BINDURA UNIVERSITY OF SCIENCE EDUCATION

FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF STATISTICS AND MATHEMATICS

A COMPARATIVE STUDY BETWEEN LSTM AND ARIMA MODELS IN FORECASTING SALES FOR SIMBISA BRANDS CHICKEN INN (2016-2023).

BY

RUMBIDZAI KIRSTEEN MTETWA

B201813B

TO BE SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR BSC.HONOURS IN STATISTICS AND FINANCIAL MATHEMATICS

SUPERVISOR: DR T.W. MAPUWEI

2024

APPROVAL FORM

The undersigned certify that they have read and recommended to the Bindura University of Science Education for acceptance of a dissertation entitled “**Comparative Study between LSTM and ARIMA in forecasting sales for Simbisa Brands Chicken Inn**”, submitted by B201813B in partial fulfillment of the requirements of the Bachelor of Science (Honours) Degree in Statistics and Financial Mathematics.

STUDENT

Name 10/06/23

Signature Date

Certified by:

DR T.W Mapuwei Map 11/06/24………
…………………

Supervisor Signature Date

Dr. M. Magodora Magodora 11/06/24……

Chairperson Signature Date

DECLARATION

I Rumbidzai Kirsteen Mtetwa as a result of this declare that this submission is my work apart from the references of other people's work which has duly been acknowledged. As a result, I declare that this work has neither been presented in whole nor in part for any degree at this university or elsewhere.

Author: Mtetwa Rumbidzai K

Registration Number: B201813B

Signature:

Date: 10 June 2024

DEDICATION

This work is dedicated to my parents in appreciation of their ongoing financial and moral support.

ACKNOWLEDGEMENTS

I want to thank the God of all power for giving me this chance and for his direction while I completed my task. I also want to express my gratitude to Dr. T.W. Mapuwei, my supervisor, for his guidance, mentorship, and useful advice that have helped my dissertation succeed. My friends, Pauline Munzwembiri, Tanyaradzwa Dzvetera, Alice Chamunorwa and Clement Simbanegavi, were also very important because of their encouragement.

ABSTRACT

As time goes on, an increasing number of restaurants search for answers to their data-related issues. The comparative analysis of LSTM and ARIMA in sales forecasting for Simbisa Brands from 2016 to 2023 was the focus of this research. The primary goal was to assess the accuracy and dependability of the sales estimates produced. During the 12-month internship at Simbisa Brands, the main aim was to increase sales and fight competition as more and more food outlets providing chicken-based products were being introduced. Using a data analytics method, two algorithms which are ARIMA and LSTM were applied, hundreds of time series data were analysed and external variables were incorporated. Python was utilized as the programming language for evaluating both models. After being put into use, SARIMAX was selected as the best model based on performance matrices (RMSE, MAPE AND MAE). Future developments and suggestions were also noted in light of the findings. It is advised that Simbisa Brands think about establishing additional stores in key areas to take advantage of the steady growth trend and grow their market share, according to the SARIMAX out prediction. It is advised that Simbisa Brands add new menu items to stay up to date with shifting consumer tastes and preferences while upholding the standards of quality and value. SARIMA outperformed both ARIMA and SARIMA models making it the best model for Simbisa Brands to use.

Keywords: ARIMA, LSTM, forecasting, time series, machine learning

Table of Contents

List of Figures

LIST OF TABLES

ABBREVIATIONS AND ACRONYMS

ACF	Autocorrelation Function
ADF	Augmented Dickey-Fuller
AIC	Akaike Information Criteria
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
CZI	Confederation of Zimbabwe Industries
I	Integrated
KPSS	Kwiatkowski-Phillips-Schmidt-Shin
LSTM	Long Short-Term Memory
MA	Moving Average
MAPE	Mean Absolute Percentage Error
MASE	Mean Absolute Scaled Error
MSE	Mean Squared Error
PACF	Partial Autocorrelation Function
PACF	Partial Auto Correlation Function
RMSE	Root Mean Square Error
ZIMSTAT	Zimbabwe National Statistics

CHAPTER 1: INTRODUCTION

1.0 Introduction

Sales forecasting helps businesses predict and plan for future demand, which makes it an essential part of business decision-making. Businesses can increase profitability and competitiveness by optimizing inventory management, production planning, and resource allocation through the analysis of historical sales patterns. Given that sales can be tracked over time, such as daily, one method of forecasting future sales is to examine historical sales and the patterns they have formed over time, then apply those patterns to future forecasts. This type of problem formulation is not exclusive to sales forecasting. For instance, weather forecasting (Taylor, McSharry, Buizza, 2009) and stock market prediction (Pai and Li, 2005) Utilize historical data to forecast future behavior. This is referred to as a time-series forecasting problem, and a great deal of study has been done in this area. As a result, it is logical to approach the sales prediction problem as a problem involving time series forecasting. (Yu, Wang, and Strandhagen, 2018). Several models are employed in the forecasting field, including ANN and ARIMA. However, ANN techniques have received a lot of attention because of their performance and accuracy in predictions, particularly in the past year with the growth of AI research (Krstanovic and Paulheim, 2017). According to Yu et al (2018), the LSTM model has grown in popularity since it can recall data from a very far-off point in the time series. Thus, this research will use the LSTM and the ARIMA model to address the Simbisa Brands sales forecast issue.

1.1 Background of Study

Trends in food consumption have significantly changed throughout Southern Africa, including Zimbabwe. These days, people strongly prefer to eat meals that are not made at home but consume food at restaurants. Future forecasts indicate that this trend is expected to continue increasing. The rise in several people who like to eat outside of their homes has caused the number of businesses in the food industry to grow quickly. The Hospitality Association of Zimbabwe Congress minutes (2005) say that the Fast-Food area has viewed a good-sized boom averaging 6 percent every annum since the early 1990s, although the statistic is only considered the registered Fast-Food outlets. This has resulted in stiff competition with each player trying to dominate the industry. The Confederation of Zimbabwe Industries, (CZI), is an organization in Zimbabwe whose responsibility is to strengthen and monitor business activities. The organization's annual survey for the year 2015, confirmed that the current trends of the growing food industry were taking place against a background where most industries in Zimbabwe were shutting down for the food industry (CZI, 2016). The Zimbabwe National Statistics Agency (ZIMSTAT) released data in 2015 that supported the CZI stance in a similar study. According to ZIMSTAT (2015), 26 food establishments-including grocery stores and formal and informal restaurants open nationwide every three months on average. The Indigenous Food Processors Association, an impartial group made up of native Zimbabwean food entrepreneurs, also provided statistics in 2014 showing that, on average, 70 new food outlets opened each month across the country. These figures make it clear that, as long as businesses continue to flourish, the fast-food industry will present enormous prospects for entrepreneurs.

However, for Simbisa brands to thrive, there is a need to identify underlying trends and patterns, verify the presence of seasonality effects, develop an accurate time series forecasting model for sales prediction, and offer practical insights and suggestions to enhance sales performance, giving them a competitive edge in the fast-food sector. These include fast-food retail stores, restaurants, and mobile wagons and caravans.

1.2 Statement of The Problem

Simbisa Brands Chicken Inn faces fierce competition as more businesses continue to enter the fast-food sector, endangering its market share and profitability. Given that so many establishments now sell chicken-based products, it is clear that the food sector is highly competitive and that there is less of a market for chicken inn products. Analysing and forecasting sales, which translate into demand for Chicken Inn items, is necessary to combat competition. This study intends to look into several time-series forecasting algorithms that Simbisa Brands uses to predict sales. According to earlier studies, machine learning can produce reliable and successful models for predicting product sales, which will save retailers' costs (Fildes and Kolassa, 2022). While there are several approaches to forecasting prediction, the classic ARIMA model and LSTM networks will be the main emphasis of this study. Given that ARIMA is a linear model, it is anticipated that the LSTM model will predict outcomes more accurately by accounting for the non-linearity of sales forecasting. Even though ARIMA is widely recognized for its predictive accuracy, it is amusing to compare the models. Simbisa Brands may be able to outbid rivals if they select the more accurate model, which provides a more realistic sales projection. The purpose of this study is to determine which model performs better than the other.

1.3 Research Objectives

1. To study the historical sales data of Simbisa brands.
2. To develop appropriate time series models to forecast future sales of Simbisa brands.
3. To evaluate how reliable and accurate the sales projections produced by the time series models are.
4. Perform a comparative analysis of ARIMA and LSTM

1.4 Research Questions

1. What patterns and trends can be found in Simbisa Brands' historical sales data?
2. What time series model provides the most accurate and reliable forecast for the future sales of Simbisa Brands?
3. How can the forecasting models be improved to enhance the accuracy of sales predictions?
4. How do the forecasting accuracy and performance of LSTM and ARIMA models compare the time series prediction tasks?

1.5 Scope of The Study

The research project is confined to contrasting LSTM with ARIMA, a commonly used model because there are other cutting-edge techniques for time-series forecasting. Sales forecasting at Simbisa Brands Chicken Inn will be the domain of comparison for the models. Emphasis was on the last eight years period from 2016 to 2023.

1.6 Significance of Study

Practical Contribution to Simbisa Brands: The study offers practical insights and suggestions to maximize operational effectiveness and sales forecasting accuracy, particularly for Chicken Inn. Simbisa Brands may make educated choices on marketing tactics, resource allocation, and inventory control by understanding sales trends and how they are incorporated into forecasting models. This may result in increased profitability, better operational efficiency, and a competitive edge for the Chicken Inn brand.

Comparing LSTM and ARIMA models can offer insightful information about the efficacy of traditional statistical methods (ARIMA) versus advanced deep learning techniques (LSTM) in handling sales forecasting tasks. This comparison can help identify which approach better suits the specific characteristics of Simbisa's sales data, such as seasonality, trends, and irregular patterns.

Improved Decision-Making and Strategic Planning: The insights derived from the study can assist Simbisa Brands in making informed decisions and formulating effective strategies for Chicken Inn. By understanding the factors that drive sales, to meet consumer demand, the business can optimize its marketing initiatives, menu selections, and employee numbers. This can result in improved customer satisfaction, increased sales revenues, and sustained growth for Chicken Inn.

1.7 Assumptions of The Study

- The research is conducted under the assumption that the sales data utilized for analysis is trustworthy and accurate. It is assumed that the information was gathered from the sales records of Chicken Inn and has been properly recorded and documented without significant errors or inconsistencies.
- All sales, according to the study, were quantified in US dollars.

1.8 Limitations of The Study

Due to the numerous protocol channels and lack of an authorization document, the researcher had difficulties when collecting information.

1.9 Definition of terms

1.9.1 Time series forecasting

Time series forecasting is the process of projecting future values of a variable using historical time-ordered data points. It is represented as a collection of vectors with the formula x(t)= 0,1,2,4..., where the amount of time elapsed is specified as t (Adhikari, 2014). Patterns, trends, and behaviors that change over time are analysed and understood through the implementation of historical sequences. To forecast future values, this method entails analyzing and modeling the patterns, trends, and other features found in the historical data. Future values are forecasted using historical data (Mills, 2019). Future product sales can be predicted by utilizing time-series forecasting algorithms.

1.9.2 Sales forecasting

The technique of projecting future sales to maximize revenue is known as sales forecasting. It entails forecasting a product or service's future sales using previous data market trends, and other relevant factors. A sales forecast is an estimate of future sales that is derived from market research, historical data, and other pertinent variables (Fan, Che, and Chen, 2017). It assists companies in forecasting demand and allocating resources appropriately.

1.9.3 Fast Food

Fast food is described as food that can be prepared and served quickly, typically from a restaurant or food truck, and is often prepackaged and mass-produced (Parasecoli, 2019). This type of food is known for its convenience and speed of service, often appealing to individuals looking for a quick meal option.

