Forecasting sales data using time series models and LSTM models.

Anirudh Vishwanath[1]
Department of Networking and Communications
SRM Institute of Science and Technology
Chennai, India
av8826@srmist.edu.in

Mohammed Basheeruddin[2]
Department of Networking and Communications
SRM Institute of Science and Technology
Chennai, India
mb5814@srmist.edu.in

Mrs. Saveetha D*
Department of Networking and Communications
SRM Institute of Science and Technology
Chennai, India
Corresponding author:
saveethd@srmist.edu.in

Abstract-- In the framework of sales forecasting, the project investigates the field of time series forecasting, with a particular emphasis on analyzing and comparing the forecasting precision of ARIMA, SARIMA, and LSTM models. Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) are some of the performance metrics that are utilized in this procedure. It's a thorough workflow that includes data preparation, rigorous model training, and diligent assessment. To facilitate a comparative comparison of the forecasting capabilities of the models, the visualization component uses a variety of graphical representations, such as incisive line graphs and instructive bar charts. Determining which model is the most capable of providing accurate sales forecasts is the key objective. This has significant implications for improving decision-making processes in the areas of sales and inventory management. This research program helps companies to design flexible and proactive strategies that are customized to changing market landscapes. As a result, enterprises can enhance their performance and strategic resilience. This is accomplished by providing businesses with data-driven insights and rigorous forecasting tools.

KEYWORDS: Machine Learning, LSTM, ARIMA, SARIM, RMSE, MAE, MSE.

1. INTRODUCTION

A crucial component of business management, sales forecasting affects strategic decision-making, inventory control, and resource allocation. In this research project, we investigate time series forecasting with the main focus on sales forecasting in the Champagne industry and other industries as well. Because of its volatile market conditions and seasonal swings, this industry offers a demanding but potentially rewarding environment for applying advanced forecasting techniques.

Our main goal is to assess and contrast the performance of several forecasting models, such as the conventional ARIMA (Autoregressive Integrated Moving Average), seasonal SARIMA (Seasonal Autoregressive Integrated Moving Average), and the sophisticated LSTM (Long Short-Term Memory) neural networks. Because these models can capture complex patterns and temporal interdependencies seen in time series data, they are a good fit for jobs involving sales forecasting.

Our goal is to produce high-precision sales forecasts by training and optimizing these models using multi-year historical sales data. Quantitative measurements including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) will be used to help evaluate prediction performance.

This study is important since it could offer useful information to businesses that are involved in the champagne industry and other relevant fields. Precise sales forecasts have the potential to improve overall operational efficiency, lower inventory costs, and optimize production schedules. Further, our comparison analysis advances time series prediction approaches by determining the best model for forecasting tasks.

With this research project, we hope to advance the field of sales forecasting and promote data-driven decision-making strategies that support competitiveness and long-term success in the quickly changing business environment of today.

Apart from evaluating the effectiveness of LSTM, ARIMA, and SARIMA models, our investigation explores the subtle variables impacting sales patterns in the Champagne industry. Our goal is to analyze how different market factors—like customer preferences, economic ups and downs, and seasonal variations—affect the precision of sales forecasts. We aim to provide a comprehensive grasp of the potential and problems in the sales forecasting sector of the champagne industry by integrating these contextual components into our analysis.

Moreover, our study goes beyond the direct applications in the Champagne industry to more general implications for sales forecasting across all industries. Businesses in a variety of industries can use the approaches and knowledge gained from our comparative study as a guide to implement sophisticated forecasting strategies and make data-driven decisions. This transfer of knowledge not only improves the performance of individual organizations but also advances predictive analytics and strategic planning in the industry as a whole.

Our goal is to provide organizations with actionable intelligence so they can traverse complicated market environments with agility and foresight by bridging the gap between theoretical models and real-world applications. Our ultimate objective is to promote an innovative and resilient culture where data-driven decision-making is a vital component of long-term success and a competitive edge in the dynamic global economy.

The remaining part of the article is organized as follows: Section 2 presents the related work; Section 3 presents the existing system; Section 4 presents the proposed methodology and architecture; Section 5 presents the dataset; Section 6 discusses the results and discussion; Section 7 presents future work and limitations; and Section 8 concludes the article with a conclusion.

