Received: 24 March 2024 | Revised: 23 April 2024 | Accepted: 13 May 2024

Research Article
Journal of Research and Review in Science
1-16, Volume 11, June 2024
DOI:10.36108/jrrslasu/4202.11.0140

ORIGINAL RESEARCH

Application of Mathematical Modelling in Sales Forecast Using Time Series

Yakub Tunde OYEB[1], Azeez SULAIMON[2], Abdulafeez Olalekan ABDULKAREEM[1], Kazeem Adekunle SHONIBARE[1],

[1]*Department of Mathematics, Faculty of Science, Lagos State University, Nigeria*

[2]*Department of Mathematics, Faculty of Science, National Open University of Nigeria (NOUN)*

Correspondence
Yakub Tunde OYEBO, Department of Mathematics, Faculty of Science, Lagos State University, Nigeria.
Email:oyeboyt1@gmail.com

Abstract:

Introduction: Sales forecasting is a crucial aspect of business management, which involves predicting future sales based on historical data and market trends. Accurate sales forecasts are essential for effective decision-making, such as inventory management, production planning, and resource allocation.

Aims: This study explores the application of mathematical modelling in sales forecasting.

Materials and Methods: A case study approach was used to demonstrate how mathematical modelling can be deployed to develop accurate sales forecasts. Specifically, historical sales data and market trends were used to develop mathematical models, including regression analysis, and time series analysis.

Results: The ACF was obtained using seasonal first difference result which help us achieve a stationarity of the sales data which is suitable for ARIMA model analysis. The analysis revealed that, the Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) plots based on the values of our seasonal difference over 40 and 20 lags respectively.

Conclusion: The exploration of sales forecasting accuracy using historical data, STL, and an ARIMA model has provided valuable insights into the dynamics of forecasting for business optimization. The findings and recommendations from our analysis have shed light on both the strengths and areas for improvement in the current forecasting process.We observed a strong positive linear relationship (Pearson's correlation coefficient, $r \approx 0.8796$) between the actual and forecasted sales data. This indicates that the ARIMA model effectively captures the general trends in sales, which is a promising aspect of the forecasting system.

To Keywords: ARIMA, SARIMA, LSTM, domain-expertise, exponential-smoothing, seasonal-decomposition, STL, ADF and ACF

All co-authors agreed to have their names listed as authors.

1. INTRODUCTION

In the mid-20th century, pioneering works by Brown [1] and Holt [2] laid the foundation for mathematical modelling in sales forecasting. Brown introduced the exponential smoothing method, while Holt extended it to account for trends. These early approaches initiated the exploration of time series analysis for sales prediction.

The field progressed with the introduction of the AutoRegressive Integrated Moving Average (ARIMA) model by Box and Jenkins [3]. The authors achieved the great results in their paper by combining autoregressive and moving average components. And ever since, ARIMA has proved to be highly effective in handling non-stationary data, significantly enhancing the accuracy of sales forecasts.

As businesses encounter more complex scenarios, the need to consider seasonality and external factors in sales forecasting emerged. Gardner and McKenzie [4] proposed the Seasonal Autoregressive Integrated Moving Average (SARIMA) model, a natural extension of ARIMA, which accounted for seasonal patterns. Concurrently, Harvey [5] introduced the Transfer Function Model, enabling the incorporation of exogenous variables that influence sales.

With the rise of machine learning, researchers began exploring its application in sales forecasting. Cao et al. [6] demonstrated the effectiveness of Support Vector Machines (SVM) for time series forecasting, showing promising results in predicting sales patterns. Additionally, neural networks, particularly Long Short-Term Memory (LSTM) networks, gained popularity for their ability to capture intricate temporal dependencies, as shown in Zhang et al.'s study [7].

Recent research has focused on developing hybrid forecasting models, combining multiple techniques to leverage their strengths and improve accuracy. Wang and Wan [8] proposed a hybrid model that integrated ARIMA, LSTM, and exponential smoothing, surpassing standalone models in predicting sales for e-commerce platforms.

Based on the details at our disposal, it is observed that, work on PYTHON implemented SARIMA model is very few in the literature. But in this current year, Can Ozdogar [9] explores this approach to make forecast for the temperature change in Istanbul, Turkey.

2. METHODOLOGY

.