1.10 Chapter Summary

This included an overview of the research, a statement of the problem, objectives, research questions, assumptions, study significance, study boundaries, and definitions of key terms. This chapter is crucial since it provides the overall study's direction. The literature overview of previous studies on comparative study between LSTM and ARIMA will be presented in Chapter 2. Chapter 3 explains the applicable methodology. Chapter 4 will analyze and present the findings. Lastly, chapter 5 will cover the research's findings and recommendations. A thorough summary of the most recent theoretical and empirical research on time series analysis and sales will be given in the following section.

CHAPTER 2: LITERATURE REVIEW

2.0 Introduction

It was discovered that time series analysis was an effective method for predicting sales trends, offering businesses insights into their past performance and future trends. This literature review concentrates on using time series analysis to examine the sales data of Simbisa Brands Chicken Inn, a well-known fast-food chain in Zimbabwe. By examining previous studies and methodologies employed in similar analyses, the aim is to shed light on different time series models in understanding and predicting sales fluctuations in Simbisa Brands. Time-series forecasting is done with a variety of models, each applied in a distinct context and area. Adhikari (2015), cites ARIMA, ANN, and Support Vector Machines as a few examples. Models fall into one of two categories which are linear or non-linear. Non-linear models can more nearly match the shape of a more complicated function than linear models, whose presumption of linear behavior places limitations on them.

2.1 Theoretical Literature

This literature review focuses on examining and analysing the theoretical frameworks, concepts, and models related to this study. It provides a framework for understanding the research topic and guiding future research

2.1.1 Time Series Analysis

According to Box et al. (2015), a sequence of observations taken one after the other over a given amount of time is called a time series. Determining the fundamental reasons for variations in sales, including trend, seasonality, and other cyclical or irregular elements, is helpful. Time series data are analysed and modeled using many statistical methods like ARIMA, seasonal ARIMA or state space models, and exponential smoothing approaches. Understanding the process of generating the data, giving a succinct explanation of it, and modeling the mechanism that generates it are the main objectives of time series analysis. A time series model is the joint distribution of a set of random variables $\{X_t\}$, given each seen data point $\{x_t\}$, where xt is considered to be an occurrence. Figure 1 below shows the steps involved in time series analysis.

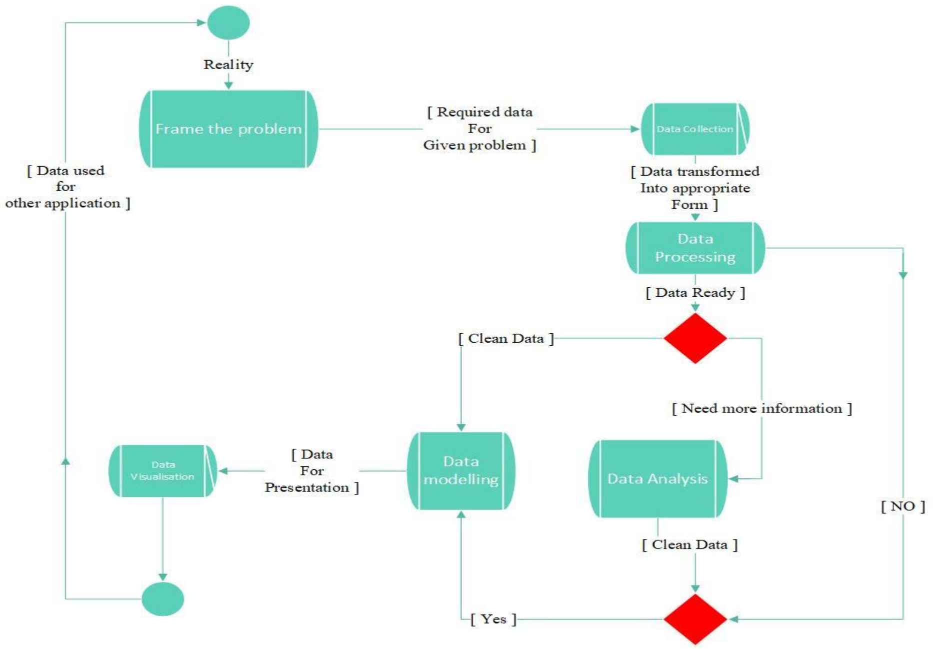


Figure 2. 1 Data analysis process

Gathering data over time is the first stage. Some of the data gathered is inaccurate, lacking, or repeated and can have some errors. Therefore, data preparation is done by managing missing values and correcting of errors in the following step. To gain a deeper knowledge of the data, exploratory data analysis is done after the data is ready, (Bevalig, 2024).

2.2 Time series and its components

A sequence or collection of observations x_0, x_1, x_2, · · · generated successively throughout time is called a time series. The equation provides it:

$$\text{Let } y(t) = x(t)\beta + \varepsilon(t), \ldots\ldots\ldots\ldots\ldots \ (2.1)$$

for every t = 0, ±1, ±2, $y(t) = y_t$. The sequence · · · has the time subscript as an index t. It consists of an unobservable white-noise sequence $\varepsilon(t) = \varepsilon_t$ of randomly distributed random variables combined with an observable signal sequence $x(t) = x_t$.

When information on one or more variables can be tracked, collected, and examined over an extended period, we can discuss time series data (Zhang, Lin, and Li, 2015). Data in time series might be random, cyclical, or seasonal. T can be used to represent the number of months, quarters, or years because the data is classifiable on a monthly, quarterly, or annual basis. Time series data shows unequivocally that what we saw this year regarding Chicken Inn sales is impacted by what we saw last year or years prior, even though it cannot be analyzed from the standpoint of the dimension of the data acquired through random sampling.

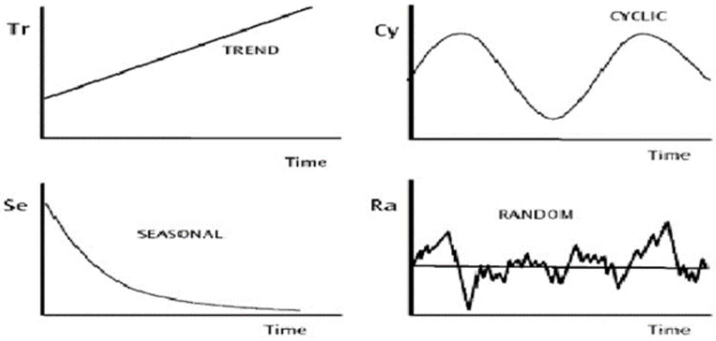


Figure 2. 2 time series components

2.2.1 Trend components

In time series analysis, the trend component describes the continuous movement or direction of the data across time. It depicts the underlying pattern that indicates if the data is rising, falling, or remaining unchanged over an extended length of time. The trend element is critical for forecasting and pattern recognition, as well as for comprehending the general behavior of the time series data. The trend component is crucial since it aids in forecasting, long-term pattern identification, and comprehending the general direction of the data.

2.2.2 Seasonality

Seasonality pertains to a component of data in a time series wherein the observations display consistent and foreseeable alterations that transpire at particular intervals, frequently repeating annually. These regular changes tend to repeat over a predetermined length of time and can be impacted by a variety of circumstances, including weather, holidays, and cultural events. Seasonality affects both the forecast's accuracy and the way the data behaves.

2.2.3 Cyclical Components

The movements in such a time series are irregular and take the form of a circle or oscillating movements over a length of time greater than a year. The ups and downs associated with cyclical components, which are not periodic but duplicate after some time typically result in global booms and busts of numerous economic systems.

2.2.4 Random Components

A random component time series is defined as data that does not reflect any form of trend or pattern, such as a seasonal or cyclical trend. This means that there are no discernible tendencies, and they are therefore unpredictable in general.

Sales

Any organization must prioritize sales since they have a direct bearing on earnings and profitability. According to the research by Suardika and Dewi (2021), sales is the exchange of products or services for money or its equivalent. Sales are a major factor in the expansion and sustainability of businesses. Organizations can enhance market share, foster customer connections, and maintain their competitiveness in the market by using efficient sales methods. To produce income for the business, salespeople are in charge of finding prospects, qualifying leads, fostering connections, and closing deals.

2.3 Autoregressive model (AR)

The autoregressive model (AR) is a particular kind of regression model in which the dependent variable is based on past values. This implies that the forecast and previously observed values have to be related. The foundation of this method is the concept of partial autocorrelation. The following equation represents an AR model:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \omega_t \ldots\ldots\ldots\ldots\ldots (2.2)$$

Just like before, ω_t represents an unobservable white-noise sequence of randomly dispersed variables that are equally and independently distributed.

2.4 The Moving Average (MA)

It is devoid of seasonal components that have a certain amount of randomness (Pannerselvam, 2005). Moving average models offer the advantage of anticipating stocks or items with a steady demand, when there is a little trend or season, and are also useful in identifying areas of support and resistance. On the other hand, MA is unresponsive to variables that occur for a cause, such as seasonal influences and cycles. The following is the formula for calculating the Moving Average:

$$y_t = c + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q} \ldots\ldots\ldots\ldots\ldots\ldots (2.3)$$

ε_t represents white noise and this is the MA(q) model (Ruey, 2010).

2.5 Autoregressive Integrated Moving Average (ARIMA)

The ARIMA model is a popular statistical analysis model for time series data. To estimate future values and capture the complex dynamics of time series data, it integrates the ideas of autoregression, differencing, and moving averages. It is made up of three components which are moving average (MA), integrated (I), and autoregressive (AR). The ARIMA model is used to forecast values for the future using historical data, identify patterns, and analyse time series data. The following is the ARIMA equation:

$$y't = \mu + \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} - \theta_1 e_{t-1} - \ldots - \theta_q e_{t-q} \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.4)$$

Unit-root nonstationary nature is attributed to the ARIMA model (Lin, and Han, 2018). ARIMA model is important in that it uses only the data from the relevant time series. First off, this capability comes in handy when forecasting a lot of time series. Secondly, this circumvents an issue that occasionally arises with multivariate models. ARIMA, however, may only be able to predict extremely high or low values. Outliers are challenging to predict for ARIMA since they fall outside of the overall trend that the model captures, even though the model is good at modeling seasons and trends.

2.6 Seasonal Autoregressive Integrated Moving Average (SARIMA)

The SARIMA model is a time series forecasting method that extends ARIMA by including seasonal components. It is a powerful tool for modeling and predicting time series data. To implement a SARIMA model, you would typically identify the AR and MA as well as the seasonal components, and fit the model to the data to make forecasts. The equation of SARIMA is shown below

$$(1-\phi_1 B)(1-\Phi_1 B^4)(1-B)(1-B^4)\, y_t = (1+\theta_1 B)(1+\theta\ _1 B^4)\mathcal{E}_t, \ldots \ldots (2.5)$$

The Backshift operator is represented by B.

2.7 Artificial Neural Networks (ANN)

ANNs attempt to imitate neural networks and their behavior, drawing inspiration from the neural connections found in the human brain. Because they are resilient and self-adjusting, artificial neural networks (ANNs) can typically solve complex non-linear problems that are difficult to implement implicitly. These issues include speech recognition, natural language processing, and forecasting (Peng, Yao, You, and Chi, 2017). Between input and output, the network contains what are known as layers, and the more levels there are, the more complicated the network is. The network is trained by weight selection concerning the expected output of the input by feeding it known data. To reduce the error between the output that the network concluded and the actual desired output, the network employs the gradient of error concerning the weights to change the scalar weights. Sequential data, however, is not suited for these kinds of networks. Analysis of the sequential data is hampered by the network's inability to represent dependencies due to its lack of memory for prior time steps. It is therefore desirable to have some kind of memory.

