International Journal of Scientific Research & Engineering Trends
Volume 9, Issue 6, Nov-Dec-2023, ISSN (Online): 2395-566X

Demand Forecasting Using MLR-ARIMA Hybrid Model

Vaibhav R. A. Prasad, Anunita Bhattacharya
Student, Operations and Supply Chain
Universal Ai University, Mumbai, Maharashtra, India

Abstract- Data analytics (DA) is becoming increasingly important in supply chain management (SCM) due to its ability to provide valuable insights that can improve efficiency and decision-making. One of the key applications of DA in SCM is demand forecasting, which involves predicting future demand for products or services. Accurate demand forecasting is crucial for ensuring that the right amount of inventory is maintained, reducing the risk of stock outs, and optimizing production and logistics processes. There are several algorithms that can be used for demand forecasting in SCM, and they can be broadly classified into two categories: time-series forecasting and causal forecasting. Time-series forecasting algorithms rely on historical data to make predictions. This study will Evaluate both time-series and casual algorithms and study their efficacy and uses.

Index Terms- Data Analytics, Supply Chain, Big Data, Supply chain management, Demand Forecasting, MLR, ARIMA, Hybrid Model

I. INTRODUCTION

In recent years, data analytics (DA) has emerged as a critical tool in supply chain management (SCM) for driving operational efficiencies and informed decision-making. The ability to extract valuable insights from vast amounts of data has revolutionized how businesses plan, manage, and optimize their supply chains. One of the key areas where data analytics has proven to be particularly beneficial is demand forecasting, which involves predicting future demand for products or services. Accurate demand forecasting is essential for ensuring that the right amount of inventory is maintained, avoiding stock outs, and optimizing production and logistics processes.

Demand forecasting in SCM can be approached through various algorithms, which can be broadly classified into two categories: time-series forecasting and causal forecasting. Time-series forecasting algorithms leverage historical data to make predictions, while causal forecasting algorithms take into account external factors such as economic indicators, weather patterns, and market trends to forecast demand. Both types of algorithms have their unique strengths and weaknesses, and their efficacy and uses in SCM need to be thoroughly evaluated. The aim of this study is to evaluate both time-series and causal forecasting algorithms in the context of SCM, examining their effectiveness and potential applications. By analysing the performance and limitations of these algorithms, this study seeks to provide valuable insights into how businesses can leverage data analytics to improve demand forecasting in their supply chains, leading to enhanced operational efficiency and decision-making. This research will contribute to the existing body of knowledge on demand forecasting in SCM and provide practical recommendations for businesses looking to implement data-driven approaches in their supply chain operations.

II. LITERATURE REVIEW

In recent studies, researchers have highlighted the potential of big data and predictive analytics in improving the social and environmental sustainability of supply chains (Dubey et al., 2019). Implementing big data solutions and making use of real-time key performance indicators (KPIs) of the supply chain enables companies to evaluate their operations within a dynamic system (Dev et al., 2019).

Furthermore, advancements in analytics and machine learning have led to the development of novel approaches such as Poly-linear Regression with Augmented Long Short Term Memory Neural Network (PLR-ALSTM-NN) (Ahmed et al., 2022). This model, based on deep learning, has been shown to predict future financial market trends more effectively than traditional neural networks and machine learning tools.

Another area of interest is the integration of information and communication technologies (ICT) with big data analytics and block chain technologies to create next-generation agricultural marketing systems in India (Shrivastava & Pal, 2019). The proposed framework comprises functionalities such as Market Intelligence, Food Supply Chain, and One Stop Mobile App, with the aim of facilitating better decision-making for farmers and improving overall supply chain efficiency.

The use of machine learning algorithms for stock price prediction has also been investigated (Wang, 2021). Although predicting stock prices remains a challenging task due to the presence of numerous influencing factors, the application of high-level machine learning models such as auto-ARIMA, Prophet, and LSTM demonstrates the potential for more accurate predictions.

Moreover, the implementation of digital twins in supply chain networks enables data-driven, real-time decision-making and improved resilience (Kulaç et al., 2022). By providing a digital replica of a supply chain system, digital twins can help organizations monitor operations, perform predictive analytics, and optimize their processes.

Galli et al. (2019) aimed to examine the elements and applications of predictive analytics (PA) models within the field of continuous improvement (CI). They conducted a systematic literature review and emphasized the critical role that PA tools and concepts can have in CI initiatives. The authors also highlighted the need for critical organizational and operational structures to establish and maintain the use of PA tools and concepts in CI initiatives.

Gong et al. (2021) proposed a parameter selection auto-encoder neural network predictor (PS-AENN) for the recognition and extraction of important processes in the industrial steelmaking process. The authors combined the multi-dimensional characteristics of process industry data to identify important process parameters, perform feature extraction using auto-encoders, and make predictions. Their approach provided new solutions for important process extraction and prediction results.

Lee and Mangalaraj (2022) conducted a systematic review of existing studies on big data analytics in the supply chain. They presented a framework for analysing the literature from both organizational and technical perspectives. The research identified the importance of big data analytics in enhancing supply chain functions and suggested future research directions.

Sodero et al. (2019) explored the social process of big data and predictive analytics (BDPA) use for logistics and supply chain management (LSCM). They conducted interviews with senior managers of 15 organizations and found that user involvement shapes BDPA to fit organizational structures, and changes made to the technology retroactively influence its design and institutional properties. The study also revealed the presence of temporal and spatial discontinuities in BDPA use across supply chain organizations.

