Sales Forecasting using Machine Learning Algorithms

*Time Series Analysis

1st Gunisha Giri
AU2020210
Bachelor of Science (Computer Science)
Ahmedabad University
gunisha.g@ahduni.edu.in

2nd Freya Shah
AU2020039
Bachelor of Science (Computer Science)
Ahmedabad University
freya.s@ahduni.edu.in

3rd Soham Jagrit
AU2040060
B.Tech (Computer Science)
Ahmedabad University
soham.j@ahduni.edu.in

4th Nitya Panchal
AU2040249
B.Tech (Computer Science)
AHmedabad University
nitya.p2@ahduni.edu.in

***Abstract*—This paper presents the findings of a sales forecasting study that included machine learning techniques including Triple Exponential Smoothing(TES), Autoregressive Integrated Moving Average(ARIMA), and Seasonal ARIMA. The research seeks to forecast a retail company's revenues for the following months using past data. The TES, SARIMA, and ARIMA machine learning techniques were utilized to construct the sales forecasting models. SARIMA and ARIMA were employed to capture the seasonality and trend of the data. We have measured the accuracy using the Mean Absolute Percentage Error (MAPE). The outcomes demonstrated that SARIMA and TES performed better than ARIMA, with SARIMA having the greatest performance. In order to help stakeholders comprehend the patterns and make data-driven decisions, the projections were then graphically represented. Overall, this project shows how well machine learning algorithms work for predicting sales and gives organizations helpful information to enhance their planning and decision-making procedures.**

***Index Terms*—sales, data, ARIMA, regression, accuracy, forecasting, SARIMA, prediction, linear, collinear, auto-correlation.**

I. Introduction

Sales forecasting is a crucial component of strategic planning that helps businesses to plan effectively for their operations and resource allocation. Accurate sales forecasting enables businesses to prepare for future development and expansion while also assisting them in optimizing their production, inventory, and personnel levels. However, sales forecasting may be difficult because of the complexity and variety of the underlying elements that affect sales. Time series analysis and machine learning algorithms have become effective tools for sales forecasting in recent years. While machine learning algorithms utilize statistical models to create predictions based on historical data, time series analysis examines and analyses the temporal patterns in data. Both methods can capture the intricate correlations and patterns seen in sales data, making them useful tools for sales forecasting. This research report investigates how to anticipate sales for a retail organization using time series analysis and machine learning techniques. With regard to estimating sales for the following months, we specifically examine the effectiveness of three well-known machine learning algorithms: Triple Exponential Smoothing, ARIMA, and SARIMA. For businesses to make wise choices regarding inventory management, marketing initiatives, and personnel levels, accurate sales forecasting is essential. Inefficient use of resources, missed opportunities and lost money can all result from inaccurate estimates. Traditional forecasting strategies frequently depend on previous sales data and straightforward statistical procedures, including exponential smoothing and moving averages. Seasonality, trends, and cycles are only a few examples of the temporal patterns in data that may be examined and modeled using time series analysis. Time series models may be used to make predictions and calculate the ranges of uncertainty surrounding such predictions. On the other hand, machine learning algorithms create predictions based on past data using statistical models. These algorithms are effective tools for sales forecasting because they can identify intricate patterns and linkages in the data. With regard to estimating sales for the following months, we specifically examine the effectiveness of three well-known machine learning algorithms: TES, ARIMA and SARIMA. We then use a performance metric which is MAPE to evaluate the models. These measures allow us to evaluate the performance of the various models and offer a gauge of how accurate the forecasts are. Depending on the peculiarities of the sales data, the algorithms' performance varies. ARIMA and SARIMA are

Identify applicable funding agency here. If none, delete this.

created to capture the temporal patterns and seasonality in the data and function effectively when the sales data display these features. This project's main aim is to show how time series analysis and machine learning algorithms work well for predicting sales. We offer insights into which algorithms are most suited for various forms of sales data by contrasting the performance of three well-known algorithms.