2.7.1 Recurrent Neural Networks (RNN)

Recurrent neural networks, or RNNs, are a series of interconnected layers of networks that carry over data from past time steps to subsequent ones along with the output. The output from earlier layers is taken into consideration when each layer's settings and input are processed., providing the network with a kind of memory (Rojas, 2013).

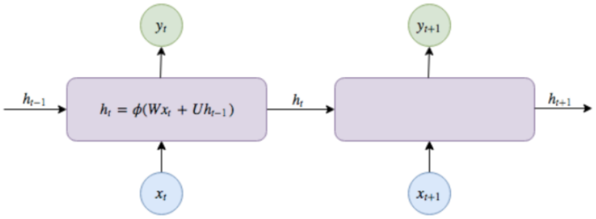


Figure 2. 3 RNN structure

Figure 2.3 demonstrates an RNN with a non-linear weight matrix (W), an activation function (φ), and a matrix (U) made up of so-called biases.

The gradient can get very complex because RNNs require the incorporation of a time dimension. This leads to the information from previous time steps either disappearing or dramatically improving. These phenomena are referred to as exploding and disappearing gradients, respectively. These observations imply that if data is kept far back in the sequence, the network can have problems correctly retrieving it from previous time steps (Li, Peng, Yao, Cui, You, and Chi, 2017). The answer to these issues is LSTM.

2.7.2 Long Short-Term Memory (LSTM)

The LSTM networks use gated cells, that are similar to computer memory, to store data. LSTM cells, in contrast to the previously stated networks, also determine when to allow data from previous time steps to be read and written (Kumar, Goomer, and Sing 2018). Consequently, the LSTM model solves the disappearing or exploding gradient problem (Krstanovic and Paulheim, 2017). The internal parts of the LSTM cell are shown in the figure below. A single cell processes data for a single time step and forwards chosen data to be sent to the next cell at

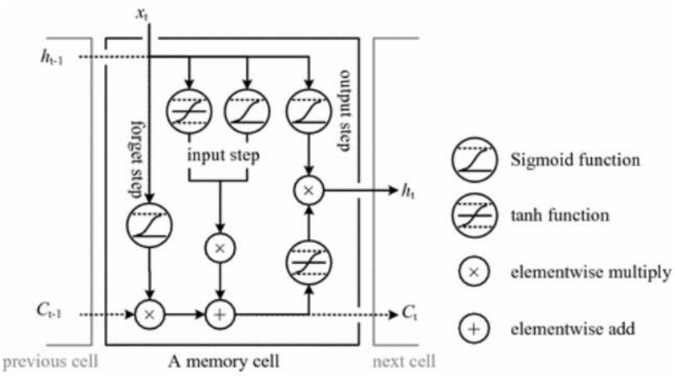


Figure 2. 4 LSTM Cell and its components (Wang and Lou, 2018)

Input at time-step t is represented by X_t.

H_t is the result of a single time step.

C_t adds to the standard output any additional dependencies that are recalled from earlier time steps. (Yu, Strandhagen, and Wang, 2018).

Depending on the gates, the next time step demonstrates an example where the fourth time step is predicted using the previous three. In contrast to the previous diagram, which just employs h_t, take note of the extra output C_t.

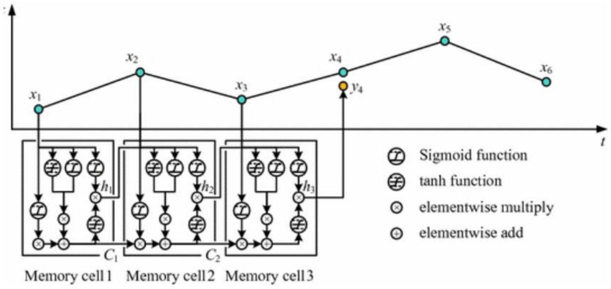


Figure 2. 5 time series forecasting using LSTM

Figure 2.5 illustrates what time-series forecasting using LSTM-cells could look like.

2.8 Empirical Literature Review

According to Yu et al. (2018), convolutional neural networks based on deep learning can be used to approximate the pricing of used homes in Beijing (Sugiartawan, Pulungan, & Sari, 2017). To forecast the trend of a time series, Lin et al. (2017) introduced TreNet, a brand-new hybrid neural network, which was motivated by the recent achievements of artificial neural networks. In their investigation, they analysed three different data sets: electricity use, chemical sensor records exposed to dynamic gas mixtures at different concentrations, and daily stock transaction data from Yahoo Finance and the New York Stock Exchange. There was a comparison between the outcomes from the CNN, LSTM, and ARIMA models. To compare the three models in use, they employed a logical regression model. They concluded that other approaches were inferior to the LSTM's prediction accuracy, which accounts for the time series. (Jiao, Xin, Wang, & Wang, 2018).

To decrease food waste, they performed a comparison study between LSTM and ARIMA to predict food product sales at grocery shops. Two scenarios have been used in the experiment. Using the provided data, both approaches have been applied, in one scenario to predict one day into the future and in the other to anticipate seven days into the future. The findings indicate that LSTM performed similarly to ARIMA when predictions were made just one day in advance of the future and only surpassed ARIMA in one case where forecasts were made seven days ahead of time. The models' effectiveness was assessed using the RMSE and MAE.

Additionally, Barddal et al. (2019) examined the use of LSTM and ARIMA models for retail industry sales forecasting. The study concluded that LSTM models were more suitable for capturing the non-linearities and complex relationships present in sales data, leading to better forecasting performance compared to ARIMA models.

Another study by Kourentzes and Petropoulos (2019) compared various forecasting methods, including LSTM and ARIMA, in the context of demand forecasting. The findings demonstrated that LSTM offered more precise sales projections compared to traditional statistical models like ARIMA, especially when handling large amounts of data with complex patterns.

Moreover, Gao et al. (2020) conducted a study in the fast-food sector but focused specifically on leveraging machine learning algorithms for sales forecasting. They compared the performance of traditional time series models with sophisticated machine learning methods, like neural networks. Their results indicated that machine learning models outperformed traditional methods in capturing complex sales patterns and achieving more accurate forecasts. This research shed light on the potential of incorporating innovative approaches in time series analysis for sales prediction.

2.9 Research gap

When comparing LSTM and ARIMA for sales forecasting, there is a research gap in the assessment of model interpretability. LSTM models are renowned for their capacity to identify long-term relationships and irregular trends in data, but because of their opaque structure, it can be difficult to comprehend the variables that influence their predictions. However, ARIMA models offer greater clarity when it comes to comprehending how historical values affect projected future values.

2.10 Proposed Conceptual Model

This study's main goal is to evaluate the precision of sales forecasting using the LSTM and ARIMA models. Sales is the dependent variable, while the LSTM and ARIMA models are the independent variables. The models were trained and tested using time series data, or historical sales data.

2.11 Summary

The chapter reviewed the empirical and theoretical literature on earlier research on time series analysis of sales. The data utilized in this chapter served as the foundation for analyzing and assessing Simbisa Brands' time series analysis. The next chapter will detail the study's methodology in addition to the methods for collecting and interpreting data.

CHAPTER 3: RESEARCH METHODOLOGY

3.0 Introduction

This section concentrates on the methods implemented to meet specified study objectives. The methodology is not limited to research methods, but can also be used as a basis for strategies used in the discourse of this study and to explain why the researcher used some techniques and not another, thus making the researcher the only person who can evaluate the results of this study (Kothari, 2017). The research methodology gives a framework and tools used for data collection and the plan for data analysis (Dawson, 2014). The time series model is to be injected to unveil and predict future sales of Simbisa Brands Chicken Inn.

3.1 Research Design

A study plan, by Parahoo (2015), is a strategy that outlines the how, when, and where of gathering data or conducting an analysis of the study hypothesis. Time series analysis of sales was investigated using a quantitative study strategy to guarantee the research's understandability to the intended audiences and to show how applicable it is to different economic systems' real-world situations. A quantitative research methodology is used in this study's research design. According to Lisa (2018), a systematic study of phenomena using computer, mathematical, or statistical methods and the collection of measurable data is known as quantitative research. The degree of influence from every element under examination was shown to be best executed using a quantitative research approach, ensuring that the problem under investigation is thoroughly researched and defined. The goal of the quantitative analysis approach is to improve comprehension and potentially allow for a more perceptive and better interpretation of the findings from the quantitative data.

3.2 Data Sources

Data gathered for various purposes with another person on behalf of a user is secondary. The research applied the data from Simbisa Brands Chicken Inn from the period of 2016 to 2023. All variables in this study are monthly. Time series data was selected because it captures variability in time rather than heterogeneity between variables. The time series can provide insight into what was happening between the variables of concern over the period under study.

3.3 Study population and sampling procedures

According to Asiamah, Mensah, and Oteng-Abayie (2017), the term "target population" refers to the group of participants in a study who are specifically chosen for the research project and are typically identified by one or more distinguishing characteristics. Thus, clients who deal with Simbisa Brands are included in the study's target population. The Harare chicken inn stores were included in the study. The study made use of secondary data, specifically sales ($) on Simbisa brands Chicken Inn. The eight-year period covered by the secondary data is from 2016 January to 2023 December. The information was given monthly. The moment an analyst makes use of recently collected information by others is known as secondary data. Data analysis is the act of looking through, cleaning, transforming, and modeling data to identify pertinent information and help with making decisions. The Gaap (Simbisa Brands website) provided the data used in this investigation.

3.4 Research Instruments

A research instrument is a device that you can use to gather, quantify, and examine information on a particular area of interest. To enter and view data and to receive the results of the various models, the researcher used the Microsoft Excel package. The code was created in the Python programming language utilizing various packages, including pandas, keras, and tensorflow.

3.5 Data Collection

Information that has already been gathered for a particular reason at a particular moment in time is referred to as secondary data. Secondary data was employed in the study since it was simple to collect and evaluate. Secondary data in this study was confirmed as it can be biased. The data was gathered across eight years. The information was stored on an Excel spreadsheet on the computer.

3.6 Description of variables

Table 3. 1 Description of variables

Symbol	Variable	Explanation
Y	Sales	This is the dependent variable to analyze and forecast. It represents the revenue generated by Simbisa brands.

3.7 Data Analysis

It is known as the methodical application of statistical tools to characterize the structure of data and derive specific conclusions about the variables under investigation. To conclude the study, this involves transforming the data with the main goal of finding certain correlations between variables.

3.8 Diagnostic test

3.8.1 Serial Autocorrelation

Data that shows the degree of similarity between the same variable's values over a series of time intervals is known as autocorrelation. One method used to check for serial correlation was the Box-Ljung test. It is a statistical analysis used to ascertain if the observed values in a time series show serial correlation or are independently distributed.

3.8.2 Test for Normality

The data in any statistical study or survey should conform to a normal distribution. The Q-Q plot and the histogram of the residuals were utilized to check for normalcy. A graphical tool called the Q-Q plot is used to determine whether a given collection of data is representative of a certain distribution, like the normal distribution.

3.8.3 Testing for stationarity: Augmented Dickey-Fuller

The models chosen to represent a series are chosen by testing against raw data. The series required to accomplish this includes autocorrelation functions and partial autocorrelation functions. If the raw data does not cease, data differencing can be used, and ADF can be used to confirm stationarity. The more negative the hypothesis is, the more the notion that there exists a unit root, at whatever confidence level, is rejected.

3.8.4 Kwiatkowski–Phillips–Schmidt–Shin (KPSS)

A statistical test called the KPSS test can be used to ascertain whether or not a time series is stationary. It is said to be stationary around the deterministic trend according to the null hypothesis of the KPSS test. It may not be stationary, which is the alternate theory.