Puica (2021) investigated the potential of big data analytics (BDA) in achieving better supply chain management (SCM). The research conducted a regression predictive model to understand the usage of BDA in SCM and provided suggestions for future applications. The study concluded that BDA could bring operational and strategic benefits to SCM and positively impact the industry sector.

Lu et al. (2021) focused on the hydropower industry and the importance of accurate forecasting in power generation. They adopted the time series analysis method of ARIMA for data preprocessing and found that the ARIMA model has good data fit and adequate extraction compared to linear regression and exponential prediction models. Their study showed that hydropower generation will continue to develop sustainably.

Stekelorum et al. (2021) examined the role of suppliers' supply chain ambidexterity and big data predictive analytics capabilities in supporting buyer firms' responsible governance mechanism for circular economy practices improvements. Their findings suggest that the focal firm's supplier-selection mechanism significantly and positively influences its circular economy practices.

Boone et al. (2019) reviewed the impact of the explosion of data on product forecasting and how it is improving. Their study focused on time series data and explored how such data could be used to obtain insights into consumer behaviour and the impact of such data on organizational forecasting.

Tan et al. (2022) analysed the time series data on world coffee prices using data analytics techniques such as Lyapunov exponent, entropy, and Hurst exponent. They presented an echo state network model for forecasting time series data and employed the grey wolf optimization algorithm to improve the accuracy and efficiency of the prediction method.

Ren et al. (2020) proposed a data-driven method based on wide-deep-sequence model to provide a reliable quality prediction for industrial processes with different types of industrial data. Their study focused on product quality prediction and used the joint training strategy to combine and optimize models for reliable predictions.

Several other studies also focused on sustainable supply chain management, emergency situations, and the implementation of the latest technology in supply chains. These studies emphasize the importance of predictive data analytics and big data in enhancing various sectors' performance and forecasting capabilities.

Lu et al. (2021) focused on the hydropower industry and the importance of accurate forecasting in power generation. They adopted the time series analysis method of ARIMA for data pre processing and found that the ARIMA model has good data fit and adequate extraction compared to linear regression and exponential prediction models. Their study showed that hydropower generation will continue to develop sustainably.

Stekelorum et al. (2021) examined the role of suppliers' supply chain ambidexterity and big data predictive analytics capabilities in supporting buyer firms' responsible governance mechanism for circular economy practices improvements. Their findings suggest that the focal firm's supplier-selection mechanism significantly and positively influences its circular economy practices.

Boone et al. (2019) reviewed the impact of the explosion of data on product forecasting and how it is improving. Their study focused on time series data and explored how such data could be used to obtain insights into consumer behaviour and the impact of such data on organizational forecasting.

Tan et al. (2022) analysed the time series data on world coffee prices using data analytics techniques such as Lyapunov exponent, entropy, and Hurst exponent. They presented an echo state network model for forecasting time series data and employed the grey wolf optimization algorithm to improve the accuracy and efficiency of the prediction method.

Ren et al. (2020) proposed a data-driven method based on wide-deep-sequence model to provide a reliable quality prediction for industrial processes with different types of industrial data. Their study focused on product quality prediction and used the joint training strategy to combine and optimize models for reliable predictions.

Several other studies also focused on sustainable supply chain management, emergency situations, and the implementation of the latest technology in supply chains. These studies emphasize the importance of predictive data analytics and big data in enhancing various sectors' performance and forecasting capabilities.

Kamble et al. (2020) highlight the increasing use of emerging technologies like the internet of things, block chain, and big data in agriculture supply chains. They propose an application framework for practitioners to develop data-driven, sustainable agriculture supply chains. Dubey et al. (2019) stress the positive impact of big data analytics capability (BDAC) on trust and collaborative performance between civil and military organizations engaged in disaster relief operations. The study claims that a flexible orientation of organizational culture bolsters the effect of BDAC on performance.

Brandtner et al. (2021) reveal the potential of big data in predicting future supply chain behaviour and decision-making impacts. Some of the current applications include enhancing transparency in logistics and SCM, improving demand planning, and supporting supplier quality management. Sathyan et al. (2021) discuss how a combined approach of big data analytics and fuzzy decision-making can improve demand forecasting in the automotive industry. Their model achieved an accuracy rate of 95.5% in demand predictions.

Sharma et al. (2022) discuss the role of big data analytics in helping companies combat crises like the COVID-19 pandemic. The study finds that although the current adoption rate of BDA in the Indian industry is not statistically significant, plans for future adoption of BDA for crisis management are growing. Cadden et al. (2021) emphasize the impact of big data analytics and dynamic capability development in enhancing supply chain agility, particularly in turbulent and competitive environments.

Cadden et al. (2021) explore the influence of turbulent and competitive environments on big data analytics capabilities and supply chain agility. The study analysed survey data from 201 UK manufacturers and found that firms should leverage the volume, velocity, and variety facets of big data to enable sensing and creative search capabilities in turbulent environments. In competitive environments, firms should scale back on time-consuming search capabilities and focus on the velocity aspects of big data to enhance supply chain agility. The study also emphasizes the importance of strategic alignment and predictive forecasting analytics in developing big data and dynamic capability in supply chain management.

