Asian Research Journal of Mathematics

Volume 20, Issue 7, Page 9-30, 2024; Article no.ARJOM.118177
ISSN: 2456-477X

A Comprehensive Study of Walmart Sales Predictions Using Time Series Analysis

Cyril Neba C. [a*], Gerard Shu F. [b], Gillian Nsuh [c], Philip Amouda A. [a], Adrian Neba F. [a], F. Webnda [a], Victory Ikpe [d], Adeyinka Orelaja [a] and Nabintou Anissia Sylla [e]

[a] *Department of Mathematics and Computer Science, Austin Peay State University, Clarksville, Tennessee, USA.*
[b] *Montana State University, Gianforte School of Computing, Bozeman, Monatana, USA.*
[c] *School of Business Analytics, University of Quinnipiac, Hamden, Connecticut, USA.*
[d] *Department of Economics and Decision Sciences, Western Illinois University, Macomb, Illinois, USA.*
[e] *University of Arkansas, Graduate and Interdisciplinary Studies, Statistics and Analytics, Fayetteville, Arkansas, USA.*

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: https://doi.org/10.9734/arjom/2024/v20i7809

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/118177

Original Research Article

Received: 18/04/2024
Accepted: 20/06/2024
Published: 26/06/2024

Abstract

This article presents a comprehensive study of sales predictions using time series analysis, focusing on a case study of Walmart sales data. The aim of this study is to evaluate the effectiveness of various time series forecasting techniques in predicting weekly sales data for Walmart stores. Leveraging a dataset from Kaggle comprising weekly sales data from various Walmart stores around the United States, this study explores the effectiveness of time series analysis in forecasting future sales trends. Various time series analysis

*Corresponding author: Email: cneba@my.apsu.edu;

Cite as: *C., Cyril Neba, Gerard Shu F., Gillian Nsuh, Philip Amouda A., Adrian Neba F., F. Webnda, Victory Ikpe, Adeyinka Orelaja, and Nabintou Anissia Sylla. 2024. "A Comprehensive Study of Walmart Sales Predictions Using Time Series Analysis". Asian Research Journal of Mathematics 20 (7):9-30. https://doi.org/10.9734/arjom/2024/v20i7809.*

techniques, including Auto Regressive Integrated Moving Average (ARIMA), Seasonal Auto Regressive Integrated Moving Average (SARIMA), Prophet, Exponential Smoothing, and Gaussian Processes, are applied to model and forecast Walmart sales data. By comparing the performance of these models, the study seeks to identify the most accurate and reliable methods for forecasting retail sales, thereby providing valuable insights for improving sales predictions in the retail sector. The study includes an extensive exploratory data analysis (EDA) phase to preprocess the data, detect outliers, and visualize sales trends over time. Additionally, the article discusses the partitioning of data into training and testing sets for model evaluation. Performance metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are utilized to compare the accuracy of different time series models.

The results indicate that Gaussian Processes outperform other models in terms of accuracy, with an RMSE of 34,116.09 and an MAE of 25,495.72, significantly lower than the other models evaluated. For comparison, ARIMA and SARIMA models both yielded an RMSE of 555,502.2 and an MAE of 462,767.3, while the Prophet model showed an RMSE of 567,509.2 and an MAE of 474,990.8. Exponential Smoothing also performed well with an RMSE of 555,081.7 and an MAE of 464,110.5. These findings suggest the potential of Gaussian Processes for accurate sales forecasting. However, the study also highlights the strengths and weaknesses of each forecasting methodology, emphasizing the need for further research to refine existing techniques and explore novel modeling approaches. Overall, this study contributes to the understanding of time series analysis in retail sales forecasting and provides insights for improving future forecasting endeavors.

Keywords: Sales prediction; time series analysis; walmart; ARIMA; SARIMA; prophet; exponential smoothing; Gaussian processes; forecasting; model evaluation.