II. Literature Survey

Sales forecasting is a critical component of corporate planning and in recent years there has been a lot of study in this area. Time series analysis and machine learning algorithms have become well-liked techniques for sales forecasting, enabling companies to estimate future sales with accuracy. A description of some of the important works in this field is provided in the literature review that follows. One of the most commonly used statistical methods for sales forecasting is regression, which involves modeling the relationship between sales and various factors that influence it. ARIMA and ARMA[7] are two other popular techniques that are widely used for time series forecasting. However, these statistical methods have limitations when it comes to handling complex data with external and internal factors. A hybrid seasonal quantile regression[8] approach and ARIMA were proposed for daily food sales forecasting in. In this, sales forecasting was performed on newly published books in an editorial business management environment using computational methods. However, the literature in this field suggests that not much work has been done on using swarm intelligence techniques to train prediction models . Genetic Algorithm (GA)[9] is a potential candidate for training ANN models.

In a research published in 2011, De Livera and colleagues assessed the effectiveness of several time series models for predicting data on retail sales. SARIMA models beat ARIMA models in the study's comparison of their performances in capturing seasonal variation in sales data. According to the study, SARIMA models were more effective in forecasting sales data with seasonality.

An extensive overview of time series forecasting techniques, including ARIMA and its variants like SARIMA, was published in a research by Hyndman and Khandakar in 2008. The study came to the conclusion that SARIMA models were good for predicting data with a seasonal trend whereas ARIMA models were acceptable for stationary trend data.

The literature review emphasises the use of time series analysis and machine learning techniques for sales forecasting. The research examined in this review show that several machine learning techniques, including decision trees, artificial neural networks and ensemble learning methods are excellent in forecasting sales data. Additionally, the study demonstrates that the ARIMA and SARIMA models are efficient in identifying temporal trends and seasonality in sales data. These results might help firms choose the right algorithms and models for their activities including sales forecasting.

III. Implementation

We started with checking if the data is stationary or not. For any time series modeling, it is a prerequisite to make the data stationary if it is not. To check whether the data is stationary or not, we performed the Augmented Dickey–Fuller test. The result of this test suggested that the data is stationary. Henceforth, after plotting the data, it is clearly visible that the data has trends and seasonality. To understand this better, we decomposed the data using the seasonal$_{d}ecompose()functionandplottedit.Belowistheplot$:

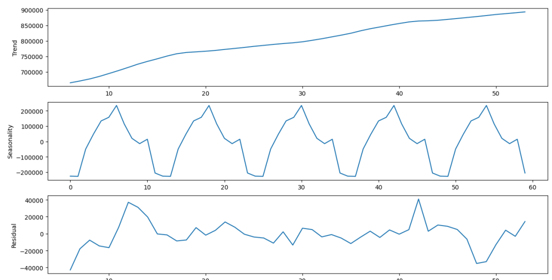

Fig. 1. Seasonal Decomposition

We examined the autocorrelation and partial autocorrelation plots in order to get the SARIMA model's ideal parameters. The model was subsequently developed using the training set of data and predictions was then made using the test set.

IV. Results

After having a clear understanding of how the data is changing with respect to time, we selected three models to implement. The models are Triple exponential smoothing, ARIMA, and Seasonal ARIMA (SARIMA).

1. Triple Exponential Smoothing: We have divided our data in 80:20 train-test ratios for all the algorithms. For TSE, we are required to enter the values for 'trend', 'seasonal', and'seasonal periods'. According to the above decomposition graph, we can easily interpret that the trend is additive (not exponential), which means that the trend is linear to some extent. We can also see that the seasonality is also additive and the period is 12 months. Below is the output of the TES implementation:

2. ARIMA: ARIMA is an exceptional model for time series analysis. However, it does not work well with seasonality in the data. As we are required to find the values of p, d, q for the implementation, we made the Autocorrelation and Partial autocorrelation plots for finding q and p respectively. The hyperparameter d will have a value of 1 as to remove seasonality, differencing is needed to be performed. To make sure the values obtained are correct, we also checked it with the inbuilt function of stattools which is auto$_{a}rima().Theorderwecameupwithwasp = 1, q = 2, andd = 1$.