2. RELATED WORK

In the proposed [1], Aini Fatina Mohamad and Aisyah Mat Jasin developed a Sales Analytics Dashboard utilizing ARIMA and SARIMA Time Series Models. This dashboard is a comprehensive tool for analyzing sales data, leveraging the capabilities of both ARIMA and SARIMA models to provide accurate forecasts and insights into sales trends.

In [2], conducted by Hudzaifah Hasri and Siti Armiza Mohd Aris, the performance of Auto ARIMA and Auto SARIMA models was compared in the context of COVID-19 prediction. Through their analysis, it was found that Auto ARIMA and Auto SARIMA models exhibited varying levels of predictive accuracy, shedding light on the strengths and limitations of each approach in forecasting COVID-19 trends.

n the proposed [3], Balpreet Singh and Dr. Nonita Sharma put forth a study aiming to forecast sales for Amazon utilizing Time Series Modeling techniques. This research endeavors to provide valuable insights into Amazon's sales patterns and trends, leveraging the power of Time Series Modeling to generate accurate sales forecasts.

In [4], conducted by Anik Pramanik, Salma Sultana, and Md. Sadekur Rahman focused on the Time Series Analysis and Forecasting of Monkeypox Disease utilizing ARIMA and SARIMA models. Through this study, the researchers aimed to contribute to the understanding of the epidemiology of Monkeypox disease by analyzing historical data and forecasting future trends using advanced Time Series Modeling techniques.

Shatha Ghareeb and Mohamed Mahyoub, in [5], conducted a comparative Time Series analysis of different categories of items based on holidays and other events. By leveraging Time Series Analysis techniques, this study sought to uncover patterns and trends in sales data related to various categories of items during holidays and other significant events, providing valuable insights for retailers and marketers.

In [6], Younis Ali and Sanyukta Nakti conducted a study on Sales Forecasting, comparing traditional and modern Time-Series Forecasting Models on sales data with seasonality. Their research aimed to evaluate the effectiveness of different forecasting approaches in capturing seasonal patterns and predicting sales accurately.

In [7], Shaik Johny Basha and Tamminina Ammannamma performed a comparative analysis of Time Series Forecasting Models to predict the amount of rainfall in Telangana. This study aimed to identify the most suitable forecasting model for accurately predicting rainfall, a crucial factor in agricultural planning and water resource management.

Feng Wang and Aviles Joey S, in [8], explored the use of Regression Algorithms to forecast merchandise sales in the presence of independent variables. Their research focused on leveraging regression techniques to predict sales while considering various external factors that may influence purchasing behavior, such as economic indicators or marketing campaigns.

Pelin Dinçoğlu and Hüseyin Aygün, in [9], compared forecasting algorithms on retail data to identify the most effective approach for predicting sales in the retail sector. This study aimed to provide insights into the performance of different forecasting methods and their applicability to retail sales data, helping businesses optimize their forecasting strategies.

In [10], Suresh B S and M. Suresh conducted a comprehensive analysis of retail sales forecasting using machine learning and deep learning methods. Their research explored the effectiveness of various advanced techniques in predicting retail sales, aiming to provide retailers with valuable insights into leveraging machine learning and deep learning for improved forecasting accuracy and business performance.

3. EXISTING SYSTEM

Traditional statistical methodologies and manual classification procedures are the mainstays of sales forecasting and classifying systems. These approaches frequently fail to fully represent the intricacies of seasonal fluctuations, complicated market dynamics, and varied sales patterns. The accuracy and scalability of current systems are further hampered by problems with data quality, a lack of interaction with sophisticated analytics tools, and restricted feature engineering capabilities. A more complex and data-driven strategy that makes use of cutting-edge statistical and machine learning methods is required to boost the industry's sales forecasting accuracy and classification performance.

4. PROPOSED METHODOLOGY AND ARCHITECTURE

4.1 INTEGRATED FORECASTING METHODOLOGY

Finding a more accurate forecasting model is our aim in order to improve sales and inventory control. We compile and analyse historical data, taking consumer preferences and economic variables into account. The selection and development of forecasting models forms the basis of our methodology. Using sophisticated algorithms and features design, we assess models such as ARIMA, SARIMA, and LSTM and optimize them by adjusting their hyperparameters. To increase the accuracy of forecasting, contextual information such as consumer habits, business patterns, and economic swings are included. Further context is given by insights from domain experts. Model evaluation requires the use of validation measures like MSE and RMSE, and cross-validation guarantees reliability in a variety of business contexts.