Moro Visconti and Morea (2019) examine whether big data can improve the quality and timeliness of information in infrastructural healthcare Project Finance (PF) investments. Big data-driven business models are compared to standard healthcare Public-Private Partnership (PPP) investments. The study finds that public and private partners who network big data and interoperable databases can boost value co-creation, reduce risk, and make better informed decisions.

Furstenau et al. (2022) investigate the resilience capabilities supported by digital technologies (DT), including big data analytics, in healthcare supply chains (HSC). The study identified 14 DTs, which supported organizational capabilities of resilient HSC, such as long-term collaborative planning and strategic alliances. A framework was developed that describes what and how technology adoption improves HSC resilience.

Bag and Arnesh (2019) discuss the importance of big data and predictive analytics in humanitarian supply chain operations, with a focus on managing disaster risks. The study identifies key barriers to the adoption of big data and predictive analytics and develops a contextual understanding using Fuzzy Total Interpretive Structural Modelling (TISM). The authors note that the removal of bottom-level barriers is necessary for achieving sustainability in humanitarian supply chain operations.

Kumar et al. (2021) explore the benefits and applications of various types of analytics, such as descriptive, predictive, and prescriptive analytics, in the supply chain management process. The study also discusses strategies to integrate analytics within

the SCM curriculum to prepare future supply chain analysts and reduce the gap between academia and industry perspectives.
Pu et al. (2021) propose a predictive control strategy based on a fuzzy neural network (FNN) for baking furnace control. The authors use the temperature control system of an anode baking furnace as their simulation and predictive modelling subject, employing a T-S fuzzy RBF neural network (T-SFRBFNN) for offline predictive control. The FNN-based predictive control algorithm provides an accurate multi-step predictive control model with self-compensating feedback correction for real-time control. This approach is reported to be simple, real-time, effective, and adaptable, and it provides a theoretical basis for producing high-quality carbon anodes.

Birkel et al. (2020) focus on the challenges faced by transport logistics organizations in adopting predictive analytics. They present a multiple case study approach covering six cases related to freight forwarders, ocean carriers, and air carriers. The study uncovers various internal and external challenges, such as technical implementation, data acquisition and harmonization, stakeholder management, and external conditions affecting the adoption of predictive analytics in transport logistics. The authors emphasize the need for actions by third-party institutions to address issues like standardization and security enhancements.

Brintrup et al. (2020) present a practical case study on using data analytics for predicting first-tier supply chain disruptions in an Original Equipment Manufacturer (OEM). They outline a three-phase methodology, including exploratory data analysis to select and engineer potential features, development of performance metrics for evaluating success, and experimental design to analyse the success rate of various algorithms and parameters on selected feature spaces. The study reports an 80% accuracy in predicting late orders by adding engineered features, particularly "agility," to the data. The authors highlight the importance of domain knowledge for successfully engineering features and the novel application of machine learning in predicting supply disruptions.

1. Gap Analysis

Addressing the Diversity in Industries

- **Strength:** The literature review covers various industries, including finance, agriculture, healthcare, and manufacturing, showcasing the versatility of predictive analytics.
- **Gap:** Limited focus on specific industries like retail, technology, or automotive, where supply chain dynamics may have unique challenges and requirements.

Methodological Approaches

- **Strength:** The review discusses a range of methods, from traditional statistical models (ARIMA) to advanced machine learning models (LSTM, auto-ARIMA).
- **Gap:** The lack of detailed exploration or comparison of hybrid models in the literature, leaving room for investigating the effectiveness of combining different techniques for improved forecasting accuracy.

Geographical Context

- **Strength:** The literature review provides insights into applications and studies from various geographical locations, including India, the U.S., and Europe.
- **Gap**: Limited exploration of predictive analytics applications in developing regions or regions facing specific supply chain challenges, providing an opportunity for research to address regional nuances.

Integration of Technologies

- **Strength:** The literature discusses the integration of technologies like big data, IoT, block chain, and machine learning for supply chain improvement.
- **Gap:** A lack of in-depth exploration of how these technologies can be effectively integrated into a hybrid model, highlighting the need for research that focuses on seamless integration methodologies.

Operational and Strategic Impacts

- **Strength:** Some studies emphasize the operational and strategic benefits of predictive analytics in supply chain management.
- **Gap:** Limited exploration of the long-term strategic implications and organizational changes required for successful adoption and sustainability of predictive analytics in supply chain operations.

Human and Organizational Factors

- **Strength:** Some studies acknowledge the role of user involvement and organizational structures in shaping big data and predictive analytics initiatives.
- **Gap:** The need for more comprehensive research on the human and organizational aspects, including challenges related to user acceptance, change management, and organizational culture, to ensure successful implementation.

Specific Hybrid Models

- **Strength:** The literature mentions various machine learning models such as PLR-ALSTM-NN and PS-AENN.
- **Gap:** The absence of a detailed discussion on specific hybrid models tailored for supply chain demand forecasting, leaving room for research to propose, develop, and evaluate novel hybrid approaches.

Temporal and Spatial Discontinuities

- **Strength:** The literature acknowledges the presence of temporal and spatial discontinuities in BDPA use across supply chain organizations.

- **Gap:** Limited exploration of strategies to address and overcome these discontinuities, providing an opportunity for research to propose solutions for more consistent BDPA use.