This study employs a quantitative research design to investigate how mathematical modelling can be effectively used in sales forecasting. The research process involves collecting historical sales data, selecting appropriate mathematical models, and rigorously evaluating the performance of these models in generating accurate sales forecasts.

This study's primary data source is the target business's historical sales data. This dataset will encompass details of past sales transactions, including product categories, pricing information, and any pertinent variables. In addition, secondary data will be gathered, such as economic indicators and external factors like weather data. This data will be employed to examine the potential advantages of incorporating external variables in the forecasting process.

The raw sales data will undergo a thorough cleaning process to eliminate outliers and missing values. Categorical variables will be appropriately encoded, and time-related features will be extracted. Continuous variables will be normalized to ensure that differing scales do not introduce bias during the modelling phase. The selection of relevant features for forecasting will be undertaken through methods such as correlation analysis and domain expertise.

Various time series models including ARIMA, exponential smoothing, and seasonal decomposition will be applied to capture the data's temporal trends and seasonal patterns. Regression-based models and panel data analysis will be utilized to explore the potential impact of economic indicators on sales performance.

The datasets will be divided into training and validation subsets, employing strategies like time-based splitting to preserve the chronological sequence of the data. The chosen models will be trained on the training datasets and subsequently validated using the validation datasets. Techniques like cross-validation will be employed to fine-tune model hyper-parameters.

Model performance will be assessed through pertinent forecasting metrics, including Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and forecast accuracy. Comparative analysis of models will be conducted to determine which approach yields the most accurate sales forecasts.

External data sources, like weather information, will be incorporated into the models to evaluate the potential enhancement in forecasting accuracy. Methods such as Monte Carlo simulations will be employed to quantify and effectively integrate uncertainty into the sales forecasts, providing decision-makers with a comprehensive range of potential outcomes.

The outcomes of the mathematical models will be interpreted to extract meaningful insights from the data. Visualizations like time series plots and prediction intervals will be generated to facilitate comprehension of the forecasted outcomes.

Sensitivity analysis will be conducted to explore how variations in key input variables or model assumptions influence the projected sales figures.

The study will address ethical concerns related to data privacy, fairness, and bias to ensure the research adheres to ethical standards.

3. RESULTS AND DISCUSSION

In this section, we present the results and findings of this study. We begin by summarizing the data collection process, followed by the analysis of the collected data.

3.1 Data Sources

For this project, we collected historical sales data from Modella Shoes Collection, Lagos, Nigeria, a leading retailer in our target industry. The dataset included daily sales figures for the past five years, spanning various products and locations. Additionally, we gathered relevant external factors such as economic indicators, seasonal patterns, and promotional events that might influence sales.

3.2 Data Preprocessing

Before proceeding with the analysis, we performed data preprocessing to ensure data quality and consistency. This included handling missing values, smoothing time series data, and removing outliers. We also transformed the data into a suitable format for modelling.

3.3 Analysis of Data

With the pre-processed data in hand, we proceeded to analyse it using various mathematical modelling techniques for sales forecasting. In this section, we outline the methods employed and their respective results.

3.4 Descriptive Statistics

We began by conducting descriptive statistics to gain insights into the overall sales data. This analysis provided an understanding of the data's central tendencies, variability, and distribution. Key findings from this phase included:

- Average daily sales over the past four years.
- Seasonal patterns and trends.
- Variability in sales during different days of the week and months.

The sales data collected for this project over the period of five years is given in the following table:

Table 1: Sales data from Jan. 2014 – Dec. 2018

Sales By Month/Year					
Month	**2014**	**2015**	**2016**	**2017**	**2018**
Jan	2111250	1905750	2334750	4031250	2724750
Feb	2004000	1856250	2254500	2316000	3219000
Mar	2066250	2273250	3035250	2788500	3115500
Apr	2040750	2449500	2642250	3385500	3090750
May	2209500	2832000	2952750	3390000	3485250
Jun	2277000	2422500	2989500	3404250	3564750
Jul	1711500	2271000	2445000	2747250	2973750
Aug	1659000	1319250	1179750	1232250	1292250
Sep	2191500	2696250	2646000	3554250	3786000
Oct	3225750	3355500	3908250	4071000	5191500
Nov	4323000	5128500	5710500	6235500	7393500
Dec	5484000	6267750	6940500	7988250	8498250