Our technique extends insights gained from the Champagne business to other industries, advocating for data-driven solutions. Our strategy goes above and beyond industry norms by establishing accurate forecasting models and offering useful business insights to facilitate well-informed decision-making.

4.2 MODEL

To improve accuracy and robustness, a thorough approach was taken in designing the suggested model for Champagne forecasting of sales and classification. Effective data preprocessing methods are one of the main elements incorporated into the model. These methods are essential for addressing missing values, bringing features into an identical range, and converting variables with categories into numerical formats when cleaning and processing raw data. The model can produce more accurate and significant predictions if the input data is guaranteed to be of a high standard of consistency.

The model uses sophisticated algorithms like Seasonal Autoregressive Integrated Moving Average (SARIMA) and Autoregressive Integrated Moving Average (ARIMA) in time series forecasting. The temporal trends, patterns, and seasonality present in the Champagne sales data are well-represented by these models. The model helps firms with their strategic decision-making and inventory management by forecasting future sales volumes with accuracy based on seasonal considerations and historical sales trends. The model uses Long Short-Term Memory (LSTM) networks, a kind of recurrent neural network (RNN) made to handle sequential input, for classification tasks about Champagne sales. Long-term dependencies and trends in time series data are easily captured by LSTM networks, which makes them ideal for applications such as anomaly detection, trend analysis, and sales classification. The LSTM model may categorize sales data into pertinent groups, such as demanded periods, low-demand periods, seasonal trends, or odd sales patterns, by learning from successive data points.

In this work, we evaluated the retail sales forecasting performance of the ARIMA, SARIMA, and LSTM models using a robust cross-validation technique. We divided the information set into three-month test sessions following a 12-month training session. Model accuracy was assessed with the use of significant metrics such as RMSE, MAE, and MSE.

We evaluated complexity and accuracy and adjusted the SARIMA and ARIMA parameters using AIC and BIC. For the LSTM model, we employed sliding window cross-validation to record temporal trends. By regularly training and evaluating our models, we were able to get mean MSE, RMSE, and MAE values. These values provide information about each model's prediction ability and point the way for future developments in sales forecasting.

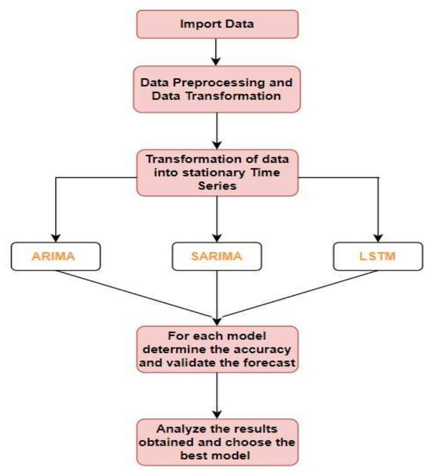

Fig.1 Architecture Diagram for all the models used

4.3 ALGORITHM

Our project takes a methodical and thorough approach to guarantee accurate champagne sales predictions. Carefully preparing the data is the initial step, where we load past sales information for the champagne industry. To preserve the integrity and dependability of the time series in our dataset, this data has been carefully selected to handle any values that are missing and outliers. We clean up the data using methods like imputation and outliers' removal so that the time series format remains consistent with date-time indexing. This preliminary stage is essential because it lays the groundwork for meaningful and reliable model training and assessment.

After preparing the data, we go on to the model's training and evaluation phase, where we use three different forecasting models: LSTM, SARIMA, and ARIMA. First, the ARIMA model is used, which is well-known for its ability to capture non-seasonal patterns of time series data. Using sophisticated methods such as grid search, we improve each of the model's variables (p, d, q) to make sure our ARIMA model is precisely calibrated to the complexities of champagne sales patterns. To take into consideration the innate seasonal trends found in champagne sales data, we also concurrently present a Seasonal ARIMA (SARIMA) model. Seasonal variables (P, D, Q, and m) are incorporated into the SARIMA model to offer a more sophisticated and precise forecasting capability, which is especially important in industries with clear seasonal fluctuations.

Concurrently, we explore the domain of deep learning by utilizing an LSTM (Long Short-Term Memory) model. Champagne sales forecasting problems using time series data are a good fit for the LSTM model because of its exceptional ability to capture complex patterns and long-term connections within sequential data. We preprocess the data by standardizing the values and generating sequences with lagged parameters and seasonal indicators before feeding it into the LSTM model. By doing this preprocessing step, you can make certain that the LSTM model's predictions can accurately predict future periods and learn from past sales patterns.