3.9 Analytical model

The research aims to investigate the sales forecasting performance of ARIMA and LSTM models. By employing historical sales data, these analytical models will be used to predict future sales figures, which will aid in the decision-making process related to inventory management.

3.9.1 ARIMA Model

ARIMA models are very useful for precisely forecasting seasonal time series and capturing their behavior. The ARIMA-based forecasts should serve as a minimum standard if competing models or approaches are to be employed for forecasting comparison. It provides insights into historical patterns and enables informed decision-making based on future projections. We must devise a plan to identify and include unique events in the forecasts in addition to normal recurring trends. These exceptional occasions could be recognized as holidays or festivals, athletic events, other planned neighborhood gatherings, etc. Typically, the model is called ARIMA (p,d,q), in which

-p is the order of AR terms

-d is the degree of differencing

-q is the order of MA terms

The ARIMA model can be represented as:

$$y't = \mu + \varphi_1 y_{t-1} + ... + \varphi_p y_{t-p} - \theta_1 e_{t-1} - ... - \theta_q e_{t-q}$$

Whereby

$y't$ represents the value of the time series at time t.

C - constant (intercept)

Ɛ - error term at time t.

Ø - parameters of autoregressive components.

ϕ -parameters of the moving average component, representing the correlation between the error terms at different lags.

Model Building Using ARIMA

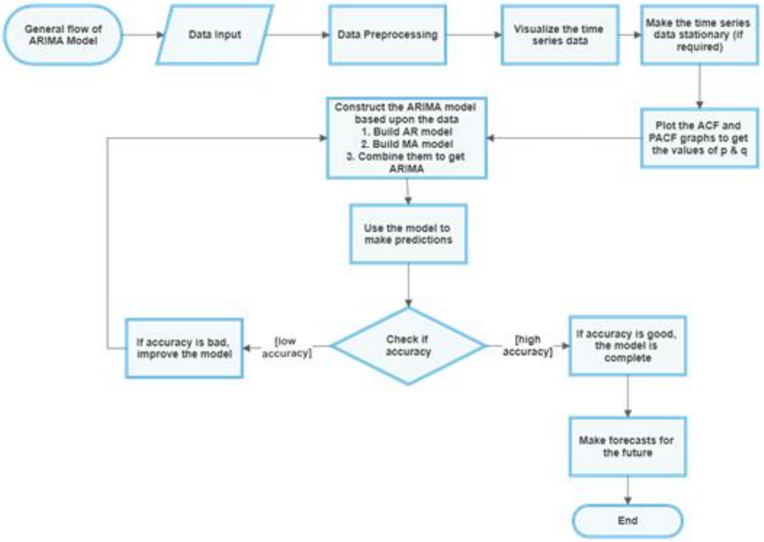

General flow of ARIMA Model
Data Input
Data Preprocessing
Visualize the time series data
Make the time series data stationary (if required)
Plot the ACF and PACF graphs to get the values of p & q
Construct the ARIMA model based upon the data
1. Build AR model
2. Build MA model
3. Combine them to get ARIMA
Use the model to make predictions
Check if accuracy
[low accuracy]
If accuracy is bad, improve the model
[high accuracy]
If accuracy is good, the model is complete
Make forecasts for the future
End

Figure 3. 1 ARIMA model building

3.9.2 Long Short-Term Memory (LSTM)

A specific kind of RNN called an LSTM is made to identify long-term dependencies. Its hidden layer has a complicated structure known as the LSTM unit. One kind of RNN architecture called LSTM was created to overcome the shortcomings of conventional RNNs in terms of identifying and gaining knowledge from long-range dependencies in sequential data. Figure 3.2 shows a basic illustration of this structure. They are extensively employed nowadays because they perform exceptionally effectively on a wide range of issues.

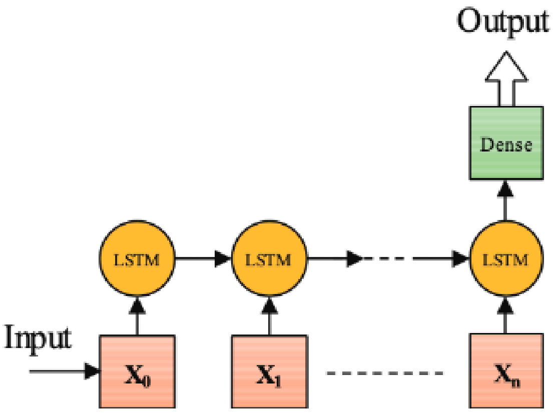


Figure 3. 2 LSTM Basic structure

An LSTM structure consists of an RNN cell and a memory. This memory allows for the retrieval and transmission of data from one moment to the next. The information that the model uses for training is determined by it. In actuality, these networks do not actively seek to learn how to remember information for extended periods; rather, this is their natural habit. In Figure 3.4, an LSTM unit is displayed.

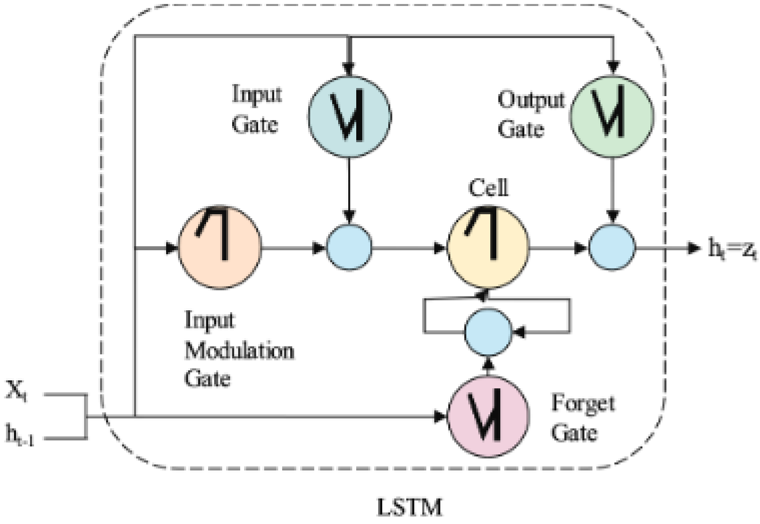


Figure 3. 3 LSTM unit

Model building using LSTM

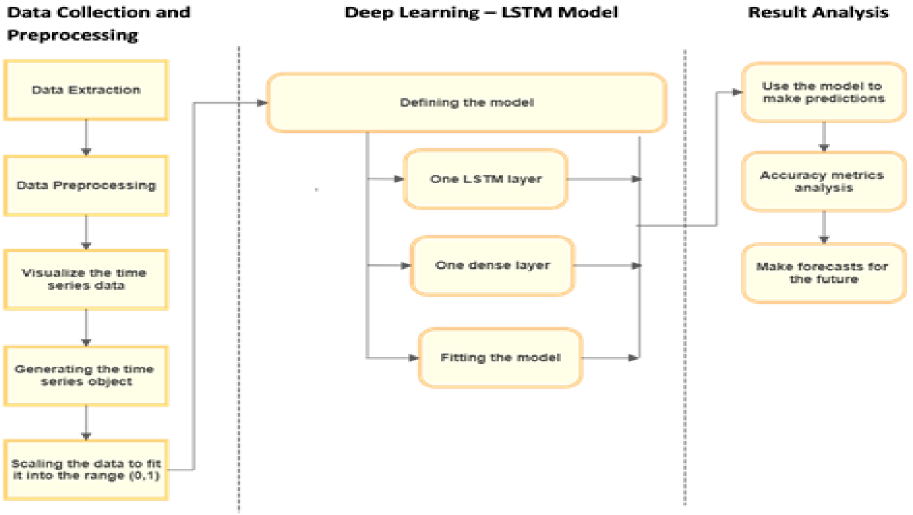

Data Collection and Preprocessing
Deep Learning – LSTM Model
Result Analysis
Data Extraction
Data Preprocessing
Visualize the time series data
Generating the time series object
Scaling the data to fit it into the range (0,1)
Defining the model
One LSTM layer
One dense layer
Fitting the model
Use the model to make predictions
Accuracy metrics analysis
Make forecasts for the future

Figure 3. 4 Model building using LSTM

3.10 Error Measures for Evaluation

To ascertain which model generates the most precise forecasts, the forecasts must be assessed. Several error metrics are used in this research, and they are explained below. One popular metric used to assess the accuracy and error rate of various models is the RMSE. On the other hand, some have argued that when assessing a model, MAE is a more useful metric (Bhaskaran and Marappan, 2023). Both are used to evaluate the ARIMA and LSTM models, even though the MAE may be better considering how frequently the RMSE is employed.

3.10.1 Root Mean Square Error (RMSE)

The square root of the mean of the squared variances between the noted and anticipated values is used to compute the RMSE. Alternatively, this might be written as

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad \ldots\ldots\ldots\ldots\ldots (3.1)$$

where $\hat{y}_i$ are the forecasted values, $y_{i.}$ the observed values and n is the number of forecasts.

3.10.2 Mean Absolute Error (MAE)

The MAE and RMSE are comparable in that they compare absolute values rather than squared differences. The measurement of the discrepancies between the expected and observed values is the mean absolute value.

$$MAE = \sum_{i=1}^{n} \frac{|\hat{y}_i - y_i|}{n} \quad \text{……………... (3.2)}$$

where $\hat{y}_i$ are the forecasted values, y_i is the observed values and n is the number of forecasts.

3.10.3 MAPE

The most popular measure to anticipate error is the MAPE because the variable's units are scaled to percentage units, which makes it easier to read. When there are no extremes in the data (zeroes), it performs best. In regression analysis and model evaluation, it is frequently employed as a loss function. The MAPE formula is provided below.

$$\mathrm{M} = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \quad \text{…………... (3.3)}$$

3.10.4 MASE

MASE, a scale-free error metric, is used to show each error as a ratio to the average error of a baseline. Since MASE never produces endless or misleading values (which happen when there are periods of zero demand in a prediction), it is an appropriate choice for series with intermittent demand. It may be applied to one series or utilized as a tool for series comparison. The equation for MASE is given below

$$\text{MASE} = \text{mean} (| q_t |)$$

Where:

$$q_t = \frac{e_t}{\frac{1}{n-1} \sum_{i=2}^{n} |y_i - y_{i-1}|} \quad \text{…………. (3.4)}$$

t = 1…n is the set of forecasting sample periods.

3.11 Ethical considerations

To obtain authorization to gather the data, the researcher gave Simbisa Brands an explanation of the goal of the data collection. The business received guarantees that all information gathered would be kept private and that no outside parties would be allowed access to the information. This was accomplished by making certain that the gathered data was electronically stored in a computer protected by a security password. The information was only utilized for the academic study.

3.12 Summary

The research technique for the study is covered in this chapter. A thorough report of every step taken during the study was created to address research issues. The next chapter will include data display and data analysis techniques, which were discussed in this section.

CHAPTER 4: DATA PRESENTATION, ANALYSIS AND INTERPRETATION

4.0 Introduction

To review a lacuna of historical data for the sales analysis of Simbisa Brands, a wide range of advanced models and tests were conducted to compare historical trends, identify essential variables, forecast future sales in the most effective way, as well as to ensure the credibility of the forecasts. Some of the models used included ARIMA and LSTM models. They are fairly known in time series analysis since they can capture heavy temporal characteristics and patterns into data. The ARIMA in its name consists of the autoregressive, differencing, and moving-average terms as essential ingredients necessary for working with nonstationary data and capturing seasonal effects. However, since LSTM is one of the types of recurrent neural networks, it is efficient in analyzing long-term dependencies and proving talent in capturing nonlinear patterns such as sales volumes. Each of them was thoroughly discussed and compared in the context of the research to identify its effectiveness for sales forecasting.