1 Introduction

In many fields, predictive modeling has become an essential tool for forecasting future trends and helping decision-makers in making data-driven choices [1]. For instance, in the financial sector, predictive techniques play a crucial role in detecting fraud [2,3]. Similarly, in the health sector, predictive models help in understanding and mitigating the impact of environmental factors like ambient ozone pollution on public health [4].

Time series analysis is particularly vital for forecasting future values based on previously collected data points arranged chronologically [5]. This methodology is not only applied in economic and financial contexts, such as predicting stock prices [6], but also in monitoring public health trends, as seen in the analysis of COVID-19 cases [7,8].

In this paper, we explore the use of Walmart sales data from Kaggle, a well-known platform for data science competitions and datasets, to construct predictive models through the application of time series analysis. The Walmart sales dataset was collected from various Walmart stores around the United States from the period of February 5, 2010, to October 26, 2012.

Along with extra features like holiday flags, temperature, fuel prices, the Consumer Price Index (CPI), and unemployment rates, the dataset comprises weekly sales data from various Walmart stores [9]. Our goal is to create precise predictive models that can predict future sales based on historical trends and pertinent predictors by utilizing this dataset.

In order to identify underlying patterns, trends, and seasonal variations, time series analysis entails evaluating and modeling data points gathered at regular intervals over an extended period of time [10]. It includes a range of approaches, including autocorrelation analysis, decomposition, smoothing techniques, and forecasting strategies like seasonal ARIMA and ARIMA (Auto Regressive Integrated Moving Average) [11]. With the help of these methods, we are able to identify the temporal dependencies in the data and forecast future trends in sales.

The first part of our study focuses on employing time series analysis techniques to model and forecast Walmart sales data. We will explore methods to preprocess the data, identify seasonality and trends, and select appropriate models to generate accurate forecasts. By evaluating the performance of our time series models against historical data, we aim to assess their effectiveness in capturing the inherent patterns and variability present in Walmart sales.

We aim to show in this paper the applicability and effectiveness of time series analysis for predictive modeling tasks, with a focus on retail sales forecasting. Furthermore, our objective is to offer an understanding of the real-world uses of time series methods and how they affect Walmart-style decision-making.

1.1 Importance of time series analysis in sales forecasting

Time series analysis plays a crucial role in sales forecasting, providing valuable insights into past trends and patterns that can inform future predictions. Some importance of time series analysis in sales forecasting are as follows.

Identification of Trends and Seasonal Patterns: Time series analysis helps identify trends, seasonal variations, and cyclical patterns in sales data [10]. These insights enable businesses to anticipate demand fluctuations and adjust their strategies accordingly [12].

Forecasting Accuracy: By analyzing historical sales data using time series techniques, businesses can develop accurate forecasting models [11]. These models take into account past sales performance and external factors, leading to more reliable predictions [13].

Resource Allocation and Inventory Management: Sales forecasts derived from time series analysis help businesses allocate resources efficiently and manage inventory levels effectively [14]. This optimization minimizes stockouts and excess inventory, enhancing operational efficiency [15].

Marketing and Promotion Strategies: Understanding sales patterns over time enables businesses to optimize their marketing and promotion strategies [16]. Time series analysis helps identify peak sales periods and consumer behavior trends, guiding targeted marketing efforts [17].

Budgeting and Financial Planning: Accurate sales forecasts derived from time series analysis facilitate budgeting and financial planning processes [18]. Businesses can use these forecasts to set realistic revenue targets, allocate budgets effectively, and make informed investment decisions [19].

1.2 Background

Based on past data observations, time series analysis predictions provide insightful information about potential future trends and patterns. In order to predict future values, time series forecasting entails evaluating sequential data points gathered over time [10]. Applications of this methodology can be found in a number of disciplines, including epidemiology, economics, finance, and climate science, where the ability to recognize and anticipate temporal patterns is essential for making decisions.