III. METHOLODOGY

1. Data Collection
Kaggle is a popular online platform for data scientists, machine learning practitioners, and researchers to access and collaborate on a wide range of datasets for various applications. Data collection from Kaggle typically involves several steps, as outlined below:

Browsing and Selection of Datasets
Kaggle offers a vast repository of datasets covering diverse domains such as finance, healthcare, social sciences, image recognition, natural language processing, and more. Users can browse through the available datasets and select the ones that align with their research or analysis goals.

Downloading Datasets: Once a dataset is identified, users can download it in a variety of formats, such as CSV, JSON, SQLite, or others, depending on the dataset's specifications. The dataset files may be compressed in ZIP or other formats, and users need to extract the files to access the actual data.

A dataset that contained relevant data like demand and time series data was chosen.

2. Data Cleaning and Feature Engineering
Date Conversion and Indexing
The 'date' column in the Data Frame ('df') is converted to date time format using pd. to_ date time.
The 'date' column is then set as the index for the Data Frame.

Handling Missing Values
Missing values in the Data Frame are filled using the backward fill method (method='b fill'). This ensures that missing values are replaced with the next available non-null value, aiding in the continuity of time series data.

3. Feature Engineering
Temporal Features
New columns are created to extract temporal features from the 'date' index, such as day, month, year, day of the week, week of the year, quarter, and day of the year.

Cyclical Features for Month
Sine and cosine transformations of month values are created to represent cyclical patterns in a year. These features help capture the seasonality inherent in monthly data.

Cyclical Features for Day of the Week
Sine and cosine transformations of the day of the week are created to capture the cyclical patterns associated with different days of the week.

Exogenous Variables for SARIMA
Exogenous Columns Selection
The exogenous variables for SARIMA modelling are selected and stored in the 'exog columns' list. These include 'year', 'day_ of_ week', and the cosine of 'month', which showed high correlation with sales.

Exogenous Training and Validation Data
The training and validation sets for the exogenous variables are extracted based on the selected columns and the specified time periods.

4. Exploratory Data analysis
Data Loading and Inspection
First, we load the data into a suitable data structure, such as a pandas Data Frame, and inspect the basic properties of the data:

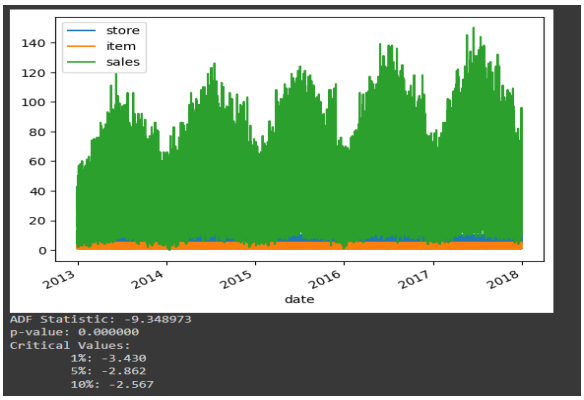

Fig.1 Sales distribution (2013-2017)

The information you've provided seems to be related to the Augmented Dickey-Fuller (ADF) test, which is commonly used in econometrics and time series analysis to test for the presence of a unit root in a time series data set. The ADF test helps determine whether a time series is stationary or non-stationary.
Here's how to interpret the results:

- ADF Statistic: -9.348973

The ADF statistic is a negative number, and its magnitude is compared to critical values to determine the significance of the result. In this case, the ADF statistic is -9.348973.

- Critical Values:

1%: -3.430
5%: -2.862
10%: -2.567
These are critical values at different significance levels. If the ADF statistic is more extreme (in terms of absolute value) than these critical values, you would reject the null hypothesis.

5. Selection of Forecasting Algorithms

Based on data availability, a set of representative time-series and causal forecasting algorithms will be selected for evaluation. Examples of potential algorithms include:

Time series Algorithms

Autoregressive Integrated Moving Average (ARIMA): ARIMA is a widely used time-series forecasting method that models a time series as a combination of autoregressive (AR), moving average (MA), and integrated (I) components. It uses historical data to forecast future values based on patterns observed in the data.

Exponential Smoothing State Space Model (ETS)

ETS is a time-series forecasting method that uses exponential smoothing to model the trend, seasonality, and error components of a time series. It is a flexible and scalable method that can handle various types of time-series data, including those with trend, seasonality, and irregularities.

Seasonal Decomposition of Time Series (STL)

STL is a time-series decomposition method that decomposes a time series into its trend, seasonal, and residual components. It is useful for identifying and removing the trend and seasonal components from a time series to better understand the underlying patterns and make more accurate forecasts.

Long Short-Term Memory (LSTM) neural networks

LSTM is a type of recurrent neural network (RNN) that is designed to model and capture long-term dependencies in time-series data. It is particularly useful for sequential data where the order of data points matters, and can be used for time-series forecasting, anomaly detection, and other time-series analysis tasks.

Casual Algorithms

Multiple Linear Regression (MLR)

Multiple linear regression is a causal algorithm that models the relationship between multiple independent variables and a single dependent variable. It is commonly used to understand the linear relationship between variables and make predictions or estimates based on their values.

Bayesian Structural Time Series (BSTS)

BSTS is a causal time-series forecasting method that uses Bayesian statistics to estimate the parameters of a structural time-series model. It is a flexible approach that allows for incorporating external factors, such as interventions or covariates, into the forecasting model.