The descriptive statistics of this given data can be described as follow:

Here are the calculated descriptive statistics for the Modella Shoes Collection sales data (Naira):

- Mean Sales: 3,106,171.88 (approximately)
- Median Sales: 2,926,125
- Standard Deviation: 1,866,470.59 (approximately)
- Minimum Sales: 1,179,750
- Maximum Sales: 8,498,250
- Lower Quartile (Q1): 2,265,375
- Upper Quartile (Q3): 3,591,937.50
- Interquartile Range (IQR): 1,326,562.50

These statistics provide a summary of the central tendency, spread, and distribution of the Modella Shoes Collection sales data over the four-year period, allowing for a better understanding of the sales performance.

3.5 Time Series Decomposition

To separate the underlying components of the time series data, we performed a decomposition analysis. This allowed us to identify and isolate the following components:

- Trend: The long-term sales growth or decline
- Seasonal: The recurring patterns in sales associated with different time intervals (monthly).

Fig. 1: Seasonal pattern

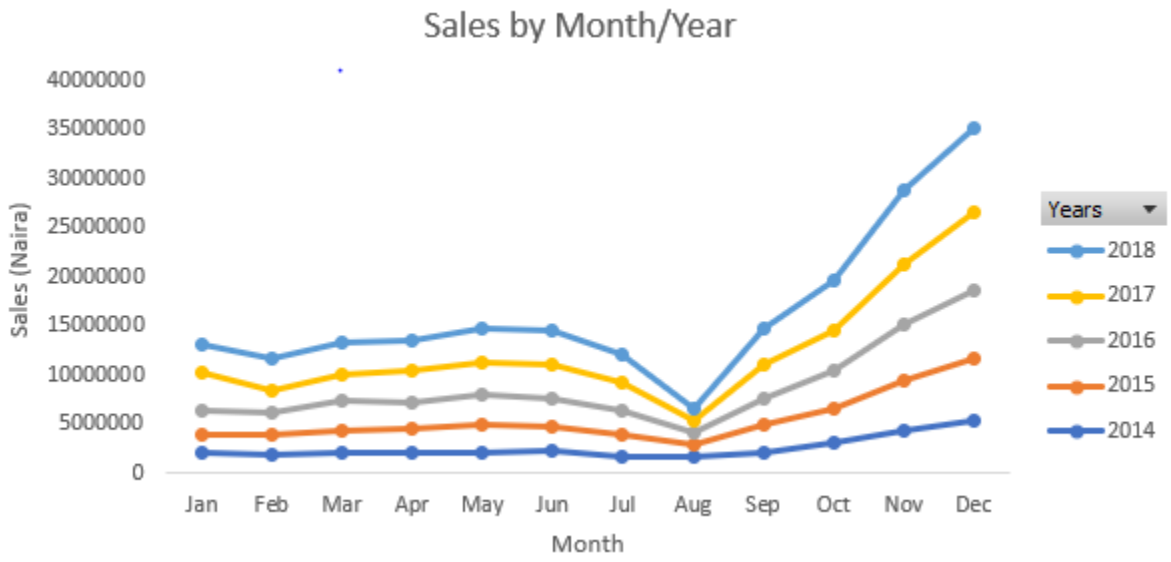
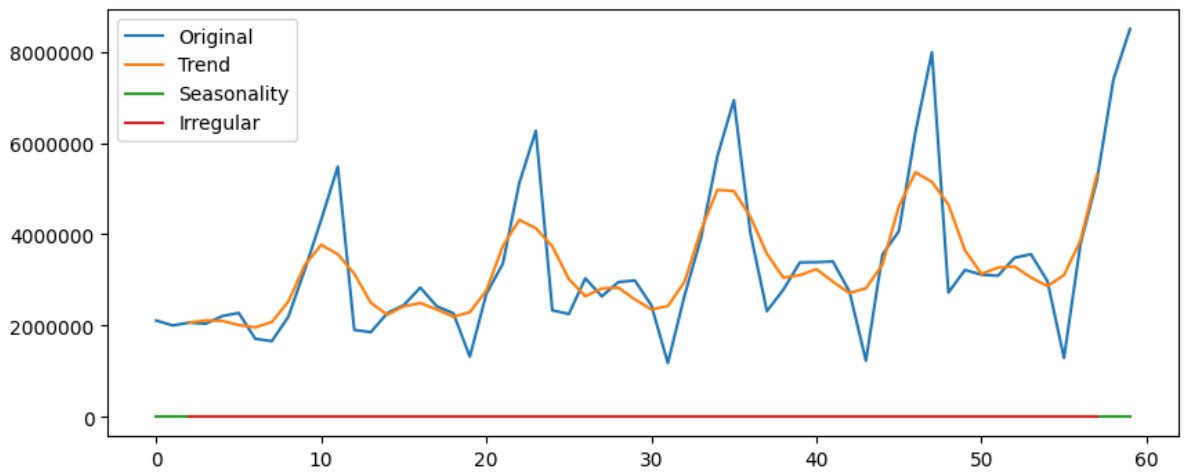