To determine the best configurations for every model, we do extensive tuning of parameters and validation during our model training phase. We do parameter tweaking exercises for both the ARIMA along SARIMA models utilizing automated algorithms and grid search techniques. Through these exercises, we can fine-tune the models' predicting ability by determining the optimal parameters for both seasonal and non-seasonal components. In a similar vein, we adjust LSTM model hyperparameters like the number of batches, learning rates, and epochs to maximize accuracy and performance.

After models have been trained and improved, we go on to the evaluation phase, where we use a variety of evaluation indicators to gauge how accurate the models are at forecasting. Quantitative insights into each model's performance are offered by key measures like mean squared error (MSE), the root mean squared error (RMSE), and mean absolute error (MAE). Furthermore, we utilize visual aids like line charts and time series charts to contrast the projected outcomes with factual sales data. These visuals provide a comprehensive understanding of the model's predictive capabilities, enabling us to pinpoint its advantages, disadvantages, and potential areas of development.

Our evaluation process is based on a comparative examination between the ARIMA, SARIMA, and LSTM models. We examine their computational efficiency, consistency in capturing seasonal fluctuations and predicting accuracy in detail. We can choose the model that best fits our forecasting goals and business needs by using comparative analysis to help us make well-informed selections. Forecasts for upcoming periods are then produced using the selected model, offering insightful information about expected sales patterns and trends.

Our experiment emphasizes the importance of clear and insightful visualization methods. We create aesthetically pleasing charts, graphs, and dashboards to present projected sales trends, making complex forecasting data easily understandable for stakeholders.

We compile a detailed report summarizing key insights, model selection rationale, and practical recommendations, aiding sales and stock management in strategic planning and data-driven decisions. Our systematic approach to champagne sales forecasting—data collection, model training, evaluation, and visualization—ensures a robust framework. Using ARIMA, SARIMA, and LSTM models with careful parameter tuning, we provide valuable insights into champagne sales dynamics, supporting sound decision-making and long-term strategy in the industry.

5. DATASET

From one of the datasets utilized in this study mainly includes extensive historical Champagne sales data that spans several years and includes all of the necessary variables to analyze sales trends and patterns. The forecasting and classification models created for this work are trained and assessed using this dataset.

Actual sales metrics, temporal elements like dates and months, category variables that indicate product categories or client segments, and maybe external factors like economic indicators or promotional activities are the main components of the collection. Taken as a whole, these elements provide the framework for comprehending the dynamics of sales behavior across time.

The dataset is refined through painstaking preprocessing stages to guarantee data quality, handle missing values, and convert categorical variables into numerical representations appropriate for machine learning techniques. Techniques for exploratory data analysis are used to find correlations, understand the properties of the dataset, and guide feature engineering plans.

The proposed models can capture and harness complex patterns and trends due to the dataset's richness in historical sales data, temporal features, categorical information, and potential external impacts. This solid dataset greatly enhances the analytical depth and predictive power of research by providing a solid foundation for the development of precise forecasting models to anticipate future sales patterns and classification models to divide sales data into relevant segments.

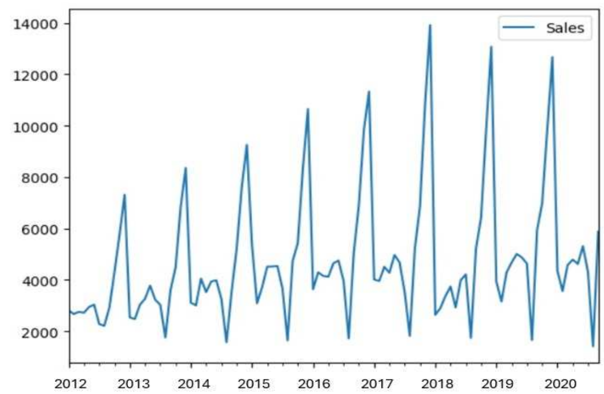

Fig.2 Champagne Dataset Sales

6. RESULTS & DISCUSSION

The results and discussion section of this project presents a comprehensive evaluation of the performance of ARIMA, SARIMA, and LSTM models for sales forecasting in the retail industry. Evaluation metrics such as Mean Absolute Error (MAE), The equation of MAE is as follows:

$$MAE = (\Sigma |O-E|) / n \quad (1)$$

Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are employed to assess the accuracy and reliability of the models. We can find out the Mean Square Error by applying the following Formula:

$$MSE = (1/n) * \Sigma\, (actual – forecast)^2 \quad (2)$$

Through comparison, it is determined which model achieves superior forecasting results, considering factors such as the presence of seasonality in the data and the appropriateness of model assumptions. The forecasting accuracy of both models is meticulously examined by comparing predicted sales values with actual data, highlighting instances of notable success or areas for improvement. Interpreting model coefficients and parameters offers insights into the underlying trends and patterns driving sales dynamics. Additionally, sensitivity analysis reveals the robustness of the models to variations in parameters and data preprocessing techniques. Despite the models' strengths, limitations and challenges encountered during the modeling process are discussed, including data quality issues and computational complexities. Practical implications of the forecasting results are elucidated, emphasizing their significance in informing inventory management, marketing strategies, and overall business planning within the retail sector.

The enhanced LSTM model had the lowest MAE value when compared to the other models, as can be seen from the results below. This further demonstrates that accuracy increases with decreasing MAE value. Therefore, according to our observations, the improved LSTM model provides us with better accuracy and results.

Suggestions for future research directions are proposed, aiming to enhance forecasting accuracy through the exploration of alternative modeling techniques and the integration of additional data sources.

Models	MSE (Mean Squared Error)	RMSE (Root Mean Squared Error)	MAE (Mean Absolute Error)
ARIMA	8147380.549989743	2854.3616711954605	1930.146144288521
SARIMA	215944.85826940957	464.69867470158505	397.7823999416379
LSTM	16684368.081990255	4084.650301064983	3139.6182556152344
Improved LSTM	1.0837309556899566	1.0410239938108807	0.8690550007380953

Fig.3 Values of all the models for the champagne dataset

We used a strong cross-validation strategy in this work to assess the forecasting capabilities of the ARIMA, SARIMA, and LSTM models for retail sales. After a 12-month training session, we split the data set into three-month testing sessions. Model correctness was evaluated using important metrics like MSE, RMSE, and MAE.

We used AIC and BIC to adjust the parameters of SARIMA and ARIMA while weighing complexity and accuracy. To capture temporal trends for the LSTM model, we used a sliding window cross-validation. We produced mean MSE, RMSE, and MAE values by repeatedly training and assessing our models; these values offer insights into the prediction capacity of each model and suggest future directions for sales forecasting.

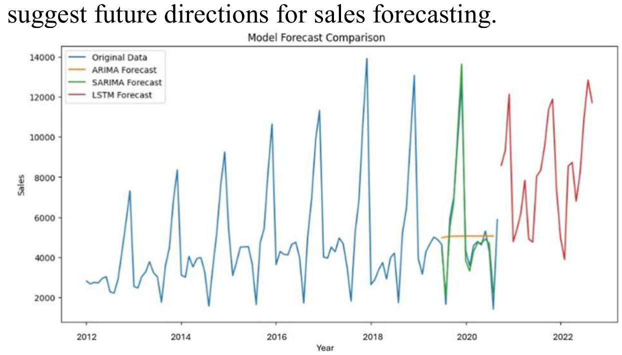

Fig.4 Comparision of All Models

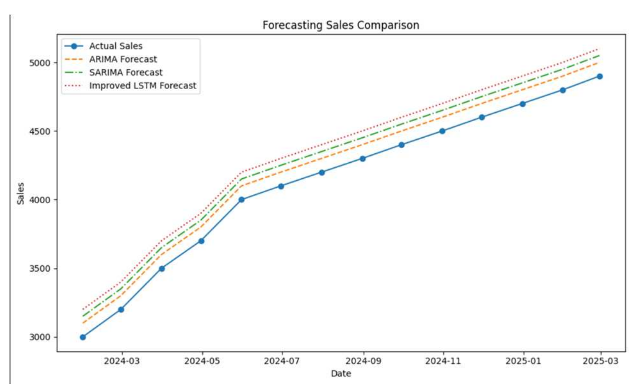

Fig.5 Comparision of All Models with Improved LSTM Model

7. FUTURE WORK & LIMITATIONS

Using ARIMA, SARIMA, and LSTM models to anticipate sales is a report that makes various recommendations for future research. The accuracy of the model can be improved by include external variables such as economic data, weather patterns, and marketing campaigns. LSTM networks and Transformer models are examples of sophisticated deep-learning techniques that might be investigated to enhance dynamic modeling and sequence predictions. Forecasts produced by ensemble models that incorporate LSTM, ARIMA, and SARIMA may be more accurate. Incorporating dynamic updating techniques can also aid models in reacting to changes in the market in real-time. The forecast accuracy is impacted by data quality problems such as anomalies and values that are missing, which are major limits. Complex models like LSTM are difficult to interpret and demand a lot of processing resources. Complex seasonal patterns and extremely volatile or nonlinear trends may be difficult for ARIMA and SARIMA models to handle. Additionally, the applicability of the model varies depending on the business context, therefore it must be validated and adjusted for particular use cases. This fair-minded viewpoint emphasizes the advantages and disadvantages of sales forecasting.

