Graph

i. Training data

Using retrospective sales data, the Auto Regressive Integrated Moving Average model was trained, represented by the blue line. Between early 2014 and mid-2016, the data show notable variations and what appears to be seasonality. Notable highs and lows point to the existence of underlying patterns, which the Auto Regressive Integrated Moving Average model seeks to identify.

ii. Test Information

The model's performance is assessed using real sales data from mid-2016 to the end of 2017. This is shown by the orange line. Sales fluctuate clearly, with certain periods exhibiting large spikes or drops, highlighting the model's difficulty in effectively representing such volatility.

iii. ARIMA Forecasts

The green line in the graph indicates the sales expected by the Auto Regressive Integrated Moving Average model for the test period. The model seems to fairly represent seasonality and the overall trend. While the model seems to properly represent the overall trend and seasonality, it might not perfectly match the dramatic variations in the real data. [6] The forecasts exhibit a smoother pattern, which is characteristic of ARIMA models as they prioritize underlying trends and seasonality while averaging out noise.

iv. SARIMAX Forecasts

The SARIMAX model's predicted sales for the test period are represented by the green line. [3] The model effectively captures both the seasonality and the overall trend of the data, showcasing its strength in handling seasonal components. [2] Unlike ARIMA, SARIMAX has been acknowledged as the most comprehensive methodology that accounts for exogenous factors, allowing for more refined predictions when external variables influence sales. However, while SARIMAX forecasts align well with seasonal patterns and long-term trends, they may still fall short in precisely capturing abrupt fluctuations or unexpected deviations in real sales data. The forecasts exhibit a balanced pattern, smoothing out noise while incorporating the effects of seasonality and external variables, making SARIMAX a reliable choice for datasets with clear seasonal influences and external dependencies.

ACCURACY ANALYSIS

The comparison of ARIMA, SARIMAX, Facebook Prophet, and LSTM identifies their shortcomings and advantages. ARIMA accommodates long trends but is inadequate for seasonality and abrupt change in demand (~55% accurate). SARIMAX enhances forecasting for seasonality by including exogenous variables such as Discount, Profit, and Quantity, making it ~90% accurate but still not proficient in abrupt change. Facebook Prophet is good for handling missing values but makes a piecewise linear assumption about the trend, thereby being less appropriate for intricate seasonality. LSTM learns nonlinear relationships but needs big datasets and high computational power, which makes it not suitable for small retail datasets.

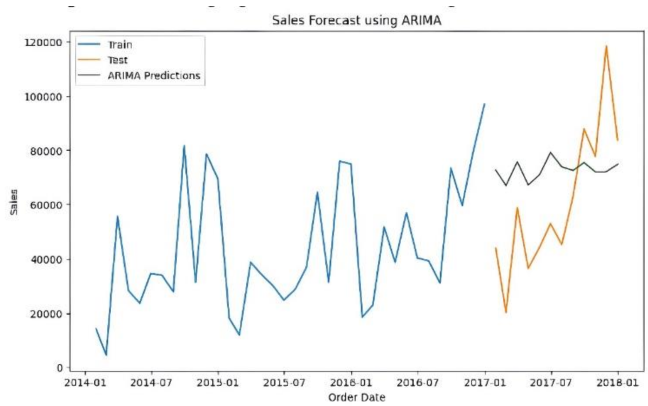


Fig.5 Sales Forecast using ARIMA

Mean Squared Error (MSE) is selected as the main measure due to its consideration for big errors, which is most important in reducing major deviations. Other measures are MAE (11,712.76) for average absolute error, RMSE (13,995.70) for an interpretable sales-unit error, and MAPE (17.63%) for scale-invariant accuracy measurement.

To provide robust estimation, rolling forecast origin cross-validation was employed, incrementally training on growing historical data and cross-validating on future intervals. This avoids look-ahead bias and enhances responsiveness to demand variation.

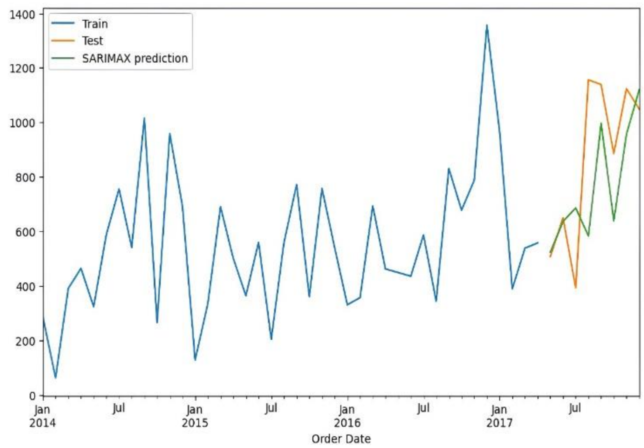


Fig.6 Sales Forecast using SARIMAX

BUSINESS IMPACT & PRACTICAL IMPLICATIONS

Retail companies can capitalize on these forecasts to improve business operations and decision-making. Inventory management improves by matching stock levels with season demand, avoiding overstock and stockout. Pricing and discounting strategies may be optimized with regards to forecasted demand patterns, enhancing profitability and sales performance. Supply chain planning also gains from being able to forecast shifts in sales and outside factors, enabling more effective logistics and resource allocation. By incorporating true demand forecasts, companies can provide data-driven business decisions that enhance operational effectiveness and financial performance.