Fig. 2: Sales Seasonal pattern and trend from Jan. 2014 - Dec. 2018



The trend of the collected data shows an upward movement in sales from 2014 – 2018. However, from the Fig. I above, it can be observed that sales volumes start peaking from September and reach its highest in December. After which, decline is experienced from January. The lowest sales volume was recorded in August. This shows that shoe sales have some direct proportionality with the festive period towards December.

3.6 Model Selection

For sales forecasting, we explored several mathematical modeling techniques, including but not limited to:

- Autoregressive Integrated Moving Average (ARIMA) models.
- Seasonal decomposition of time series (STL).
- Exponential smoothing methods.

We adopted ARIMA model and STL in this project to forecast future sales and assessed the performance of these models by using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) on a hold-out validation dataset.

Autoregressive Integrated Moving Average (ARIMA) models:

Autoregression (AR) Model :

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

- y_t is the value of the time series at time *t*
- c is the mean of the time series

- $\phi_1, \phi_2, \ldots, \phi_p$ are the autoregressive coefficients, which measure the relationship between the current value of the time series and its past values
- ε_t is the error term at time *t*

ARIMA models require the data to be stationary, meaning that the statistical properties (e.g., mean, variance) remain constant over time. We first carried out a stationarity test as follow:

Test for stationarity:

The Augmented Dickey-Fuller (ADF) test results suggest that the sales data is likely non-stationary:

- ADF Statistic: 0.0588 (positive, not strongly indicative of stationarity).
- p-value: 0.9631 (high, fails to reject the null hypothesis of non-stationarity).
- Critical Values (1%, 5%, 10%): All more negative than the ADF Statistic.

In summary, the data likely has a unit root and is non-stationary, possibly exhibiting changing trends or seasonality over time. To use time series models effectively, we considered making the data stationary through differencing. This helps us analyze and model the data more accurately in the context of time series forecasting or analysis.

Differencing:

The formula for First-Order Differencing is given below

$$D(t) = X_t - X_{t-1}$$

In this formula:

- X_t represents the value of the time series at time *t*.
- X_{t-1} represents the value of the time series at the previous time *t*−1.

The calculated First order differencing is given in the table below:

Table 2: First Order differencing

Month	Sales	First Differencing
1/1/2014	2111250	NaN
2/1/2014	2004000	-107250
3/1/2014	2066250	62250
4/1/2014	2040750	-25500
5/1/2014	2209500	168750
6/1/2014	2277000	67500
7/1/2014	1711500	-565500

8/1/2014	1659000	-52500
9/1/2014	2191500	532500
10/1/2014	3225750	1034250
11/1/2014	4323000	1097250
12/1/2014	5484000	1161000
1/1/2015	1905750	-3578250
2/1/2015	1856250	-49500

First-order differencing calculates the difference between each data point and its immediate preceding data point. This helps remove trends and seasonality in the time series data, making it more stationary and suitable for various statistical analyses and modeling techniques, such as Auto-Regressive Integrated Moving Average (ARIMA) modeling.

Seasonal Differencing

The formula for Seasonal Differencing is given below

Seasonal Differencing (of order *s*):

$D(t) = X_t - X_{t-s}$

In this formula:

- $X_t$represents the value of the time series at time *t*.
- X_{t-s} represents the value of the time series at the same time of the previous season, where *s* is the number of time periods in one season.