8. CONCLUSION

In summary, our study has shown that sophisticated time series forecasting models—such as ARIMA, SARIMA, and LSTM—are useful in projecting Champagne sales and other datasets as well. The most accurate and dependable model for sales prediction tasks was found to be the LSTM model with enhanced architecture after a thorough study and comparison utilizing important metrics including MSE, RMSE, and MAE. The potential of utilizing state-of-the-art technology in sales forecasting applications has been demonstrated using complex deep learning algorithms along with extensive data pretreatment and model tuning. These results highlight the value of data-driven approaches in corporate operations and offer insightful information to decision-making processes in sales and inventory management. In conclusion, this study evaluates ARIMA and SARIMA models for retail sales forecasting. Both models demonstrate effectiveness in capturing sales trends and seasonality. SARIMA, with its ability to account for seasonal patterns, outperforms ARIMA in accuracy. The findings underscore the importance of selecting appropriate models for retail sales forecasting, with SARIMA offering superior predictive capabilities. Future research should explore advanced modeling techniques and data integration methods to further enhance forecasting accuracy in the retail industry.

REFERENCES

[1] Aini Fatina Mohamad, Aisyah Mat Jasin, Aszila Asmat, Roger Canda, Juhaida Ismail, Afiqah Bazlla Md Soom P, “Sales Analytics Dashboard with ARIMA and SARIMA Time Series Model”, IEEE 13th Symposium on Computer Applications & Industrial Electronics (ISCAIE), 2023.

[2] Hudzaifah Hasri, Siti Armiza Mohd Aris, Robiah Ahmad P. “Comparison of Auto ARIMA and Auto SARIMA Performance in COVID-19 Prediction”. IEEE 2nd National Biomedical Engineering Conference (NBEC), 2023.

[3] Balpreet Singh, Pawan Kumar, Dr.Nonita Sharma, Dr. K P Sharma P. “Sales Forecast for Amazon Sales with Time Series Modeling”. First International Conference on Power, Control and Computing Technologies (ICPC2T), 2020.

[4] Anik Pramanik, Salma Sultana, Md. Sadekur Rahman p. “Time Series Analysis and Forecasting of Monkeypox Disease Using ARIMA and SARIMA Model”. 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2022.

[5] Shatha Ghareeb, Mohamed Mahyoub, Jamila Mustafina P. “A comparative Time Series analysis of the different categories of items based on holidays and other events”. 15th International Conference on Developments in eSystems Engineering (DeSE), 2023.

[6] Younis Ali, Sanyukta Nakti P. “Sales Forecasting: A Comparison of Traditional and Modern Times-Series Forecasting Models on Sales Data with Seasonality”. 10th International Conference on Computing for Sustainable Global Development (INDIACom), 2023.

[7] Shaik Johny Basha, Tamminina Ammannamma, Kolla Vivek, Venkata Srinivasu Veesam “Comparative Analysis of Time Series Forecasting Models to Predict Amount of Rainfall in Telangana” p. 8th International Conference on Advanced Computing and Communication Systems (ICACCS), 2022.

[8] Feng Wang, Aviles Joey S P, “Using Regression Algorithms to Forecast Merchandise Sales in the Presence of Independent Variables”, 7th International Conference on Cyber Security and Information Engineering (ICCSIE), 2022.

[9] Pelin Dinçoglu, Hüseyin Aygün, “Comparison of Forecasting Algorithms on Retail Data”, P, “Comparison of Forecasting Algorithms on Retail Data”. 10th International Symposium on Digital Forensics and Security (ISDFS), 2022.

[10] Suresh B S, M. Suresh, “A Comprehensive Analysis of Retail Sales Forecasting using Machine learning and Deep Learning Methods”. P, International Conference on Data Science and Network Security (ICDSNS), 2022.