We can clearly see that the model is not accurate. As mentioned above, because of the seasonality in the data, it

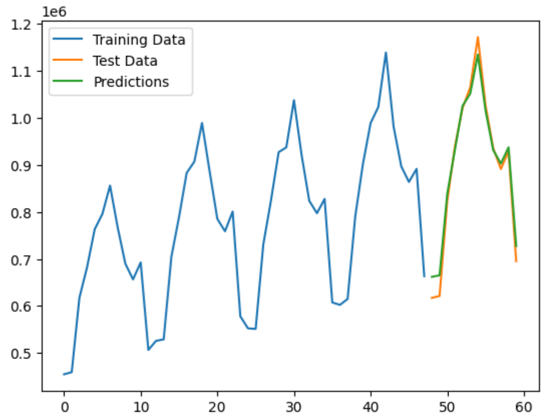

Fig. 2. Output of Triple Exponential Smoothing

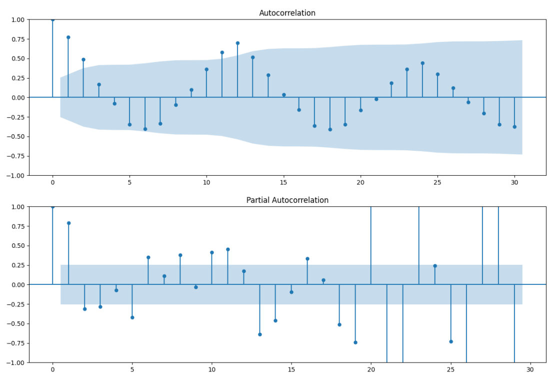

Fig. 3. Autocorrelation and Partial AUtocorrelation

does not perform well. Hence, it would be a better choice to use SARIMA here.

3. SARIMA: In this model, the p, d, and q discussed above remain the same, just a new parameter in which the period of seasonality is needed to be added. Here, as the pastern gets repeated after 12 months, we used 12 as the period. Below is the plot of sales predicted by SARIMA:

Performance analysis: To check whether the models performed well, we used the Mean Absolute Percentage Error (MAPE) metric which is each time period minus actual values divided by actual values. Below is a comparison of the performance of the three models.

Model	MAPE Value
TSE	0.02510657
ARIMA	0.35141871
SARIMA	0.01330408

Conclusion

This project's goal was to determine if time series analysis and machine learning techniques, especially TES, ARIMA, and SARIMA models, are useful for predicting sales. We used a dataset of daily sales numbers for a retail firm for the last few years. Pre-processing the data, choosing acceptable performance indicators, and assessing the effectiveness of the

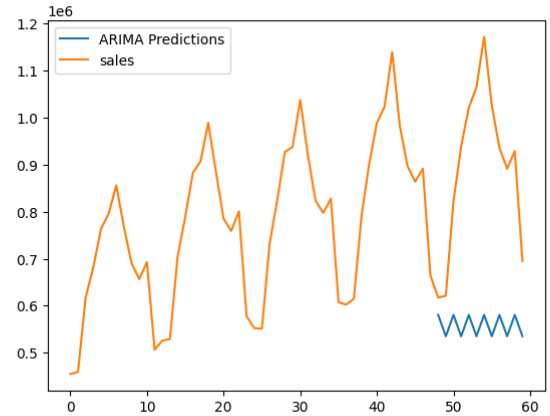

Fig. 4. ARIMA Model Predictions

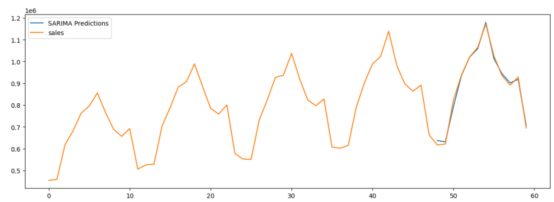

Fig. 5. SARIMA Model Predictions

three models were all parts of our investigation. We discovered that each of the three models could estimate sales accurately, but their performance varied depending on the particulars of the sales data.