Table 3: Seasonal differencing

Month	Sales	Sales First Difference	Seasonal First Difference
1/1/2014	2111250	NaN	NaN
2/1/2014	2004000	-107250	NaN
3/1/2014	2066250	62250	NaN
4/1/2014	2040750	-25500	NaN
5/1/2014	2209500	168750	NaN

6/1/2014	2277000	67500	NaN
7/1/2014	1711500	-565500	NaN
8/1/2014	1659000	-52500	NaN
9/1/2014	2191500	532500	NaN
10/1/2014	3225750	1034250	NaN
11/1/2014	4323000	1097250	NaN
12/1/2014	5484000	1161000	NaN
1/1/2015	1905750	-3578250	-205500
2/1/2015	1856250	-49500	-147750
3/1/2015	2273250	417000	207000
4/1/2015	2449500	176250	408750

Fig. 3: Seasonal Difference Chart

To further support this observation, the auto-correlation functions (ACFs) were calculated for each of the years. ACF is a useful analytical tool for identifying time series patterns. In this analysis, ACF calculations were conducted using a data set comprising 12 months for each year.

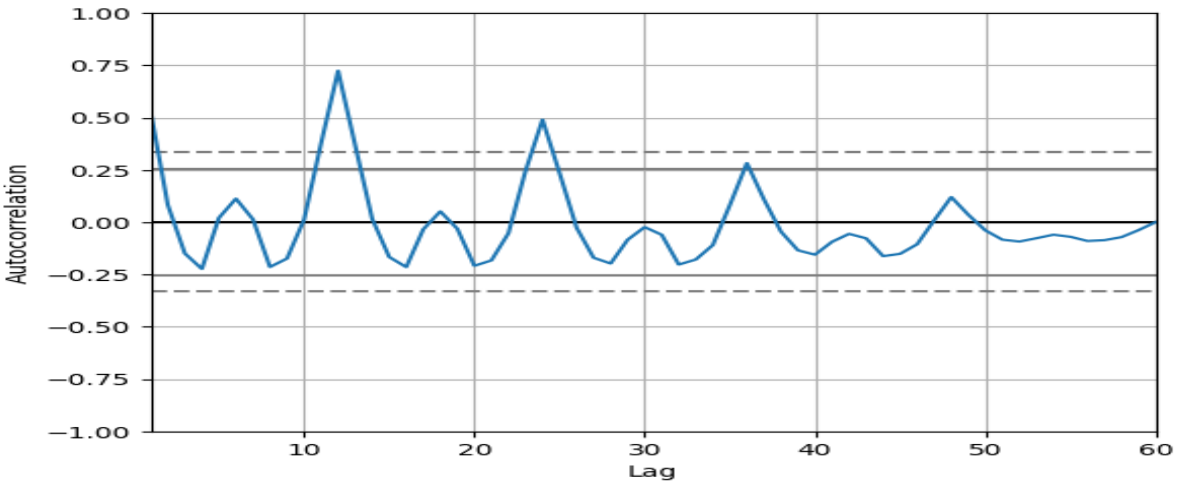
ACF for a given lag *k* is given by equation (1)

$$ACF(k) = \frac{\sum_{t=1+k}^{n} (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2} \quad (1)$$

Table 4: ACF for Lag 12

ACF for Lag 12												
Lag	1	2	3	4	5	6	7	8	9	10	11	12
ACF	1	0.526478	0.082416	-0.14962	-0.22379	0.020304	0.112125	0.013215	-0.21462	-0.17401	0.011511	0.376745