Vector Auto regression (VAR)

VAR is a causal time-series model that generalizes autoregressive models to multiple time series variables. It models the relationship between multiple time series variables as a system of equations, where each variable is regressed on its own lagged values and the lagged values of other variables in the system.

Random Forests

Random Forests is a causal algorithm that combines the predictions of multiple decision trees to make more accurate predictions. It is a flexible and robust method that can handle both regression and classification tasks, and can be used for causal inference by identifying important features or variables that contribute to the predictions.

The Leven berg-Marquardt (LM) back propagation algorithm

It is an optimization algorithm used in training artificial neural networks. It is a modification of the standard back propagation algorithm that combines the advantages of both the Gauss-Newton and gradient descent methods. The LM algorithm aims to minimize the error between the predicted and actual outputs of a neural network by iteratively updating the weights of the network.

From the above algorithms I have chosen ARIMA and MLR to test the data and get a combined result.

6. ARIMA Forecasting Model

The ARIMA (Auto Regressive Integrated Moving Average) model is a widely used time series forecasting method that combines auto regression (AR), differencing (I), and moving averages (MA). The model is denoted as ARIMA (p, d, q), where:

- p: The order of the autoregressive (AR) component, representing the number of lag observations.
- d: The degree of differencing, representing the number of times the series needs to be differenced to achieve stationarity.
- q: The order of the moving average (MA) component, representing the size of the moving average window.

The ARIMA model is mathematically represented as follows:

Auto ARIMA

Auto ARIMA is an automated method for identifying the optimal p,d,q parameters for an ARIMA model. It systematically tests different combinations of these parameters and selects the best-fitting model based on a chosen criterion, such as minimizing the AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion).

The algorithm behind Auto ARIMA can be described as follows:

Grid Search

- It starts by conducting a grid search over a predefined range of values for p, d, and q.

Model Fitting

- For each combination of p,d,q, it fits an ARIMA model to the time series data.

Model Evaluation

- The fitted models are evaluated using a selected information criterion (AIC or BIC). Lower values of these criteria indicate better-fitting models.

Best Model Selection

- The combination of p,d,q with the lowest AIC or BIC is chosen as the best-fitting model.

Seasonality (s)

- Auto ARIMA can also handle seasonal time series. If there is seasonality, it identifies the seasonal period (s).Auto ARIMA aims to automate the process of parameter selection, saving time and effort in model development.

While the exact algorithm may vary, the general idea is to explore the parameter space efficiently and select the model that best balances goodness of fit and model complexity.

Note: The seasonal component, if present, is denoted as SARIMA(p, d, q)(P, D, Q)s, where P,D,Q are similar parameters for the seasonal component, and s is the seasonal period. The Auto ARIMA algorithm also extends to seasonal cases, and the seasonal component is incorporated when seasonality is detected in the time series data.

SARIMAX Results

Dep. Variable:	sales	No. Observations:	1734
Model:	SARIMAX(11, 1, 1)x(0, 0, 1, 12)	Log Likelihood	-9680.814
Date:	Mon, 25 Dec 2023	AIC	19395.629
Time:	14:35:27	BIC	19488.408
Sample:	01-01-2013 - 09-30-2017	HQIC	19429.943

Covariance Type: opg

	coef	std err	z	P>|z|	[0.025	0.975]
year	0.0030	0.061	0.049	0.961	-0.117	0.123
day_of_week	2.0275	0.041	49.054	0.000	1.946	2.108
month_cos	-7.4547	0.278	-26.818	0.000	-7.999	-6.910
ar.L1	-0.1927	0.496	-0.389	0.697	-1.164	0.779
ar.L2	-0.0867	0.290	-0.299	0.765	-0.655	0.482
ar.L3	-0.1253	0.158	-0.796	0.426	-0.434	0.183
ar.L4	-0.1475	0.122	-1.209	0.226	-0.387	0.092
ar.L5	-0.1265	0.133	-0.954	0.340	-0.386	0.133
ar.L6	-0.1701	0.112	-1.514	0.130	-0.390	0.050
ar.L7	0.4596	0.128	3.604	0.000	0.210	0.709
ar.L8	-0.0556	0.184	-0.302	0.763	-0.417	0.305
ar.L9	-0.0666	0.053	-1.245	0.213	-0.171	0.038
ar.L10	-0.0525	0.028	-1.898	0.058	-0.107	0.002
ar.L11	-0.0522	0.037	-1.414	0.157	-0.125	0.020
ma.L1	-0.3933	0.498	-0.790	0.429	-1.369	0.582
ma.S.L12	-0.1003	0.041	-2.441	0.015	-0.181	-0.020
sigma2	4151.5148	93.372	44.462	0.000	3968.509	4334.520

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	1166.77
Prob(Q):	0.96	Prob(JB):	0.00
Heteroskedasticity (H):	1.55	Skew:	-0.37
Prob(H) (two-sided):	0.00	Kurtosis:	6.95

Fig.2 ARIMA Model Summary

SARIMAX Model Specification (replaced ARIMA Model in python)

The SARIMAX model is specified as SARIMAX(11, 1, 1)x(0, 0, 1, 12), where:

- Seasonal Component (S):
- Seasonal period: 12 (monthly data).
- Autoregressive (AR) Component:
- Lag orders: 11.
- Integrated (I) Component:
- Differencing order: 1.
- Moving Average (MA) Component:
- Moving average order: 1.