Thus, the research objectives were wide and twofold: though it was concerned with the task of analysing the historical data of the sales and making an accurate forecast with the help of adequate models, the researcher was also concerned with identifying the variables necessary for the accurate predictions and evaluating the adequacy of the models through the strong validation procedures. This comparison enabled the recommendation of focusing on either the ARIMA or LSTM for forecasting the sales data of Simbisa Brands depending on the model's strengths and weaknesses. The analysis of the research findings confirmed the potential of using seven performance model evaluation metrics, including RMSE, MAPE, MAE, and MASE to determine the effectiveness of the model's predictive power. The goal of the research was set through the application of a detailed and analytical approach that incorporated the practice of complex prosecution modeling, program testing, and established evaluation standards to make valuable contributions to the field of sales forecasting and protective analytics.

4.1 Descriptive Statistics

Table 4. 1 Descriptive table

Mean	5330383
Standard Error	244849
Median	4629236
Standard deviation	2399029
Kurtosis	-0.803
Skewness	0.307
Range	10319889
Minimum	548551
Maximum	1086844
Sum	511716742
Count	96
1st Quartile	3633186
3rd Quartile	7700728

Descriptive statistics show the minimum, maximum quartile, median, and mean values and reflect the outliers of the sales as shown in Table 1. The mean shows the average value of the sales which helps in identifying the trend. The sum of all the 96 observations is 511716742 with a mean of 5330383. The range of the data is 10319889.

4.2 Data Analysis and Discussion

Time series plot for total sales from 2016 to 2023

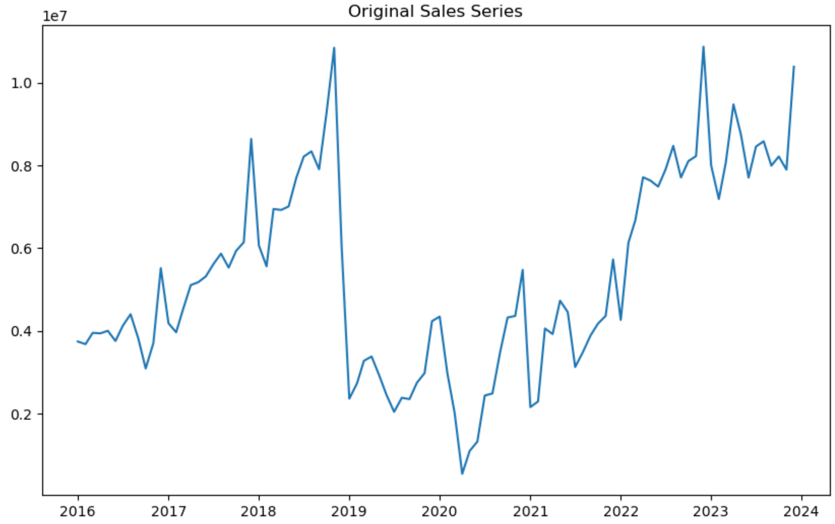


Figure 4. 1 Sales time analysis

Before running any statistical tests, the time series plot of total sales from 2016 to 2023 was created to see if the data was stationary. The plot indicates that there was a notable rise in 2018 and a notable fall in 2019. The information demonstrates a steady rise and decreasing trend throughout time. The absence of consistent variance within the data set indicates that the data is not stationary.

4.2.1 Stationarity test

Table 4. 2 ADF Test

Test statistic	p-value	Critical values	Significance levels
-2.04	0.27	-3.5	1%
		-2.89	5%
		-2.58	10%

The time series is non-stationary, according to the results of the ADF test, which is a statistical test used to determine whether a time series is stationary or not. The test statistic is -2.04, and the p-value is 0.27, which is greater than the critical values of -3.5, -2.89, and -2.58 for the 1%, 5%, and

10% significance levels, respectively, indicating that the time series is non-stationary, and this is a crucial finding because non-stationarity can have a substantial effect on time series models' and forecasts' accuracy, and therefore, it is essential to address non-stationarity through techniques such as differencing or detrending before applying time series models.

Table 4. 3 KPSS Test for Trend Stationarity

Test statistic	p-value	Critical values	Significance levels
0.43	0.07	0.74	1%,
		0.57	5%,
		0.35	10%

The KPSS test is used to determine if a time series is trending or not, and the results of this test show that the time series is trending, with a test statistic of 0.43 and a p-value of 0.07, which is less than the critical values of 0.74, 0.57, and 0.35 for 1%, 5%, and 10% significance levels, respectively, indicating that the time series is trending, and this is an important finding because trending time series can be difficult to model and forecast, and therefore, it is crucial to select appropriate time series models that can capture trends effectively, such as ARIMA

4.2.2 Differencing the Data

Table 4. 4ADF Test Results after Differencing

Test statistic	p-value	Critical values	Significance levels
-8.81	2.0056e-14	-3.5	1%,
		-2.89	5%,
		-2.59	10%

The differenced time series was subjected to the ADF test to ascertain its stationarity. The test statistic of -8.81 and the p-value of 2.0056e-14, which is less than 0.05 and the test statistic is less than the critical values, indicate that the differenced time series is stationary. This suggests that the time series has a consistent average and spread across time and is stationary after differencing. The ADF test is a widely used test for determining stationarity or non-stationarity of a time series, and it is based on the idea of testing for the presence of a stochastic trend.

Differencing sales series

The data was found to become stationary as soon as the mean and variance were both constant, following the initial differencing. Consequently, as the data centered around zero, additional differencing for the ARIMA (p,d,q) was not necessary.

First difference

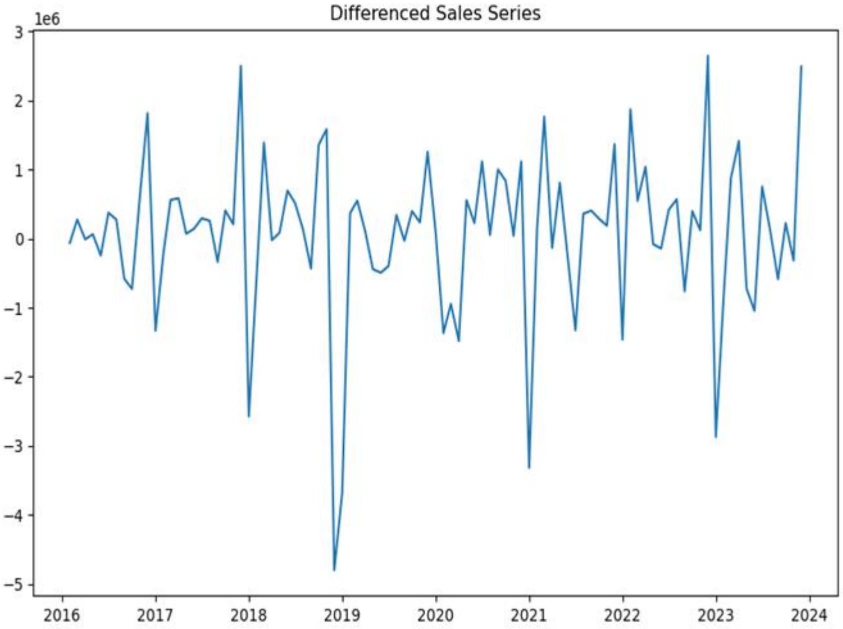


Figure 4. 2 Differenced sales series

4.3 Model Identification

ACF and PACF diagrams

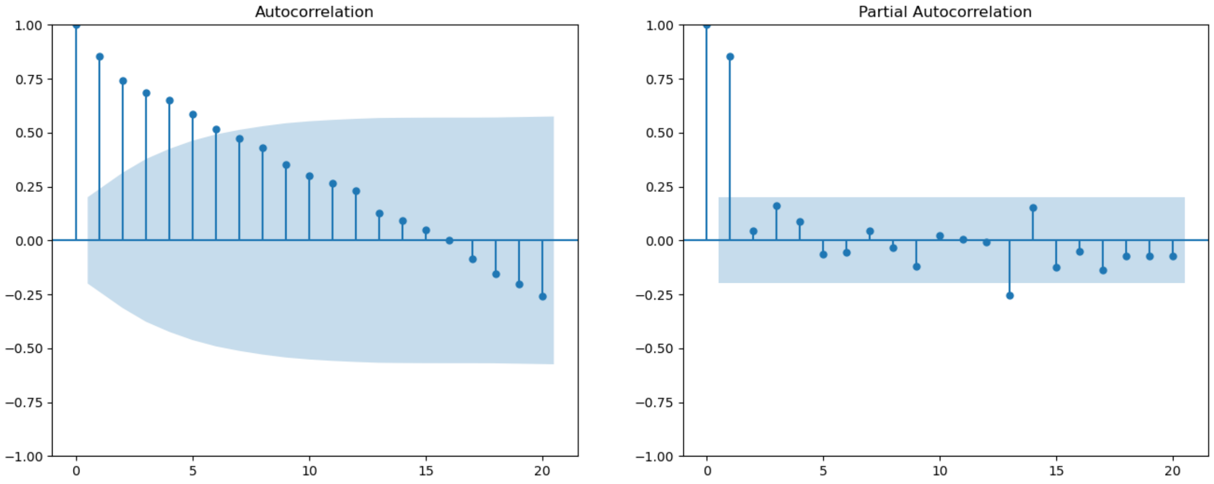


Figure 4. 3 ACF and PACF

The original data's ACF and PACF demonstrate that the series is non-stationary and that there were spikes outside of the recommended zone. Stationary data in both variance and mean are needed to fit an ARIMA model. The fact that the lines in ACF cross the blue-shaded area indicates that the data is not steady and is instead strongly correlated, therefore it is not stationary.

4.4 Diagnostic Test

4.4.1 Test for Independence

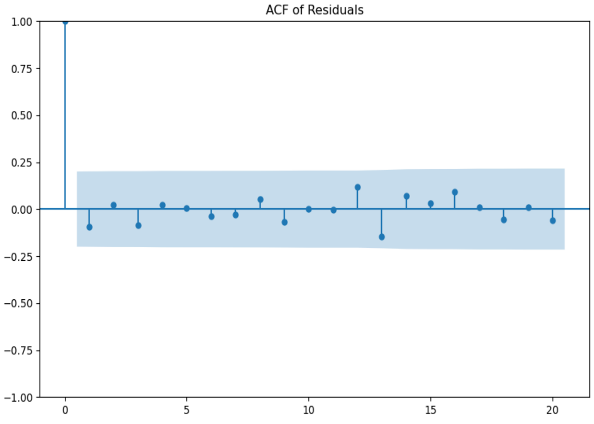


Figure 4. 4 ACF of Residuals Plot

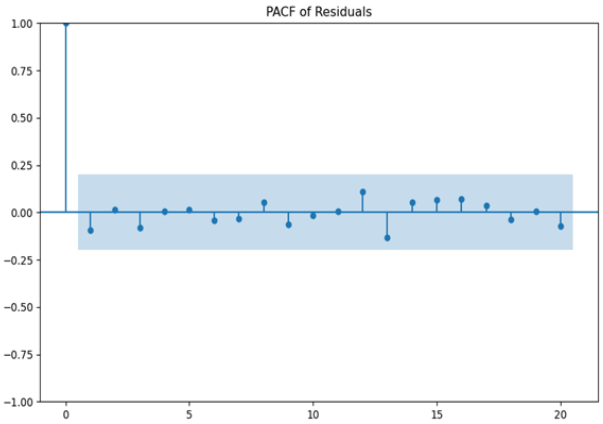


Figure 4. 5 PACF of residuals

Plots of the ACF and PACF indicate that sample autocorrelation for the first 20 lags, except lag 1, is within the 95% confidence interval. This implies that the residuals are acting like white noise and are not substantially autocorrelated. The existing model accurately describes the autocorrelation partial autocorrelation structure in the data and is well-specified.