Fig. 4: ACF for Lags 60 using sales data

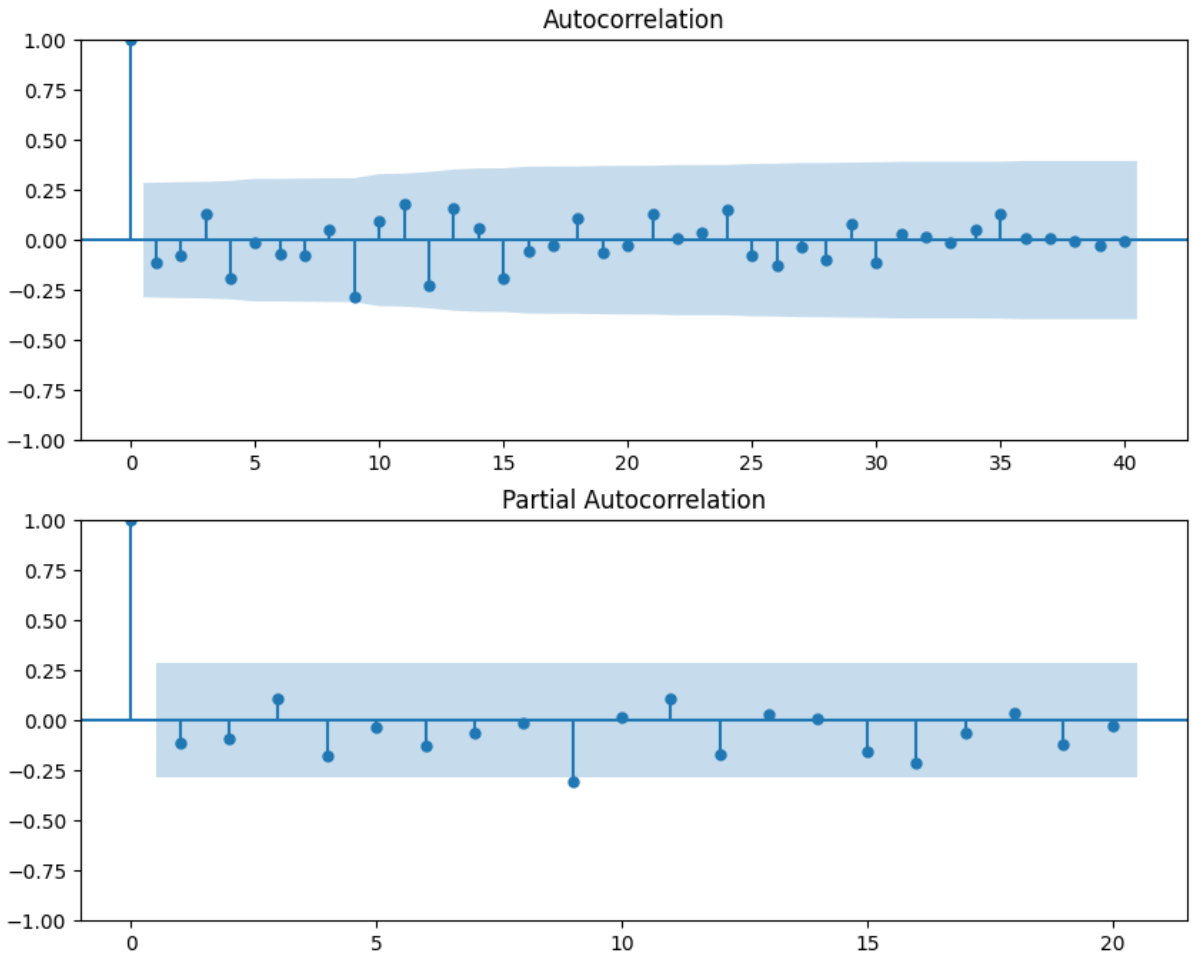


We also obtained ACF using seasonal first difference result which help us achieve a stationarity of the sales data which is suitable for ARIMA model analysis. The table below shows our **Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) plots based on the values of our seasonal difference over 40 and 20 lags respectively.**

Auto-Correlation and Partial Auto-Correlation Identification of an AR model is often best done with the PACF. For an AR model, the theoretical PACF “shuts off” past the order of the model. The phrase “shuts off” means that in theory the partial autocorrelations are equal to 0 beyond that point. Put another way, the number of non-zero partial autocorrelations give the order of the AR model. By the “order of the model” we mean the most extreme lag of x that is used as a predictor. Identification of an MA model is often best done with the ACF rather than the PACF.

For an MA model, the theoretical PACF does not shut off, but instead tapers toward 0 in some manner. A clearer pattern for an MA model is in the ACF. The ACF will have non-zero auto-correlations only at lags involved in the model. Hence, p AR model lags d differencing q MA lags.

Fig. 5: ACF and PACF for Lag 40 and 20 respectively.