ARIMA and SARIMA models were created to capture temporal patterns and seasonality in the data, which makes them suitable for sales forecasting. In our investigation, the SARIMA model performed better than the ARIMA and TES models. Our findings also demonstrated how crucial it is to pre-process data, pick sensible performance indicators, and visualize forecasts in order to support decision-making. We advise businesses to adhere to best practices in sales forecasting, which emphasize the use of high-quality data, feature engineering, and model selection, as well as the use of time series analysis and machine learning algorithms.

Our findings also shed light on the suitability of various models for various kinds of sales data. Using this information, businesses may choose the optimal model for their unique business requirements and data characteristics. For businesses trying to increase the accuracy of their sales forecasting and make better decisions regarding their operations and resources, this project report might be a useful resource. Future research should focus on overcoming the limitations of our study. For instance, since we only examined one dataset, our findings might not apply to data from other sources. In addition, we only employed three machine learning algorithms; other algorithms could be more effective in predicting sales.

In conclusion, this research shows how time series analysis and machine learning algorithms may be used to accurately predict sales. We offer insights into which algorithms are most suited for various forms of sales data by contrasting the performance of three well-known algorithms. Companies may increase the accuracy of their sales forecasting and make better judgments about their operations and resource allocation by following our best practices and recommendations.

REFERENCES

[1] B. Singh, P. Kumar, N. Sharma and K. P. Sharma, "Sales Forecast for Amazon Sales with Time Series Modeling," 2020 First International Conference on Power, Control and Computing Technologies (ICPC2T), Raipur, India, 2020, pp. 38-43, doi: 10.1109/ICPC2T48082.2020.9071463. http://toc.proceedings.com/53952webtoc.pdf

[2] Khan, R. A., Quadri, S. M. (2012). Business intelligence: an integrated approach. Business Intelligence Journal, 5(1), 64-70. https://parsmodir.com/wp-content/uploads/2018/11/birafi.pdf

[3] Sobreiro, P., Martinho, D., Pratas, A. (2018, June). Sales forecast in an IT company using time series. In 2018 13th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-5). IEEE. https://ieeexplore.ieee.org/document/8399191

[4] Nunnari, G., Nunnari, V. (2017, July). Forecasting monthly sales retail time series: a case study. In 2017 IEEE 19th conference on business informatics (CBI) (Vol. 1, pp. 1-6). IEEE. http://doi.ieeecomputersociety.org/10.1109/CBI.2017.57

[5] Cholette, P. A. (1982). Prior information and ARIMA forecasting. Journal of Forecasting, 1(4), 375-383. https://doi.org/10.1002/for.3980010405

[6] Smith, S. A., McIntyre, S. H., Achabal, D. D. (1994). A Two-Stage Sales Forecasting Procedure Using Discounted Least Squares. Journal of Marketing Research, 31(1), 44–56. https://doi.org/10.1177/002224379403100104

[7] A. S. Weigend and N. A. Gershenfeld, "Time series prediction: Forecasting the future and understanding the past", Addison-Wesley, 1994. https://scirp.org/reference/referencespapers.aspx?referenceid=2328927

[8] Mitra, A., Jain, A., Kishore, A., Kumar, P. (2022). A Comparative Study of Demand Forecasting Models for a Multi-Channel Retail Company: A Novel Hybrid Machine Learning Approach. Operations Research Forum, 3(4). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9514716/

[9] Hani Omar, Van Hai Hoang, Duen-Ren Liu, "A Hybrid Neural Network Model for Sales Forecasting Based on ARIMA and Search Popularity of Article Titles", Computational Intelligence and Neuroscience, vol. 2016, Article ID 9656453, 9 pages, 2016. https://doi.org/10.1155/2016/9656453

[10] Mitra, A., Jain, A., Kishore, A., Kumar, P. (2022). A Comparative Study of Demand Forecasting Models for a Multi-Channel Retail Company: A Novel Hybrid Machine Learning Approach. Operations Research Forum, 3(4). https://doi.org/10.1007/s43069-022-00166-4