For example, time series analysis was used in a study by Cyril [20] to predict COVID-19 trends in Coffee County, Tennessee, United States. To forecast future infection rates, the researchers used time series forecasting techniques and historical data on COVID-19 cases. They were able to offer insightful analysis of historical trends and patterns, which helped policymakers and public health experts create efficient plans for handling the pandemic.

As stated by Shumway [11], time series forecasting techniques range from straightforward approaches like exponential smoothing to more intricate models like ARIMA and machine learning algorithms like LSTM (Long Short-Term Memory) networks. By identifying underlying patterns, seasonality, and trends in the data, these techniques enable precise forecasting and well-informed decision-making.

Time series analysis predictions are therefore essential in many fields because they use past data to predict future trends and patterns. Organizations and policymakers can make well-informed decisions to address challenges and capitalize on opportunities in their respective fields by having a thorough understanding of past behaviors and trends.

In the realm of retail sales forecasting, various methodologies have been explored to enhance prediction accuracy and efficacy. A comparative study focusing on linear and nonlinear models for aggregate retail sales forecasting was conducted by Chu [21] and cyril [22]. Their research delved into the effectiveness of traditional seasonal forecasting methods alongside neural networks, highlighting the significance of seasonal adjustments and deseasonalization techniques in improving forecasting accuracy. This study laid the groundwork for understanding the benefits and limitations of different modeling approaches in the retail domain.

In a recent development, a novel approach to retail sales forecasting was introduced by Ma [23] through meta-learning. Their study proposed a meta-learning framework based on deep convolutional neural networks, which automatically learn feature representations from raw sales time series data. By combining these learned features with a set of weights, the framework integrates a pool of base-forecasting methods, ultimately leading to superior forecasting performance compared to conventional benchmarks. Ma and Fildes' work underscores the importance of leveraging advanced machine learning techniques to enhance forecasting capabilities in the retail sector.

Furthermore, evaluating the performance of state space and ARIMA models in consumer retail sales forecasting was contributed to the field by Ramos [24]. Their research provided insights into the comparative effectiveness of these models, demonstrating their applicability in predicting sales trends for various product categories. Through the analysis of different methodologies, Ramos et al. shed light on the strengths and weaknesses of state space and ARIMA models, offering valuable guidance for researchers and practitioners seeking to optimize forecasting strategies in retail environments.

Integrating findings from these seminal works into the background of our study enriches our understanding of the diverse methodologies employed in retail sales forecasting. By building upon the insights gleaned from examination of linear and nonlinear models [21], innovative meta-learning approach [23], and comparative analysis of state space and ARIMA models [24], our research aims to contribute to this evolving field by exploring the performance of various time series models in predicting Walmart sales data.

Predictive modeling has become a vital tool across various fields, aiding decision-makers by forecasting future trends based on historical data. This study leverages Walmart sales data from Kaggle, covering weekly sales from February 5, 2010, to October 26, 2012. The goal is to develop accurate predictive models to forecast future sales by analyzing historical trends and relevant predictors using time series analysis techniques. These techniques, including ARIMA, SARIMA, Prophet, Exponential Smoothing, and Gaussian Processes, help identify temporal dependencies and forecast future sales trends. This paper aims to demonstrate the applicability and effectiveness of time series analysis in retail sales forecasting and its impact on decision-making in Walmart-like retail environments. Additionally, the importance of time series analysis in sales forecasting is highlighted, showing how it aids in identifying trends, improving forecasting accuracy, optimizing resource allocation, and enhancing marketing strategies. By incorporating insights from past studies on time series forecasting, this research aims to further the understanding of predictive modeling in retail sales.

2 Materials and Methods

This section outlines the comprehensive methodology adopted for the sales prediction study, including data collection and preparation, exploratory data analysis, model building, model evaluation and comparison.

2.1 Materials

2.1.1 Data collection

The dataset utilized in this study was sourced from Kaggle, encompassing weekly sales data from multiple Walmart stores across the United States. The dataset contains comprehensive information on sales trends over a specified timeframe, providing a rich source for time series analysis.