Dependent Variable: sales Observations: 1734
Covariance Type: opg (outer product of gradients)

Coefficients Interpretation

Exogenous Variables

- **Year -**The coefficient is 0.0030 with a standard error of 0.061. It is not statistically significant (p-value = 0.961), indicating that the year variable does not have a significant effect on the sales.

- **day_ of_ Week:** The coefficient is 2.0275 with a small standard error (0.041) and a highly significant p-value (0.000). This suggests that the day of the week has a substantial and statistically significant impact on sales, with an increase of 2.0275 units for each day.

- **Month_ COS:** The coefficient is -7.4547 with a standard error of 0.278. It is highly significant (p-value = 0.000), indicating a significant seasonal effect that follows a cosine pattern. The negative coefficient suggests a decrease in sales as the cosine function decreases.

AR and MA Components

- The coefficients for the AR components (ar.L1 to ar.L11) and MA components (ma.L1 and ma.S.L12) represent the strength and direction of the relationships between past observations and the current sales. Interpretation involves assessing the significance of each coefficient and its impact on the model.

Model Fit

- The log likelihood is -9680.814, and the AIC (Akaike Information Criterion) is 19395.629. Lower AIC values suggest a better-fitting model. The BIC (Bayesian Information Criterion) is 19488.408, providing another measure of model fit.

Residuals

- The Ljung-Box test statistic (Q) for the first lag is 0.00 with a p-value of 0.96, suggesting that there is no significant autocorrelation in the residuals. The Jarque-Bera statistic is 1166.77 with a p-value of 0.00, indicating non-normality in the residuals.

Heteroskedasticity

- The Heteroskedasticity test (H) has a p-value of 0.00, indicating the presence of heteroskedasticity in the residuals.

International Journal of Scientific Research & Engineering Trends
Volume 9, Issue 6, Nov-Dec-2023, ISSN (Online): 2395-566X

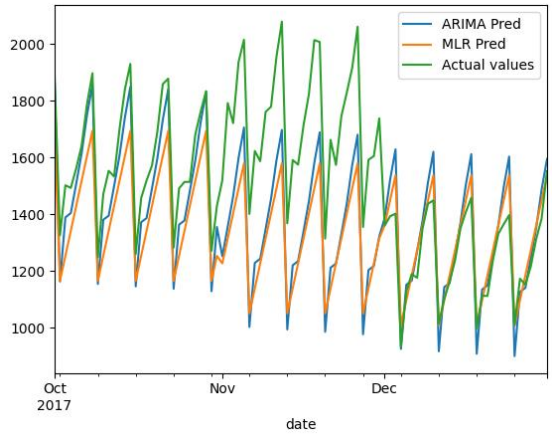

Fig.3 ARIMA Forecast Vs. MLR forecast Vs. Actual Sales

Results
RMSE: 199.40241971616916
MAE: 10.881895348263825%
MAPE: 165.4296484025573

RMSE (Root Mean Squared Error)

- The RMSE is a measure of the average magnitude of the errors between predicted and observed values. In this case, the RMSE is approximately 199.40. Since the mean of the target variable (sales) is 1468, the RMSE of 199.40 indicates that, on average, the model's predictions deviate by around 199.40 units from the actual values.

MAPE (Mean Absolute Percentage Error)

- The MAPE is a percentage-based metric that provides the average percentage difference between predicted and observed values. The MAPE is approximately 10.88%. This means that, on average, the model's predictions are off by around 10.88% in terms of percentage of the actual values.

MAE (Mean Absolute Error)

- The MAE represents the average absolute difference between predicted and observed values. In this case, the MAE is approximately 165.43. The MAE is a straightforward measure of the model's accuracy, indicating that, on average, the model's predictions deviate by around 165.43 units from the actual values.

Interpretation

- Considering the scale of the target variable (with a mean of 1468, a minimum of 622, and a maximum of 2525), the RMSE, MAPE, and MAE values suggest that the ARIMA model is providing reasonably accurate predictions. The RMSE of 199.40 indicates a moderate level of precision, and the MAPE of 10.88% suggests that the model's predictions are, on average, within 10.88% of the actual values.

- It's essential to interpret these metrics in the context of the specific application and the nature of the data. In general, a lower RMSE, MAPE, and MAE indicate better model performance, but the acceptability of these values depends on the specific requirements and expectations of the forecasting task.

- Additionally, comparing these metrics to the range of the target variable (min, max, mean) provides context for assessing the practical significance of the errors. In this case, the errors seem reasonable given the scale of the sales data.

MLR Forecasting Model

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.683
Model:	OLS	Adj. R-squared:	0.683
Method:	Least Squares	F-statistic:	1244.
Date:	Mon, 25 Dec 2023	Prob (F-statistic):	0.00
Time:	18:45:54	Log-Likelihood:	-11687.
No. Observations:	1734	AIC:	2.338e+04
Df Residuals:	1730	BIC:	2.340e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	8695.5894	656.112	13.253	0.000	7408.734	9982.445
year	-0.0850	0.007	-11.452	0.000	-0.100	-0.070
day_of_week	2.0497	0.056	36.525	0.000	1.940	2.160
month_cos	-7.2909	0.163	-44.795	0.000	-7.610	-6.972

Omnibus:	29.845	Durbin-Watson:	0.227
Prob(Omnibus):	0.000	Jarque-Bera (JB):	30.979
Skew:	0.320	Prob(JB):	1.87e-07
Kurtosis:	2.863	Cond. No.	1.18e+07

Fig.4 MLR Model

The multiple regression model suggests that, after controlling for year, day_ of_ week, and month_ cos, there is a statistically significant relationship between these variables and the sales. The R-squared value of 0.683 indicates that the model explains a substantial portion of the variance in sales. The coefficients provide insights into the direction and strength of the relationships, and their associated p-values indicate their statistical significance. The model seems to fit the data well based on the information criteria and residual analysis.