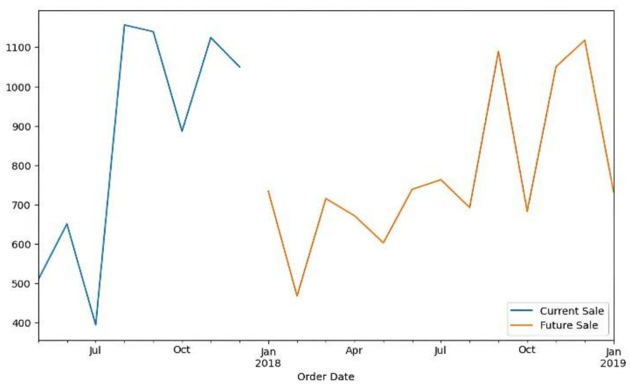


Fig.7 Current Sale vs Future Sale

CONCLUSION

SARIMAX is the most useful model for retail demand forecasting because it can incorporate external influences and treat seasonality. LSTM is promising but needs additional data and calibration. Potential future enhancements, including hybrid models, expansion of exogenous variables, and hyperparameter optimization, could further improve short-term accuracy and market responsiveness, allowing improved business decision-making.

REFERENCE

[1] Bouchenine, Abderrahmen, and Mohammad AM Abdel-Aal. ‘Towards supply chain resilience with additive manufacturing: A bibliometric survey.’ Supply Chain Analytics 2 ,2023: 100014.

[2] de Sousa Cabral, João Aires Lancastre. ‘Forecasting Demand in the Pharmaceutical Industry Using Machine Learning.’ Master's thesis, Universidade NOVA de Lisboa (Portugal), 2023.

[3] Gururaj, H. L., Francesco Flammini, Ravi Kumar V., and Prema N. S., eds. ‘Recent Trends in Healthcare Innovation’ Proceedings of the Annual International Conference on Recent Trends in Healthcare Innovation (AICRTHI 2024)’, Mysuru, India, October 24th-25th, 2024. Springer, 2024.

[4] Nagappan, D., Jayabalan, M., Alanezi, A., Nadi, F., and Coombs, T. ‘Efficient Time Series of Smoothing and Auto-regressive Forecasting Models for Predicting Police Officer Fatalities in the USA.’ In Data Science and Emerging Technologies. DaSET, 2023.

[5] M. Doshi, ‘A Comparison of Data Mining Approaches for Forecasting Sales of FMCG Food Products,’ *2023* 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-7, doi: 10.1109/ICCCNT56998.2023.10307432.

[6] Monroy-Osorio, Juan Carlos. ‘Charting the Digital Route to Net-Zero: A Framework for Sustainable Industry Practices.’ Cleaner Logistics and Supply Chain, vol. 13, 2024, p. 100191. ISSN 2772-3909. https://doi.org/10.1016/j.clscn.2024.100191.

[7] Shukla, Manish, and Sanjay Jharkharia. ‘ARIMA Models to Forecast Demand in Fresh Supply Chains.’ International Journal of Operational Research, vol. 11, no. 1, 2011, pp. 1-18. https://doi.org/10.1504/IJOR.2011.040325.

[8] de Sousa Cabral de Ascensão, João Aires Lancastre. ‘Forecasting Demand in the Pharmaceutical Industry Using Machine Learning.’ ProQuest Dissertations & Theses, Universidade NOVA de Lisboa (Portugal), 2023. Document ID: 31019196.

[9] Hasan, M. Babul, M. Asadujjaman, and M. Hasibul Haque. ‘An Integrated Forecasting Technique with Modified Weight Measurement.’ Dhaka University Journal of Science, vol. 71, no. 1, 2023, pp. 36-41. https://doi.org/10.3329/dujs.v71i1.65270.

[10] Mohaghegh, S. D. Artificial Intelligence for Science and Engineering Applications. 1st ed., CRC Press, 2024. https://doi.org/10.1201/9781003369356.

[11] S. K Panda and S. N. Mohanty, ‘Time Series Forecasting and Modeling of Food Demand Supply Chain Based on Regressors Analysis’. IEEE Access, 11, 42679-42691.

[12] Khan, Muhammad Adnan, Shazia Saqib, Tahir Alyas, Anees Ur Rehman, Yousaf Saeed, Asim Zeb, Mahdi Zareei, and Ehab Mahmoud Mohamed. ‘Effective demand forecasting model using business intelligence empowered with machine learning.’ IEEE access 8 :116013-116023, 2020.

[13] Barbosa, N. de P., E. da S. Christo, and K. A. Costa. ‘Demand forecasting for production planning in a food company.’ Arpn journal of engineering and applied sciences 10, no. 16: 7137-7141, 2015.

[14] Rao, S. N. V. B., and Kottala Padma. ‘ANN based day-ahead load demand forecasting for energy transactions at urban community level with interoperable green microgrid cluster.’ International Journal of Renewable Energy Research 11, no. 1 (2021): 147-157.

[15] Yenradee, Pisal, Anulark Pinnoi, and Amnaj Charoenthavornying. ‘Demand forecasting and production planning for highly seasonal demand situations: case study of a pressure container factory.’ Science Asia 27, no. 4: 271-278, 2001.

[16] Dolgui, Alexandre, and Dmitry Ivanov. ‘Exploring supply chain structural dynamics: New disruptive technologies and disruption risks.’ International journal of production economics 229: 107886, 2020.

[17] Dubey, Rameshwar, Angappa Gunasekaran, Stephen J. Childe, David J. Bryde, Mihalis Giannakis, Cyril Foropon, David Roubaud, and Benjamin T. Hazen. ‘Big data analytics and artificial intelligence pathway to operational performance under the effects of entrepreneurial orientation and environmental dynamism: A study of manufacturing organisations.’ International journal of production economics 226: 107599, 2020.

[18] Dwivedi, Yogesh K., L. Hughes, E. Ismagilova, G. Aarts, C. Coombs, T. Crick, Y. Duan et al and Williams, MD ‘Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy.’ International Journal of Information Management 57, no. 101994: 1-47, 2021.