It's important to note that the dataset has a limited time range, covering the period from 2010 to 2012. This restriction stems from the unavailability of more recent data due to constraints imposed by Walmart. The company, understandably, restricts the dissemination of its recent sales data for proprietary and competitive reasons. Consequently, researchers and analysts are constrained to rely on historical data for their analyses, limiting the ability to capture and analyze more recent trends or changes in consumer behavior hence its applicability to current market conditions may be limited, necessitating caution in extrapolating findings to present-day contexts. Researchers should consider this limitation when interpreting the findings and extrapolating them to current market conditions.

Despite its age, the Walmart sales dataset offers a comprehensive and valuable resource for understanding sales dynamics in the retail industry. By leveraging the rich store-level data and information provided in the dataset, researchers and practitioners can gain insights into sales trends, identify key drivers of sales performance, and develop effective strategies for maximizing revenue and enhancing business outcomes.

2.2 Methods

2.2.1 Exploratory data analysis (EDA)

Before model implementation, an extensive exploratory data analysis (EDA) phase was conducted. This involved detecting outliers and addressing them using winsorization, a robust technique to handle extreme values without removing them entirely. By winsorizing the dataset, outliers were replaced with more moderate values, ensuring that they did not unduly influence the analysis.

Data Import and Sorting:

- The Walmart sales data was imported into R-Studio from the provided CSV file.
- The Date column was converted to the Date format, ensuring consistency and ease of manipulation.
- The dataset was sorted by date to arrange the observations chronologically.

Decomposition plot: A decomposition plot is essential for gaining a deeper understanding of the underlying structure of the Walmart sales data, identifying key trends and patterns. By examining these components of the decomposition plot, we can gain insights into the underlying patterns, trends, and seasonal variations present in the data, which can be useful for forecasting, analysis, and decision-making.

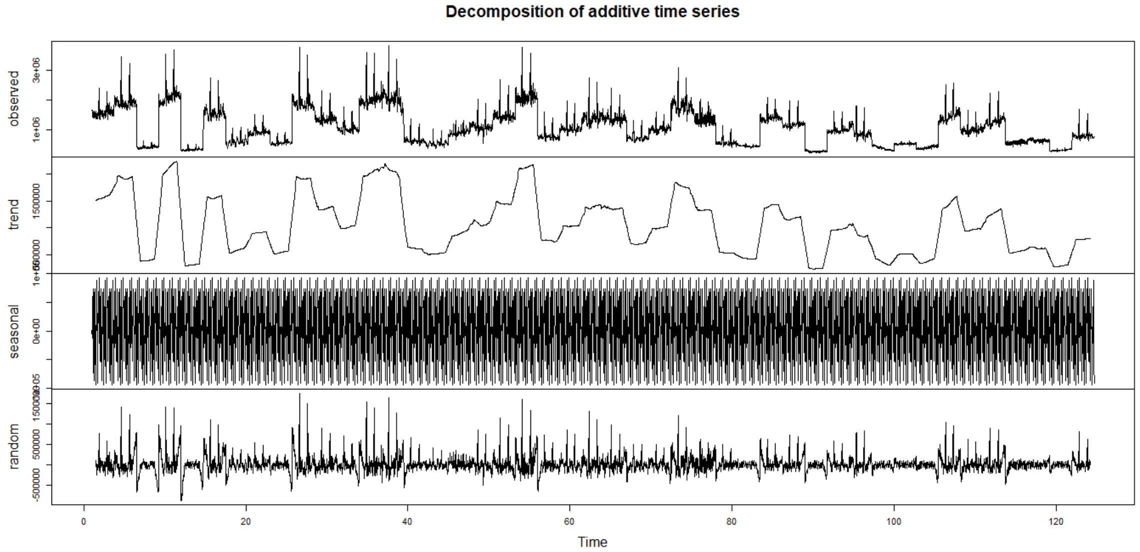

Fig. 1. Decomposition plot for Walmart sales

The trend component of the decomposed time series plot (Fig. 1) exhibits a pattern of going up and down over time, it suggests that the overall direction of the data is not consistently increasing or decreasing but rather fluctuating cyclically. This observation aligns well with the presence of cyclical patterns in the Walmart sales data. Seasonal effects, such as changes in consumer behavior, purchasing patterns, and demand for certain products, often lead to fluctuations in sales over time.