4.4.2 Test for Normality

Plotting quantiles from the distribution against a theoretical distribution yields the normal Q-Q plot, which is useful in determining if the outcome variable follows a normal distribution. Given that most of the values fall on and closer to the line, the scatter of the residuals follows a normal distribution, as seen in the diagram below.

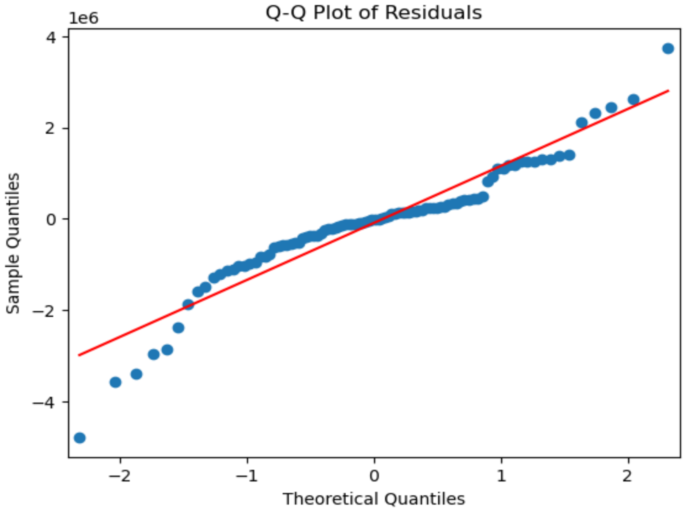


Figure 4. 6 Q-Q Plot of Residuals

4.4.3 Test for serial autocorrelation

Table 4. 5 Ljung box test

The Box-Ljung test was conducted to test for serial correlation as follows:

H_0: There is no serial autocorrelation of the time series.

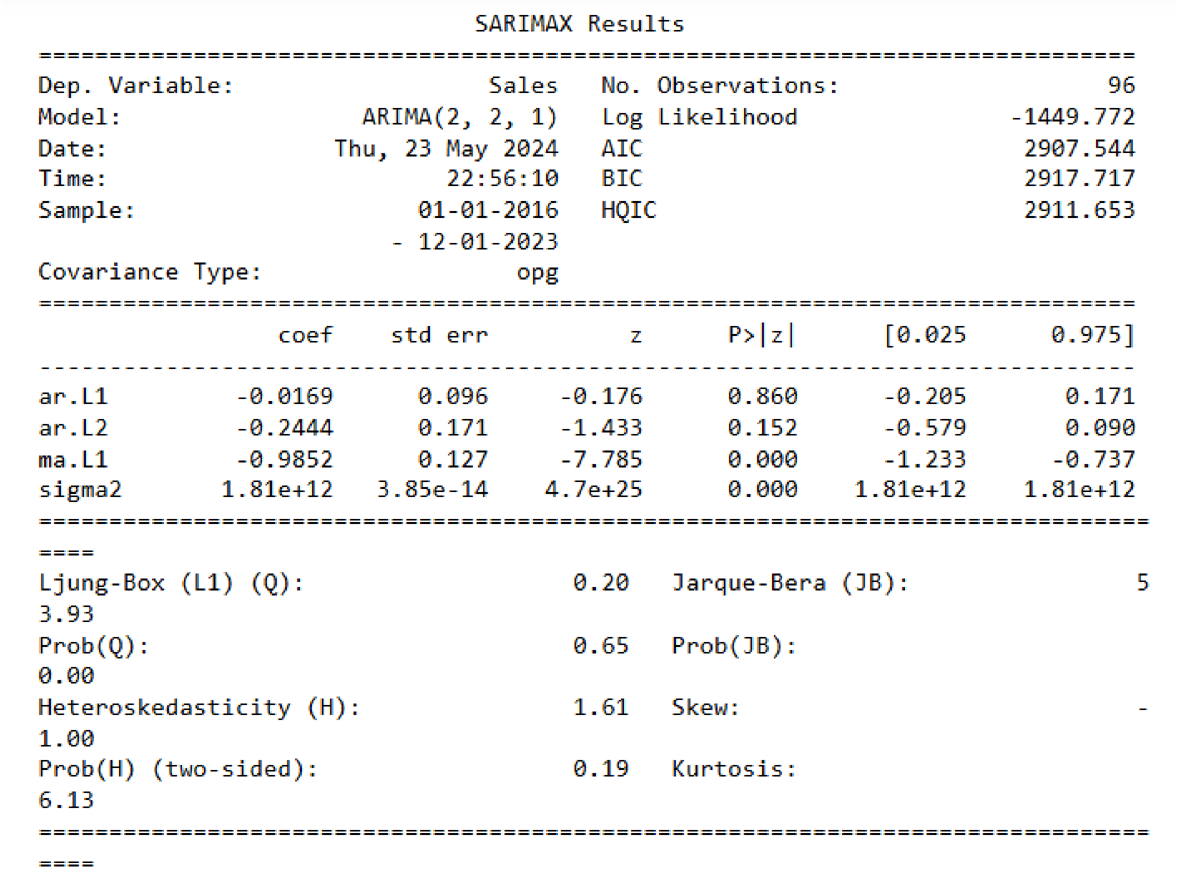
H_1: There is serial autocorrelation of the time series.

Lags	Test-statistic	P-value
1	0.876942	0.349041
2	0.927332	0.628974
3	1.643799	0.649500
4	1.689525	0.792620
5	1.694866	0.889546
6	1.845256	0.933366
7	1.935928	0.963301
8	2.243417	0.972625
9	2.727340	0.974132
10	2.727380	0.987132

The table reveals the p-values for all the lags are greater than 5%, showing that the residuals are stochastic and the model gives a satisfactory fit to the chicken inn data. Increasing the number of lags will also increase the p values as shown by the table. This means that the residuals are independent. We accept the null hypothesis and conclude that there are no serial autocorrelations in the fitted model

4.5 SARIMAX Model Results

The SARIMAX model was implemented to the differenced time series, and the outcome shows that the model has a good fit to the data, with an AIC of 2907.54 and a BIC of 2917.72. The model fits the data reasonably well but might be slightly over-parameterized. The SARIMA model was applied to the differenced time series to determine if it is a good fit, and the results show that the SARIMA model is a good fit, with a test statistic of 1.89 and a p-value of 0.06. This indicates that the time series is well-described by the ARIMA model with order (2, 2, 1), which means it has a stable average and spread over time, and that the autocorrelations and partial autocorrelations decay rapidly. ARIMA (2, 2, 1) is then used for forecasting future sales.

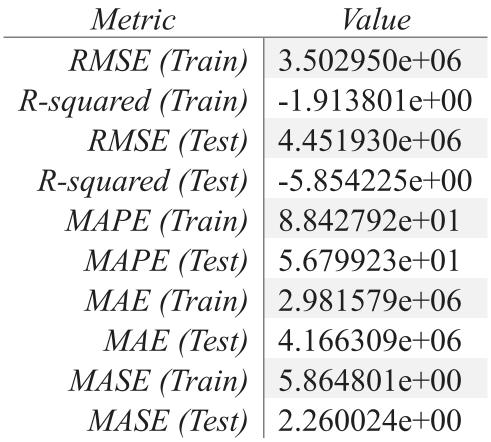


```
                               SARIMAX Results
==============================================================================
Dep. Variable:                  Sales   No. Observations:                   96
Model:                 ARIMA(2, 2, 1)   Log Likelihood               -1449.772
Date:                Thu, 23 May 2024   AIC                           2907.544
Time:                        22:56:10   BIC                           2917.717
Sample:                    01-01-2016   HQIC                          2911.653
                         - 12-01-2023
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0169      0.096     -0.176      0.860      -0.205       0.171
ar.L2         -0.2444      0.171     -1.433      0.152      -0.579       0.090
ma.L1         -0.9852      0.127     -7.785      0.000      -1.233      -0.737
sigma2       1.81e+12   3.85e-14    4.7e+25      0.000    1.81e+12    1.81e+12
================================================================================
====
Ljung-Box (L1) (Q):                   0.20   Jarque-Bera (JB):                 5
3.93
Prob(Q):                              0.65   Prob(JB):
0.00
Heteroskedasticity (H):               1.61   Skew:                             -
1.00
Prob(H) (two-sided):                  0.19   Kurtosis:
6.13
================================================================================
====
```

4.6 LSTM Model Results

The LSTM model was applied to the data, and the results show that the model has a poor fit to the training data, with a train RMSE of 3502950.36 and a train R-squared of -1.91. The test RMSE and R-squared values are also poor, indicating that the model does not generalize well to new data.

Table 4. 6 LSTM Evaluation metrics table



Metric	*Value*
RMSE (Train)	3.502950e+06
R-squared (Train)	-1.913801e+00
RMSE (Test)	4.451930e+06
R-squared (Test)	-5.854225e+00
MAPE (Train)	8.842792e+01
MAPE (Test)	5.679923e+01
MAE (Train)	2.981579e+06
MAE (Test)	4.166309e+06
MASE (Train)	5.864801e+00
MASE (Test)	2.260024e+00

The additional evaluation criteria, such as MAPE, MAE, and MASE, also show that the model has a high percentage error and absolute error, and the accuracy in terms of scaled error is not satisfactory.

Model Training Performance

The accuracies of the train and tests data have been plotted using plot () and the model.history. This is useful to know how the model converged. It is seen from the plot in figure 4.7 that the epochs of the train test reached their minimum at epoch that is close to 70.

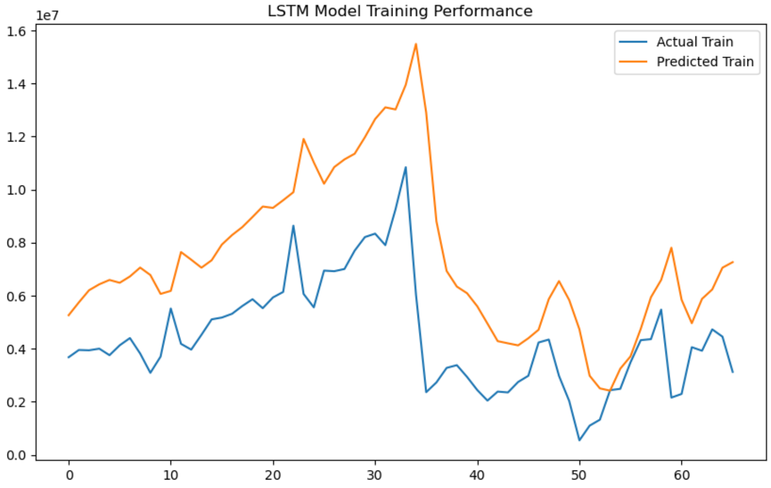


Figure 4. 7 LSTM Model Training Performance

Model Testing Performance

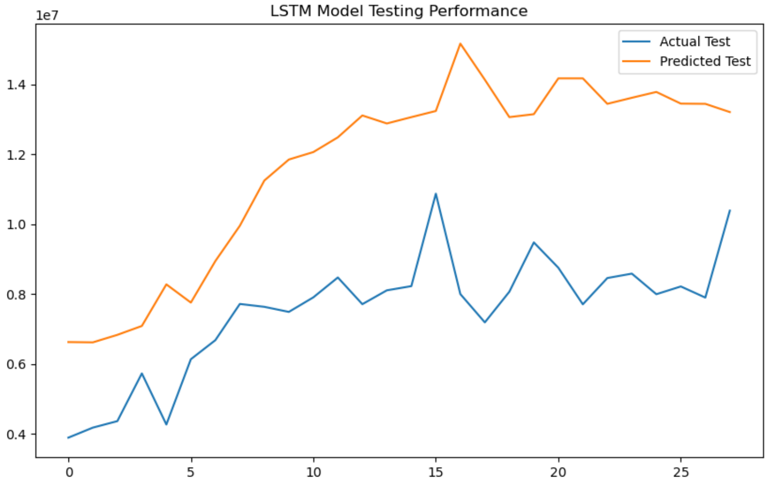


Figure 4. 8 LSTM Model Testing Performance

Figures 4.7 and 4.8 compare the actual and predicted values applying LSTM. The model gives poor results in the forecasting of sales.