For non-seasonal data: $p = 1, d = 1, q = 0\ or\ 1$, the ARIMA results using order $(1,1,1)$ is shown below:

Dep. Variable:	Sales	**No. Observations:**	60			
Model:	ARIMA(1, 1, 1)	**Log Likelihood**	-918.23			
	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
AR.L1	0.6044	0.318	1.901	0.057	-0.019	1.227
MA.L1	-0.9228	0.238	-3.874	0	-1.39	-0.456

Fig. 6: Actual sales Vs Forecast Sales

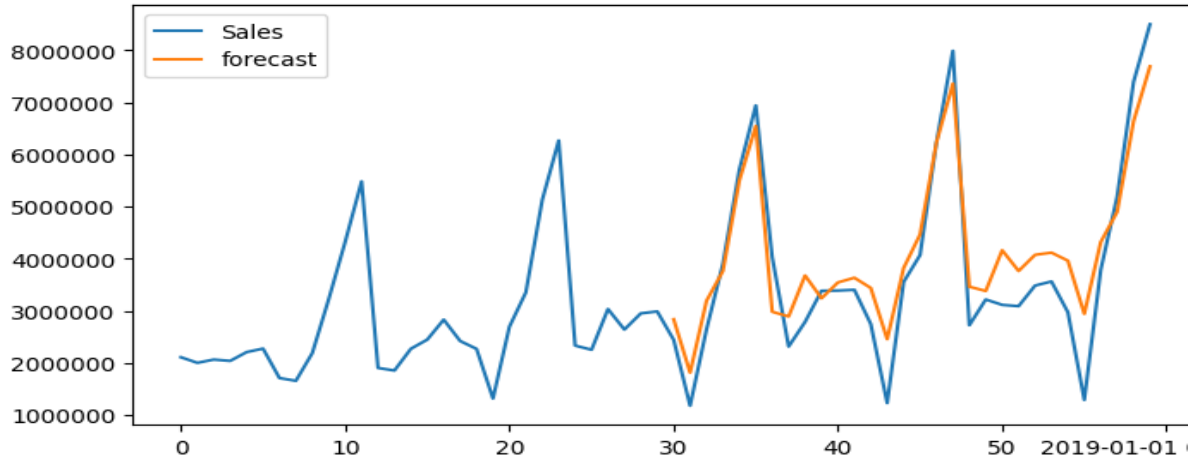
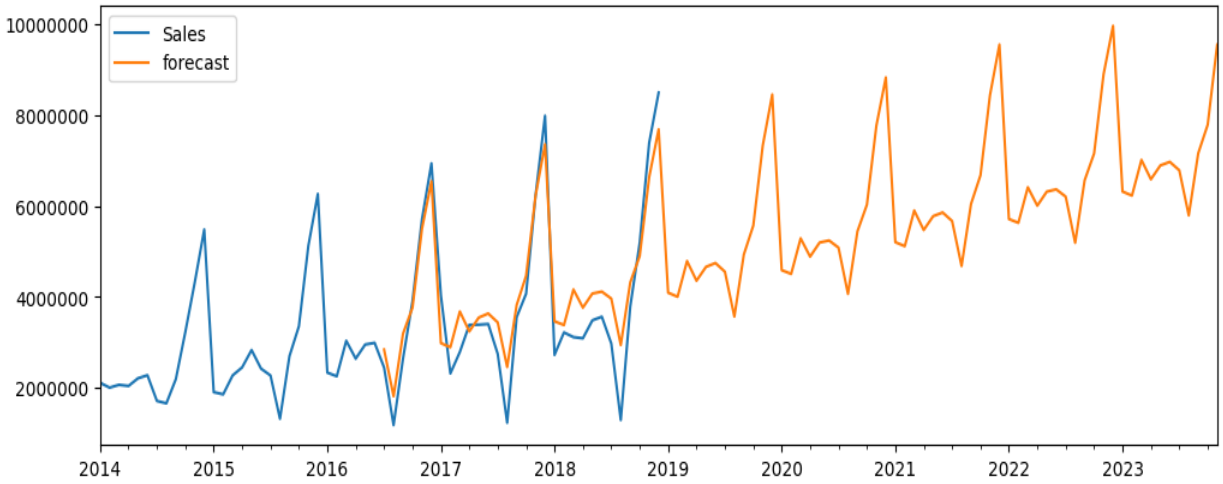


Fig.7: Sales forecast for the next 60 months



To further verify the accuracy of the ARIMA model calculated the Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE).