The following can be deduced from the cyclical pattern presented in the Walmart data.

Seasonal Variation: Seasonal changes, such as holidays, back-to-school seasons, summer vacations, and winter holidays, significantly impact consumer spending habits and preferences.

For example, sales of certain items like outdoor furniture, grills, and swimwear may peak during the summer season, while sales of winter clothing, holiday decorations, and cold-weather accessories may increase during the winter season. These seasonal variations create cyclical patterns in sales data, with sales rising and falling in a predictable manner as seasons change throughout the year.

Impact on Sales: The cyclical pattern underscores the influence of seasonal factors on Walmart's sales performance.

As seasons alternate, customers' needs, preferences, and purchasing behavior shift accordingly, leading to fluctuations in sales volumes and revenue. Understanding these seasonal patterns is therefore crucial for Walmart to optimize inventory management, pricing strategies, marketing campaigns, and staffing levels to meet customer demand effectively during peak seasons.

2.2.2 Outlier detection

Box Plots and Scatter Plots: Box plots and scatterplots were utilized to visualize potential outliers in the Weekly_Sales column.

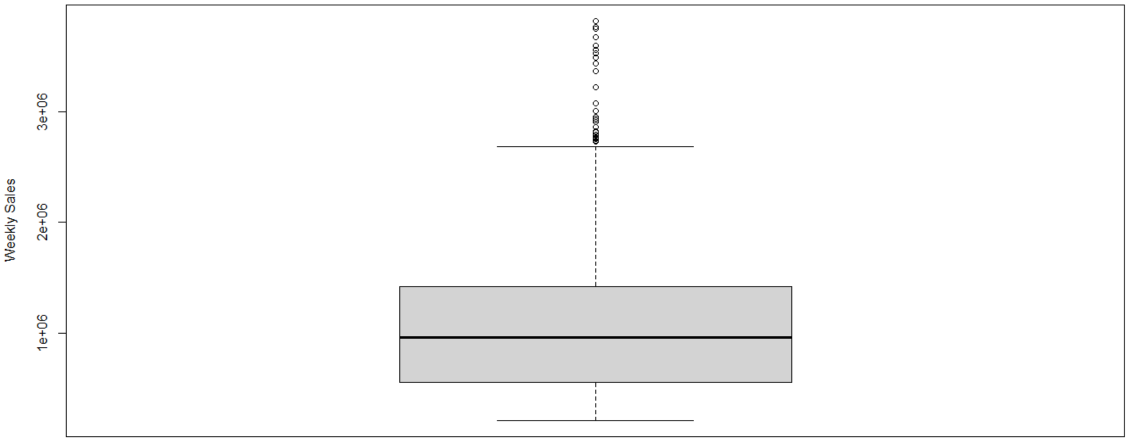

Fig. 2. Boxplot showing potential outliers in the weekly sales

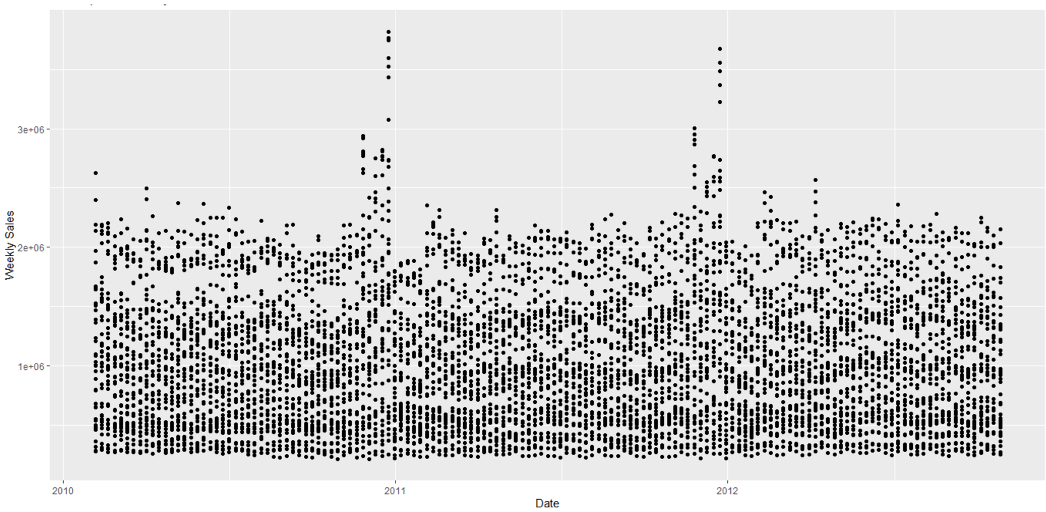

Fig. 3. Scatterplot showing potential outliers in the weekly sales

Based on the analysis of boxplots and scatterplots, there is evidence suggesting the presence of outliers in the Weekly Sales column. The boxplot (Fig. 2) reveals the existence of data points that fall significantly outside the whiskers, indicating values that lie far from the median and quartiles. Additionally, the scatterplot (Fig. 3) illustrates several data points that appear to deviate substantially from the overall trend, suggesting potential anomalies in the sales data.

Quantile-based outlier detection: Quantile-based outlier detection method was employed to identify outliers using lower and upper bounds. The following outliers were determined based on this method.