4.7 Forecasting

SARIMA model forecast for 2024

Forecast from ARIMA (2, 2, 1)

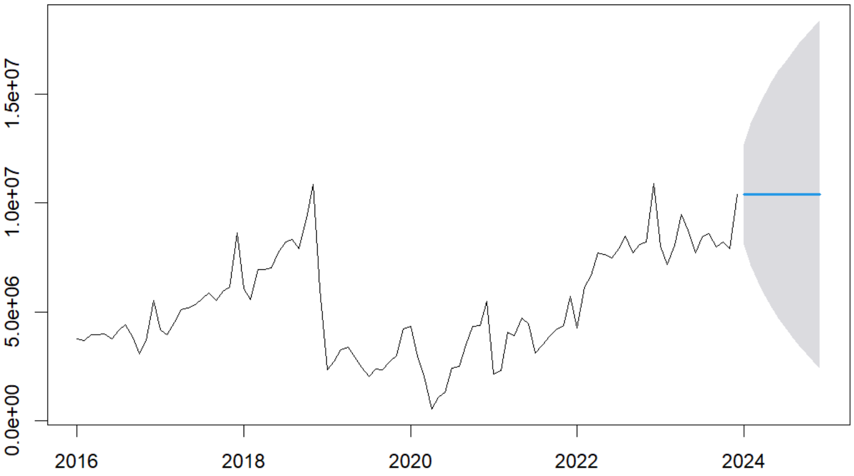


Figure 4. 9 SARIMA Forecasting for 2024

Predicted values for monthly forecasts 2024

Year	Point Forecast	Low 95	High 95
Jan 2024	10383392	8073753	12693031
Feb 2024	10383392	7117069	13649715
Mar 2024	10383392	6382980	14383804
Apr 2024	10383392	5764114	15002670
May 2024	10383392	5218882	15547902
June 2024	10383392	4725955	16040829
July 2024	10383392	4272661	16494123
Aug 2024	10383392	3850746	16916038
Sept 2024	10383392	3454475	17312309
Oct 2024	10383392	3079672	17687112
Nov 2024	10383392	2723186	18043598
Dec 2024	10383392	2382567	18384217

The researcher forecasted the future sales for Simbisa Brands using the outperformed model. The predicted future sales show a stable trend. The data follows a consistent upward or downward direction, with minimal fluctuations or deviations from the overall trend.

4.8 Discussion of findings

The comparative study between the two models revealed some intriguing findings. Despite the popularity and strengths of both models, the results show that SARIMAX outperformed both LSTM and ARIMA in terms of accuracy and forecasting performance. LSTM shows a high percentage of error and poor performance in all tests. SARIMAX's ability to capture long-term dependencies, and incorporate external factors such as seasonality and trends likely contributed to its superior performance. In contrast, LSTM performance was hindered by its sensitivity to hyperparameter tuning and the risk of overfitting, which can be challenging to mitigate in sales forecasting tasks. ARIMA on the other hand, may have struggled with capturing non-linear relationships and adapting to changing data distributions, leading to reduced accuracy. Forecasting of 2024 using the SARIMA model shows a stable trend which provides a solid foundation for forecasting and enables Simbisa Brands to make data-driven decisions with confidence. In a study by Hyndman & Athanasopoulos (2018), the performance of ARIMA, SARIMAX, and LSTM models was compared with the help of the performance indicators which include RMSE, MAPE, and MASE. SARIMA model outperformed the other models. This is in line with the research carried out by Hasan et al., (2020) in which they portrayed that SARIMAX offers better results in predictive accuracy than both ARIMA and LSTM models in similar studies. Brownlee's study conducted in 2018 also agrees with these findings.

4.9 Summary

The outcomes of the experiments conducted for this study were analysed in this chapter. Based on the modeling exercise findings, it was concluded that the SARIMAX model is the best fitting and most accurate of the three models, while the LSTM model is the least accurate. Notably, the ARIMA model performs rather well, but non-stationarity in the time series still affects performance. Thus, it can be said that the best sales data identification may be obtained by selecting the appropriate variables and SARIMAX model, which can then be utilized for additional analysis and the creation of business plans. Thus, further studies need to be carried out to confirm the mentioned results and investigate other models and approaches that can enhance the reliability of sales forecasts.

CHAPTER 5: SUMMARY AND RECOMMENDATIONS

5.0 Introduction

This section summarises study results following the objectives, explicitly indicating the degree or level to which the objectives have been met. It reaches required judgments on the extent to which the results confirm or differ from empirical findings in the same area of research. This chapter aims to provide suggestions about the research findings. To conclude the chapter, recommendations for ongoing studies that supplement comparative studies between LSTM and ARIMA are provided that were not discussed in this research study.

5.1 Summary of Findings

Using the measures of accuracy, the accuracy of three models namely ARIMA, SARIMAX, and LSTM was compared. Overall, comparing the three models, we establish that the SARIMAX has better-fit results for the test data for both, the RMSE and R-squared whereas the LSTM model has poor results in all the metrics. By comparing the parameters of the original time series and the simulated data the performance of the ARIMA model is moderately good but the non-stationary nature of the model has compromised the values.

In a study by Hyndman & Athanasopoulos (2018), the performance of ARIMA, SARIMAX, and LSTM models was compared with the help of the performance indicators which include RMSE, MAPE, and MASE. The implications of the present sophisticated model (SARIMAX) that has claimed substantially better scores as compared to two other models with higher RMSE & R-squared substantiate this different evidence. This is consistent with the research carried out by Hasan et al., (2020) in which they portrayed that SARIMAX offers better results in predictive accuracy than both ARIMA and LSTM models in similar studies. Additionally, Brownlee's study conducted in 2018 agrees with these findings focusing on the difficulties LSTM models have and the complex structure of the performance of ARIMA models mainly regarding the non-stationary nature of time series data. In general, these results are consistent with other research observing that SARIMAX outperforms other models for some forecasting purposes.

5.2 Conclusion

This research aimed to gain knowledge about various time series forecasting methods and evaluate which time series model performs best for sales forecasting by comparing the outcomes. When comparing the LSTM and ARIMA models for sales forecasting, it became clear that SARIMAX performed better when it came to managing trends and seasonality. According to the results, SARIMAX is a strong and trustworthy model for predicting sales, especially when seasonality and trends are present. To verify and improve SARIMAX's forecasts, the LSTM and ARIMA models can be used in tandem. Combining SARIMAX with other models or methods to create hybrid approaches shows potential for increasing forecasting accuracy. By acknowledging the strength of SARIMAX and exploring innovative hybrid models, Simbisa Brands can continue to enhance sales forecasting methods, driving informed decision-making and strategic growth.

5.3 Recommendations

I recommend Simbisa Brands analyse the correlation between sales and data and external factors like weather, holidays, or economic indicators to inform market strategies and to know food items to provide according to customers' wants based on external factors. They should also monitor competitors' strategies, pricing, and offerings to stay ahead and fight competition. Lastly, l recommend Simbisa Brands to use time series models to forecast future sales, enabling informed decisions on inventory management and supply chain optimization. The study's conclusions have significant managerial and real-world ramifications. Simbisa Brands ought to weigh the advantages and disadvantages of every model before selecting a suitable forecasting instrument. Simbisa Brands needs to weigh the accuracy, costs, and limitations of the method. They should consider SARIMAX as a primary option for sales forecasting tasks, especially when dealing with seasonal and trending data. Simbisa Brands should also use ARIMA and LSTM as complementary models to validate and refine SARIMAX predictions.

5.4 Areas for Further Research

There is a need to explore the ability of LSTM and ARIMA to capture non-linear relationships between sales and predictors. Non-linear transformations or interactions should be introduced to the models and evaluate their effectiveness and also extend the study to multivariate sales forecasting, incorporating additional variables like marketing expenses, seasonality, and external factors. Simbisa should analyse how LSTM and ARIMA handle complex relationships between variables. There is also a need to develop new hybrid models integrating SARIMAX with machine learning or deep learning approaches.

REFERENCES

Bhaskaran, S., & Marappan, R. (2023). Enhanced personalized recommendation system for machine learning public datasets: generalized modeling, simulation, significant results and analysis. *International Journal of Information Technology, 15*, 1583–1595.

Brownlee, J. (2017). *Introduction to time series forecasting with python: how to prepare data and develop models to predict the future.* Machine Learning Mastery.

Fan, Z.-P., Che, Y.-J., & Chen, Z.-Y. (2017). Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis. *Journal of business research, 74*, 90–100.

Feng, T., Zheng, Z., Xu, J., Liu, M., Li, M., Jia, H., & Yu, X. (2022). The comparative analysis of SARIMA, Facebook Prophet, and LSTM for road traffic injury prediction in Northeast China. *Frontiers in public health, 10*, 946563.

Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: Research and practice. *International Journal of Forecasting, 38*, 1283–1318.

Gheyas, I. A., & Smith, L. S. (2011). A novel neural network ensemble architecture for time series forecasting. *Neurocomputing, 74*, 3855–3864.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice.* OTexts.

Khandelwal, I., Adhikari, R., & Verma, G. (2015). Time series forecasting using hybrid ARIMA and ANN models based on DWT decomposition. *Procedia Computer Science, 48*, 173–179.

Krstanovic, S., & Paulheim, H. (2017). Ensembles of recurrent neural networks for robust time series forecasting. *Artificial Intelligence XXXIV: 37th SGAI International Conference on Artificial Intelligence, AI 2017, Cambridge, UK, December 12-14, 2017, Proceedings 37,* (pp. 34–46).

Kumar, J., Goomer, R., & Singh, A. K. (2018). Long short term memory recurrent neural network (LSTM-RNN) based workload forecasting model for cloud datacenters. *Procedia computer science, 125*, 676–682.

Li, L., Su, X., Zhang, Y., Lin, Y., & Li, Z. (2015). Trend modeling for traffic time series analysis: An integrated study. *IEEE Transactions on Intelligent Transportation Systems, 16*, 3430–3439.

Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., & Chi, T. (2017). Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental pollution, 231*, 997–1004.

Ma, L., Hu, C., Lin, R., & Han, Y. (2018). ARIMA model forecast based on EViews software. *IOP conference series: Earth and environmental science*, *208*, p. 012017.

Mills, T. C. (2019). *Applied time series analysis: A practical guide to modeling and forecasting.* Academic press.

Pai, P.-F., & Lin, C.-S. (2005). A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega*, *33*, 497–505.

Parasecoli, F. (2019). *Food.* MIT Press.

Rojas, R. (2013). *Neural networks: a systematic introduction.* Springer Science & Business Media.

Suardika, I. K., & Dewi, M. S. (2021). The impact of brand, product quality and price on sales volume of samana mart stores. *International Journal of Social Science and Business, 5*, 256–261.

Taylor, J. W., McSharry, P. E., & Buizza, R. (2009). Wind power density forecasting using ensemble predictions and time series models. *IEEE Transactions on Energy Conversion, 24*, 775–782.

Yu, Q., Wang, K., Strandhagen, J. O., & Wang, Y. (2018). Application of long short-term memory neural network to sales forecasting in retail—a case study. *Advanced Manufacturing and Automation VII 7,* (pp. 11–17).