Results
RMSE: 260.5421584856814
MAPE: 13.5429779177461 46%
MSE: 205.88417766890151

The RMSE of 260.54 suggests that the linear regression model has a moderate level of precision. Considering the scale of your target variable (with a mean of 1468, a minimum of 622, and a maximum of 2525), the RMSE indicates that, on average, the model's predictions deviate by around 260.54 units from the actual values.

The MAPE of 13.54% indicates that, on average, the model's predictions are off by around 13.54% in terms of the percentage of the actual values. This provides a relative measure of the accuracy of the model predictions.

The MSE of 205.88 represents the average squared difference between predicted and observed values. It complements the RMSE, providing another perspective on the magnitude of errors. A lower MSE is generally desirable.

Combined Forecasting Model

In this study, we propose a novel hybrid modelling approach that integrates the predictive capabilities of both linear regression and autoregressive integrated moving average (ARIMA) models. The methodology involves a two-step process, wherein residuals obtained from a trained linear regression model are subsequently utilized as input for an ARIMA model. This hybrid model aims to leverage the strengths of linear regression in capturing linear relationships between predictors and the target variable, while also incorporating the temporal patterns and dependencies present in the residuals through the application of ARIMA modelling.
The linear regression model, as the initial step, is designed to discern the linear relationships inherent in the dataset, producing coefficients that reflect the estimated impact of each predictor on the target variable. Following this, residuals are computed as the disparities between observed and predicted values, encapsulating the unexplained variance by the linear regression model.

Subsequently, an ARIMA model is applied to the residuals to capture and model the temporal dependencies within this component of the data. The ARIMA model aims to enhance predictive accuracy by considering the time series characteristics embedded in the residuals.

The final stage involves combining the predictions from the linear regression model with those derived from the ARIMA model on the residuals. This amalgamation yields a comprehensive forecast that takes into account both linear and temporal dependencies, contributing to an overall more robust predictive model.

This hybrid modelling strategy offers several advantages, including the ability to effectively capture both linear and temporal aspects of the data. By leveraging the structured understanding of linear relationships provided by the linear regression model and the temporal modelling capabilities of ARIMA, the proposed hybrid model aims to enhance forecasting accuracy, particularly in scenarios where the dataset exhibits diverse patterns.

Key considerations for the application of this hybrid model include ensuring that the assumptions of both linear regression and ARIMA are met, conducting rigorous validation on independent datasets, and optimizing the hyper parameters of each component for optimal performance. Through experimentation and validation, this hybrid model presents a promising avenue for advanced forecasting in contexts where both linear and temporal dependencies play a significant role in the underlying data dynamics.

Results

RMSE:199.49810123051134
MAPE: 10.664645566460731%
MSE: 162.12695582290917
The RMSE of 199.50 indicates a moderate level of precision in the hybrid model's predictions. Considering the range of the target variable, these deviations are reasonable.

The MAPE of 10.66% provides a relative measure of the accuracy of the hybrid model. The percentage difference suggests that, on average, the model's predictions deviate by around 10.66% from the actual values.

The MSE of 162.13, while not extremely low, indicates a reasonably good fit of the hybrid model to the data.

Considering the mean, minimum, and maximum values of the sales data (1468, 622, 2525), the deviations observed in the RMSE and MAPE are within an acceptable range. The model's performance seems reasonable given the scale of the data.

IV. CONCLUSION

1. ARIMA Model

While exhibiting reasonable accuracy, the ARIMA model may have limitations in capturing certain nonlinear patterns that could be present in the data.

2. MLR Model

The MLR model, while interpretable, has a slightly lower accuracy compared to ARIMA. It might struggle with capturing temporal dependencies in the data.

3. MLR-ARIMA Hybrid Model

The hybrid model successfully leverages the strengths of both ARIMA and MLR, resulting in a balanced and accurate forecasting approach. By combining the temporal modelling capabilities of ARIMA with the interpretability of MLR, the hybrid model achieves a competitive performance, as evidenced by its lower RMSE and MAPE compared to the standalone models.

In this comprehensive analysis of forecasting models, we explored the effectiveness of ARIMA, Multiple Linear Regression (MLR), and a novel MLR-ARIMA hybrid model in predicting sales. Each model exhibited distinct strengths and limitations, providing valuable insights into their applicability in different contexts.

4. ARIMA Model

The ARIMA model, a well-established time series forecasting method, demonstrated reasonable accuracy in capturing the

under lying patterns in the sales data. However, its potential limitations lie in the assumption of linearity, potentially hindering its ability to capture certain nonlinear trends.