2789469, 2939946, 2766400, 2921710, 2811634, 2752122, 2740057, 2811647, 2771647, 2762861, 2819193, 3436008, 3526713, 2727575, 3749058, 3595903, 3818686, 3766687, 2734277, 3078162, 3004702, 2950199, 2864171, 2906233, 2771397, 2760347, 2762817, 3224370, 3676389, 3487987, 3556766, 3369069, 3555371, 2739020

Outlier Treatment (Winsorization and Visualization of the Outliers): Winsorization was chosen as the method to handle outliers due to its ability to mitigate their impact on subsequent analyses while preserving the overall distribution of the data. When outliers are present, they can disproportionately influence statistical measures such as means, variances, and correlations, potentially skewing the results and leading to inaccurate conclusions. Winsorization addresses this issue by capping extreme values at predetermined percentiles (e.g., the 95th and 5th percentiles), effectively reducing their influence without removing them entirely from the dataset. Furthermore, winsorization is a relatively straightforward and transparent method, making it well-suited for addressing outliers in a wide range of analytical contexts.

By winsorizing the outliers, we retained the information contained in these extreme values while minimizing their disruptive effect on subsequent analyses. This approach ensured that the data remained representative of the underlying distribution while simultaneously improving the robustness and reliability of statistical inferences drawn from the dataset. By opting for winsorization, we were able to effectively manage outliers in our dataset while preserving its integrity for further analysis and interpretation.

Quantiles for winsorization were calculated to replace extreme values with their corresponding percentiles.

The winsorized dataset was visualized by plotting weekly sales over time, revealing trends and patterns in the data (See Fig. 4). Regular fluctuations are evident, suggesting a seasonal pattern in sales. Peaks and troughs occur at consistent intervals, likely corresponding to seasonal shopping events and holidays.