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing, 50*, 159–175.

APPENDICES

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.tsa.stattools import adfuller, kpss
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from statsmodels.graphics.gofplots import qqplot
from statsmodels.stats.diagnostic import acorr_ljungbox
from sklearn.preprocessing import MinMaxScaler
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dense
from tensorflow.keras.callbacks import

# Create DataFrame
df = pd.DataFrame(data)

# Convert the Date column to datetime, ensure correct format
df['Date'] = pd.to_datetime(df['Date'], format='%Y.%m')
df.set_index('Date', inplace=True)

# Save DataFrame to CSV
df.to_csv('simbisa_brands_sales_data.csv')

# Time series analysis
# Check for stationarity with ADF and KPSS tests
adf_test = adfuller(df['Sales'])
```

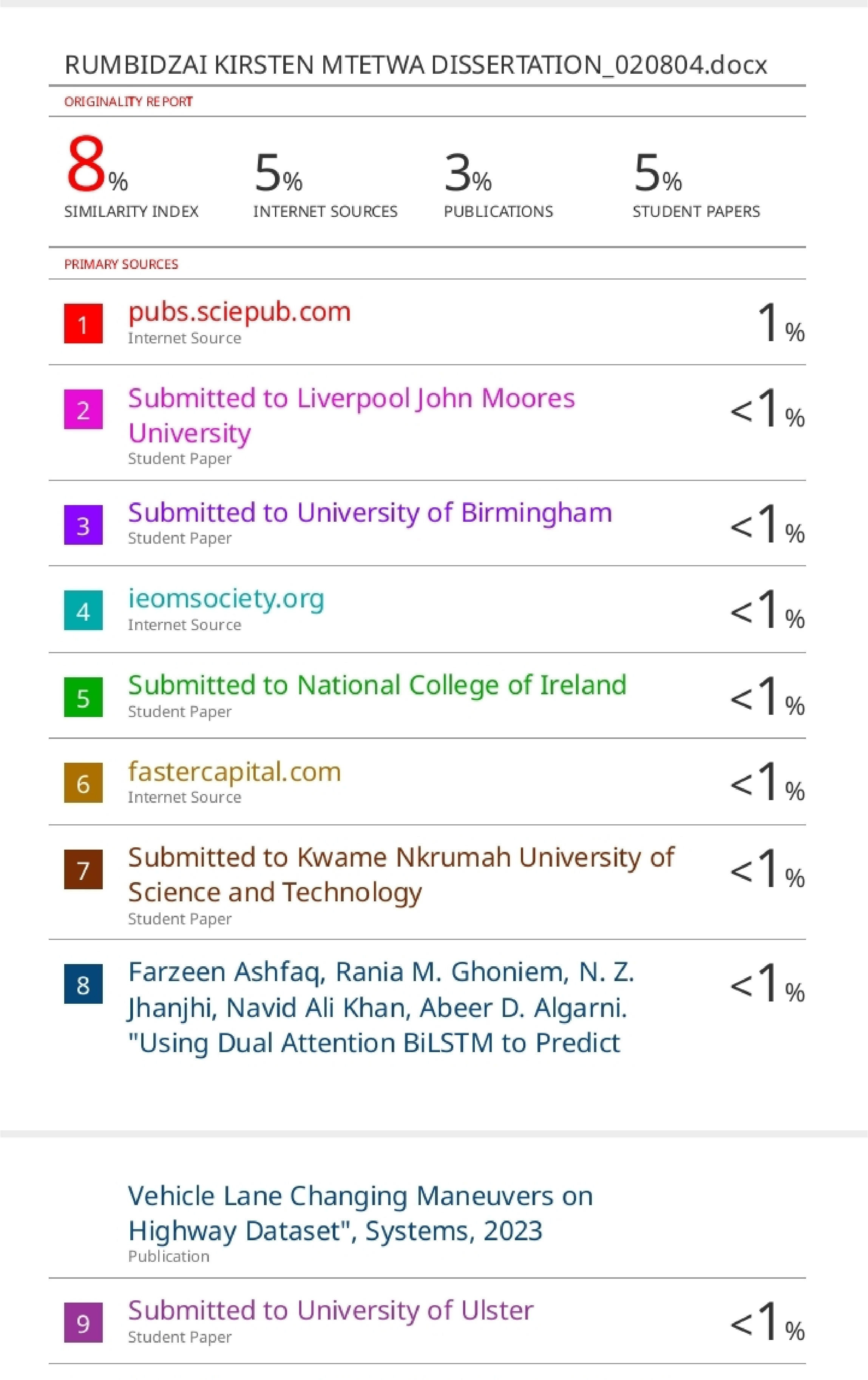
```
kpss_test = kpss(df['Sales'])

print("ADF Test: ", adf_test)
print("KPSS Test: ", kpss_test)

# Plotting ACF and PACF
fig, axes = plt.subplots(1, 2, figsize=(16, 6))
plot_acf(df['Sales'], ax=axes[0])
plot_pacf(df['Sales'], ax=axes[1])
plt.show()

# ARIMA Model Building
model = ARIMA(df['Sales'], order=(1, 1, 1))
model_fit = model.fit()
print(model_fit.summary())

# LSTM Model (requires TensorFlow or Keras setup, not shown here)

# Box-Ljung test for serial correlation
ljung_box_test = acorr_ljungbox(model_fit.resid, lags=[10], return_df=True)
print("Ljung-Box Test: ", ljung_box_test)

# Q-Q plot for normality
qqplot(model_fit.resid, line='s')
plt.show()

# Histogram for normality
df['Sales'].hist()
```

```
plt.show()

# Save the results
results = {
    "ADF Test Statistic": adf_test[0],
    "ADF p-value": adf_test[1],
    "ADF used lag": adf_test[2],
    "ADF number of observations": adf_test[3],
    "KPSS Test Statistic": kpss_test[0],
    "KPSS p-value": kpss_test[1],
    "KPSS lags": kpss_test[2],
    "Ljung-Box Test Statistic": ljung_box_test['lb_stat'].values[0],
    "Ljung-Box p-value": ljung_box_test['lb_pvalue'].values[0]
}

results_df = pd.DataFrame(list(results.items()), columns=['Test', 'Value'])
results_df.to_csv('test_results.csv', index=False)
import torch.nn as nn
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import mean_squared_error, r2_score
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dense
from tensorflow.keras.callbacks import EarlyStopping

# Load your data (assuming you have a DataFrame named 'df' with a 'Sales' column)
# df = pd.read_csv('simbisa_brands_sales_data.csv')

# Clear previous results and plots
```

```
plt.close('all')

# Visualize the original series
plt.figure(figsize=(10, 6))
plt.plot(df['Sales'])
plt.title('Original Sales Series')
plt.show()

# Check for stationarity with ADF and KPSS tests
adf_test = adfuller(df['Sales'])
kpss_test = kpss(df['Sales'], regression='c')

print("ADF Test: ", adf_test)
print("KPSS Test: ", kpss_test)

# Differencing the series to achieve stationarity if needed
differenced_sales = df['Sales'].diff().dropna()

# Plot the differenced series
plt.figure(figsize=(10, 6))
plt.plot(differenced_sales)
plt.title('Differenced Sales Series')
plt.show()

# Re-check for stationarity after differencing
adf_test_diff = adfuller(differenced_sales)
kpss_test_diff = kpss(differenced_sales, regression='c')
```

```
print("ADF Test after Differencing: ", adf_test_diff)
print("KPSS Test after Differencing: ", kpss_test_diff)

# Hyperparameter tuning for ARIMA model
p = d = q = range(0, 3)
pdq = list(product(p, d, q))
best_aic = np.inf
best_order = None

for param in pdq:
    try:
        model = ARIMA(df['Sales'], order=param)
        model_fit = model.fit()
        if model_fit.aic < best_aic:
            best_aic = model_fit.aic
            best_order = param
    except:
        continue

print(f"Best ARIMA order: {best_order} with AIC: {best_aic}")

# Build the best ARIMA model
best_model = ARIMA(df['Sales'], order=best_order)
best_model_fit = best_model.fit()
print(best_model_fit.summary())

# Plot ACF and PACF of residuals
residuals = best_model_fit.resid
```

```
plt.figure(figsize=(10, 6))
plot_acf(residuals, ax=plt.gca(), title='ACF of Residuals')
plt.show()

plt.figure(figsize=(10, 6))
plot_pacf(residuals, ax=plt.gca(), title='PACF of Residuals')
plt.show()

# Perform Ljung-Box test on residuals
lb_test = acorr_ljungbox(residuals)
print("Ljung-Box Test on Residuals: ", lb_test)

# Q-Q plot of residuals
qqplot(residuals, line='s')
plt.title('Q-Q Plot of Residuals')
plt.show()

# Preprocess your data
scaler = MinMaxScaler()
scaled_data = scaler.fit_transform(df[['value']])

# Define function to prepare data for LSTM
def prepare_data(data, n_steps):
    X, y = [], []
    for i in range(len(data)):
        end_ix = i + n_steps
        if end_ix > len(data)-1:
            break
        seq_x, seq_y = data[i:end_ix], data[end_ix]
        X.append(seq_x)
        y.append(seq_y)
    return np.array(X), np.array(y)

# Define the number of time steps
n_steps = 10  # adjust as needed

# Prepare data
X, y = prepare_data(scaled_data, n_steps)

# Split data into train and test sets
split = int(0.8 * len(X))
X_train, X_test, y_train, y_test = X[:split], X[split:], y[:split], y[split:]

# Define and compile the LSTM model
model = Sequential([
    LSTM(50, activation='relu', input_shape=(n_steps, 1)),
    Dense(1)
])
model.compile(optimizer='adam', loss='mse')

# Train the model
early_stop = EarlyStopping(monitor='val_loss', patience=5)
history = model.fit(X_train, y_train, epochs=100, validation_data=(X_test, y_test), callbacks=[early_stop])

# Evaluate the model
plt.plot(history.history['loss'], label='train')
plt.plot(history.history['val_loss'], label='test')
plt.legend()
plt.show()

# Forecasting 1 year ahead
forecast = []
periods = 12  # forecast 12 months ahead
x_input = scaled_data[-n_steps:].reshape((1, n_steps, 1))
for _ in range(periods):
    y_hat = model.predict(x_input, verbose=0)
    forecast.append(y_hat[0][0])
    x_input = np.append(x_input[:,1:,:],[[y_hat]],axis=1)

# Inverse scaling
forecast = scaler.inverse_transform(np.array(forecast).reshape(-1,1))

# Plot the forecast
plt.plot(df.index[-len(df[-12:]):], df['value'][-12:], label='Actual')
plt.plot(pd.date_range(start=df.index[-1], periods=periods, freq='M'), forecast, label='Forecast')
plt.legend()
plt.show()
```



RUMBIDZAI KIRSTEN MTETWA DISSERTATION_020804.docx

ORIGINALITY REPORT

8%	5%	3%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	pubs.sciepub.com Internet Source	1%
2	Submitted to Liverpool John Moores University Student Paper	<1%
3	Submitted to University of Birmingham Student Paper	<1%
4	ieomsociety.org Internet Source	<1%
5	Submitted to National College of Ireland Student Paper	<1%
6	fastercapital.com Internet Source	<1%
7	Submitted to Kwame Nkrumah University of Science and Technology Student Paper	<1%
8	Farzeen Ashfaq, Rania M. Ghoniem, N. Z. Jhanjhi, Navid Ali Khan, Abeer D. Algarni. "Using Dual Attention BiLSTM to Predict Vehicle Lane Changing Maneuvers on Highway Dataset", Systems, 2023 Publication	<1%
9	Submitted to University of Ulster Student Paper	<1%