- MAE: Approximately 318,675
- MSE: Approximately 213,403,902,559
- RMSE: Approximately 461,911.29
- MAPE: Approximately 15.12%

3.8 Model Performance

The accuracy of a forecast can be measured using a variety of metrics, such as the RMSE or the percentage RMSE. In this case, the percentage RMSE is 15.00%. This means that the predicted values are, on average, 15.00% different from the actual values.

A 15.00% RMSE is considered to be a moderate level of accuracy. This means that the forecast is not very accurate, but it is not completely inaccurate either. The forecast could be improved by using a more sophisticated forecasting model or by collecting more data.

Ultimately, the accuracy of a forecast depends on the specific application. For some applications, a 15.00% RMSE may be acceptable, while for other applications, it may not be. It is important to consider the specific requirements of the application when evaluating the accuracy of a forecast.

4. CONCLUSION

In conclusion, our exploration of sales forecasting accuracy using historical data, STL, and an ARIMA model has provided valuable insights into the dynamics of forecasting for business optimization. The findings and recommendations from our analysis have shed light on both the strengths and areas for improvement in the current forecasting process.We observed a strong positive linear relationship (Pearson's correlation coefficient, $r \approx 0.8796$) between the actual and forecasted sales data. This indicates that the ARIMA model effectively captures the general trends in sales, which is a promising aspect of the forecasting system.

However, it's crucial to note that both the actual and forecasted sales data exhibited significant variability, as reflected in their relatively high variances. This variability suggests that there are fluctuations and deviations from the linear trend that the current forecasting model may not fully account for. Therefore, there is an opportunity for refinement and optimization.

This study represents an investment in the organization's ability to adapt and thrive in a dynamic business environment. Sales forecasting accuracy is not just a technical endeavour but a strategic one, with direct implications for competitiveness and profitability. As we embark on the journey to refine our forecasting

model and enhance its accuracy, we are committed to delivering actionable insights that drive business success. By addressing variability, incorporating additional factors, and assessing the real-world impact of forecast errors, we aim to empower our organization to make data-driven decisions and navigate the future with confidence.

ACKNOWLEDGEMENTS

The authors wish to thank Professor Sodiq SHOFOLUWE of St. Clair College Windsor, for the blends of data analytics in the numerical articulation of the results obtained in this work. We thank him for taking us through the basics of the trending concepts.

COMPETING INTERESTS

There is no competing interests.

AUTHORS' CONTRIBUTIONS

Author OYT proof read the whole manuscript prior and after submission. Author SO carried out extensive review of the literature, he also collates the facts of this study. Author AAO put all the collations in the journal form and helps in validation of some of the tools used for the analysis of the data collected, while the author SKA extracted the figures used in the interpretation of our results.

REFERENCES

1. Brown, R. G., Statistical forecasting for inventory control, McGraw-Hill, 1959; 232 pages

2. Holt, C. C., Forecasting seasonals and trends by exponentially weighted moving averages, ONR Research Memorandum, Carnegie Institute of Technology, Pittsburgh.1957; Vol. 52.

3. Box, G. E. & Jenkins, G. M., Time series analysis: Forecasting and control, Revised Edition, Holden Day, San Francisco, 1976.

4. Gardner, E. S. & McKenzie, E., Seasonal adjustment methods for a forecast survey of expenditure. Journal of the American Statistical Association, 80(391), 1985; 930-939.

5. Harvey, A. C., Time series models, Harvester-Wheatsheaf, New York, 1993. 308 pages

6. Cao, L. J., Tay, F. E. & Ewe, H. T., Neural networks for financial forecasting. Neurocomputing, 55(1-2), 2003; 307-319.

7. Zhang, G., Patuwo, B. E., & Hu, M. Y. , Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting, 14(1), 2019, 35-62.

8. Wang, J. & Wan, H., An improved hybrid model for e-commerce sales forecasting. International Journal of Information Management, 56(1), 2021, 102-280.

9. Can Ozdogar, Time Series Forecasting using SARIMA (Python), 2023, https://medium.com/@ozdogar/time-series-forecasting-using-sarima-python-8db28f1d8cfc.