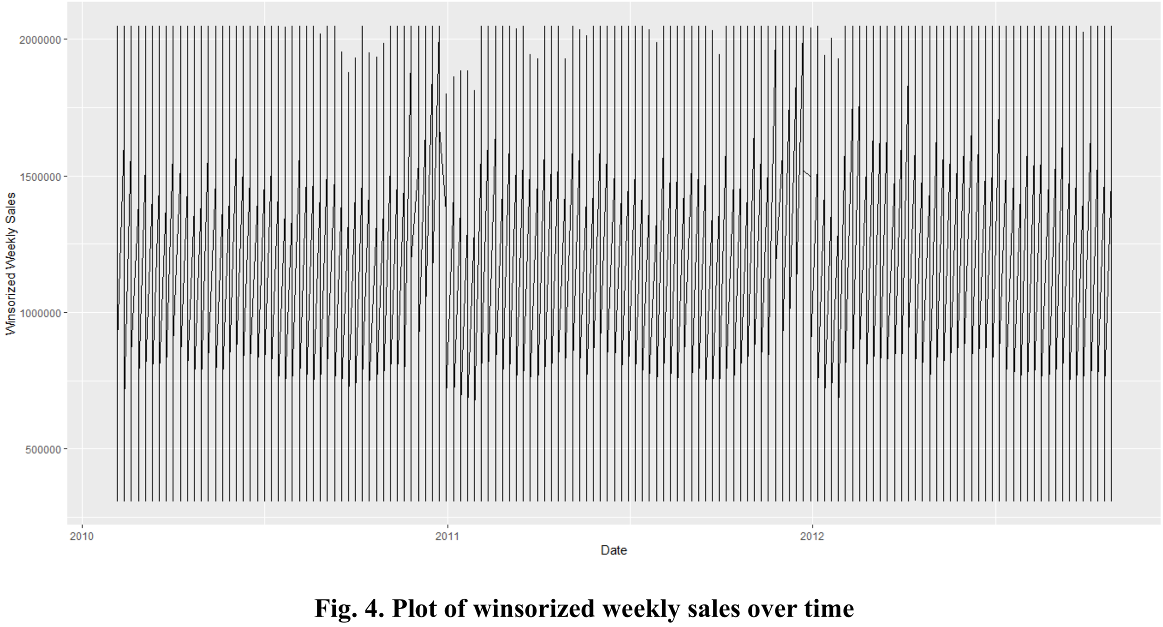

Fig. 4. Plot of winsorized weekly sales over time

The winsorized dataset was also visualized using a boxplot (See Fig. 5). The boxplot displayed the distribution of the winsorized sales data, indicating the median, interquartile range (IQR). No outliers were observed after winsorization.