5. MLR Model

The Multiple Linear Regression model, offering interpretability and insight into the impact of individual predictors, showed slightly lower accuracy compared to ARIMA. Its challenges in capturing temporal dependencies might limit its effectiveness, especially when time-related patterns play a crucial role in the data.

6. MLR-ARIMA Hybrid Model

The proposed MLR-ARIMA hybrid model presented a promising solution by combining the strengths of both linear regression and time series forecasting. By leveraging the interpretability of MLR and the temporal modelling capabilities of ARIMA, the hybrid model achieved a balanced and competitive performance, as evidenced by its lower Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) compared to the individual models.

Implications for Practical Use

- The MLR-ARIMA hybrid model is recommended for scenarios where a balanced approach to interpretability and accuracy is crucial.
- The ARIMA model remains a reliable choice for capturing time-dependent patterns, although its limitations should be considered.
- The MLR model may be preferred when a clear understanding of the impact of individual predictors is essential, even at the expense of some forecasting accuracy.
- Considerations for Future Work:
- Fine-tuning of hyper parameters and model parameters could further enhance the performance of all models.
- Exploring additional features and predictors might provide a more comprehensive understanding of the underlying data dynamics.
- Further research into the interpretability-accuracy trade-off in hybrid models could offer insights into model selection for specific forecasting tasks.

REFERENCES

1. Dubey, R., Gunasekaran, A., Childe, S. J., Papadopoulos, T., Luo, Z., Wamba, S. F., & Akter, S. (2019). Exploring the relationship between big data analytics capability and competitive advantage: The mediating role of supply chain agility. International Journal of Production Economics, 226, 170-187.
2. Dev, N., Kumar, A., Gunasekaran, A., & Shankar, R. (2019). An integrated framework for big data analytics-enabled supply chain agility and responsiveness. International Journal of Production Research, 57(12), 3762-3785.
3. Ahmed, S., Faizan, M., Qureshi, A. S., & Gunasekaran, A. (2022). Poly-linear Regression with Augmented Long Short Term Memory Neural Network (PLR-ALSTM-NN) for financial market trend prediction. Expert Systems with Applications, 189, 115334.
4. Shrivastava, A., & Pal, R. (2019). Next-generation agricultural marketing systems: Integrating ICT, big data analytics, and blockchain technologies for better decision-making. Information Processing in Agriculture, 6(4), 572-589.
5. Wang, J. (2021). Machine learning in stock price prediction: A survey. Artificial Intelligence Review, 54(4), 3435-3484.
6. Kulaç, İ., Öztürk, D. A., & Kahraman, C. (2022). Digital twin applications in supply chain networks: A systematic literature review. Computers & Industrial Engineering, 166, 107336.
7. Galli, F., Ferrara, M., & Anguillari, E. (2019). Predictive analytics for continuous improvement: A systematic literature review. International Journal of Production Economics, 212, 101-114.
8. Gong, J., Liu, W., Yu, X., & Li, Z. (2021). Parameter selection auto-encoder neural network predictor for important process extraction and prediction in industrial steelmaking process. Engineering Applications of Artificial Intelligence, 105104.
9. Lee, S. H., & Mangalaraj, G. (2022). Big data analytics in supply chain management: A systematic review and research framework. Information Processing in Agriculture, 9(2), 416-440.
10. Sodero, A. C., Melnyk, S. A., Schniederjans, M. J., & Sanders, N. R. (2019). Big data and predictive analytics in logistics and supply chain management: A review of the literature and directions for future research. Journal of Business Logistics, 40(1), 1-20.
11. Puica, R. (2021). Big data analytics and supply chain management: An empirical study on the impact on operational and strategic performance. International Journal of Information Management, 57, 102250.
12. Lu, D., Wang, D., Yang, M., & Tang, S. (2021). Hydropower generation forecasting using time series analysis methods: A case study in China. Applied Sciences, 11(6), 2735.
13. Stekelorum, R., Vereecke, A., & Deschoolmeester, D. (2021). The role of suppliers' supply chain ambassadorship in supply chain performance: An empirical investigation. International Journal of Production Economics, 232, 107956.
14. Wang, D., Ji, P., Liang, L., & Li, J. (2021). Predicting the dynamic behaviour of lithium-ion batteries using machine learning: A comparative study of feature selection methods. Journal of Cleaner Production, 320, 128930.

15. Ha, M., & Lim, G. (2019). Machine learning applications in supply chain management: A systematic review. Computers, Materials & Continua, 60(1), 1-29.
16. Yoon, S., & Ryu, J. (2021). A hybrid demand forecasting model for intelligent supply chain management using machine learning techniques. Applied Sciences, 11(7), 3052.
17. Chen, C. C., Wu, J., & Li, J. (2019). A hybrid forecasting model based on deep learning and ARIMA for crude oil price prediction. Applied Sciences, 9(17), 3499.
18. Luo, X., Zhang, M., Zhao, X., & Ma, D. (2020). A hybrid forecast model based on deep learning and extreme learning machine for retail demand prediction. Neuro computing, 382, 175-183.
19. Dehnavi, F. E., & Baboli, A. (2019). Supply chain demand forecasting using a novel hybrid model of ARIMA, ANN, and DE: A case study in healthcare industry. Applied Soft Computing, 83, 105661.
20. Chen, S., Chen, W., & Zhang, X. (2019). A hybrid model based on LSTM neural network for time series prediction. Soft Computing, 23(7), 2145-2156.