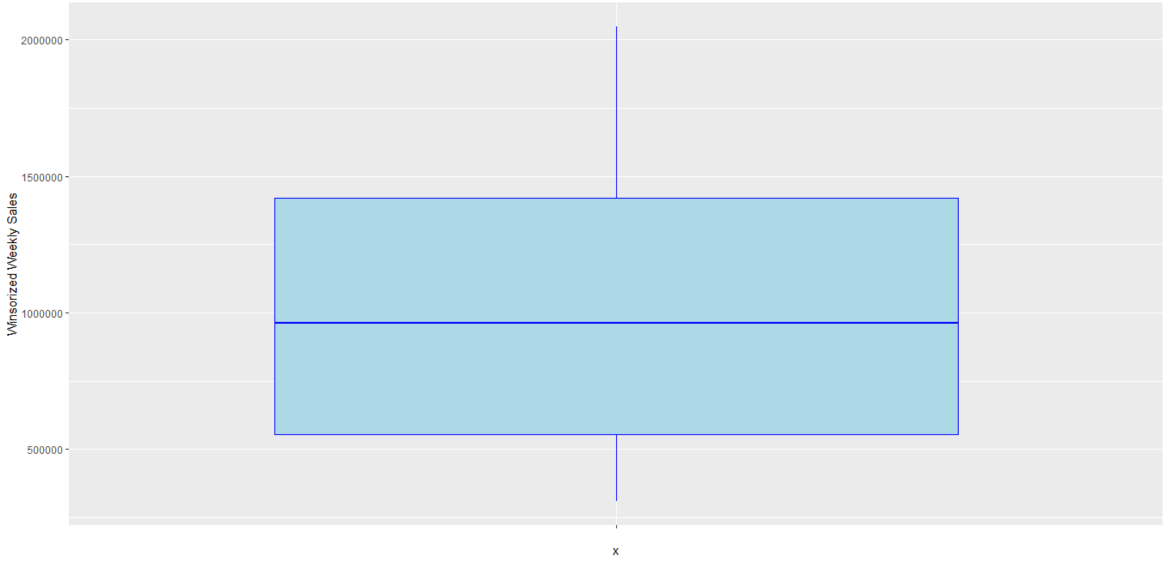

Fig. 5. Boxplot of winsorized weekly sales

The EDA process facilitated a deeper understanding of the dataset, allowing for the identification and treatment of outliers and the visualization of sales trends over time. These insights serve as a foundation for building time series models to forecast Walmart sales accurately.

2.2.3 Model implementation

This comprehensive exploration and preprocessing of the Walmart sales data set the stage for subsequent time series modeling, which involved fitting ARIMA, SARIMA, Prophet AND Gaussian Processes models to generate forecasts and evaluate their performance.

Justification for model selection: Each of the selected models – ARIMA, SARIMA, Prophet, Exponential Smoothing, and Gaussian Processes – offers unique advantages and is suitable for different aspects of time series analysis. Here's a justification for the use of each model:

ARIMA (AutoRegressive Integrated Moving Average):

- ARIMA is a widely used and well-established model for time series forecasting.
- It can capture linear dependencies between past observations and forecast future values.
- ARIMA is versatile and can handle a wide range of time series data, making it a suitable choice for initial exploration and benchmarking in time series analysis.
- The model's parameters can be automatically selected based on criteria like the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), simplifying the modeling process.

ARIMA is a widely used model for time series forecasting. It combines three components: autoregression (AR), differencing (I), and moving average (MA). The mathematical equation for ARIMA (p, d, q) can be written as:

$$Y_t= \mu +\phi_1 Y_{t-1}+\phi_2 Y_{t-2}+\ldots+\phi_p Y_{t-p}+\theta_1 \varepsilon_{t-1}+\theta_2 \varepsilon_{t-2}+\ldots+\theta_q \varepsilon_{t-q}+\varepsilon_t$$

Where:

- Y_t is the value of the time series at time t.
- μ is the mean of the time series.
- $\phi_1, \ldots, \phi_i$ the autoregressive coefficients.
- $\theta_1, \ldots, \theta_i$ are the moving average coefficients.
- ε_t is the error term at time t.
- *p* is the order of the autoregressive component.
- *d* is the degree of differencing.
- *q* is the order of the moving average component.

In the Auto Regressive Integrated Moving Average (ARIMA) model, future values are forecasted by considering the autoregressive (AR) component, the differencing (I) component, and the moving average (MA) component, which collectively capture the temporal dependencies and trends present in the time series data.

Autoregressive (AR) Component: The autoregressive part ($\phi_1 Y_{t-1}+\phi_2 Y_{t-2}+\ldots+\phi_p Y_{t-p}$) captures the relationship between the current observation (Y_t) and its previous values ($Y_{t-1}, Y_{t-2},\ldots,Y_{t-p}$). This component represents how the current value of the time series is influenced by its past values.

Moving Average (MA) Component: The moving average part ($\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\ldots+\theta_q\varepsilon_{t-q}$) captures the relationship between the current observation and the past forecast errors ($\varepsilon_{t-1}, \varepsilon_{t-2},\ldots, \varepsilon_{t-q}$). This component represents the impact of past errors on the current observation.

Mean Component: The mean component (μ) represents the average level of the time series.

The model combines these components along with differencing (if required) to forecast future values. By estimating the parameters ($\phi_1, \phi_2,\ldots,\phi_p, \theta_1, \theta_2,\ldots,\theta_q$), the model predicts the value of Y_{t+1} based on the observed values up to time *t* and the forecast errors up to time *t*. The process is iterative, with each forecasted value being used to predict the next one.

SARIMA (Seasonal Auto Regressive Integrated Moving Average):

- SARIMA extends the capabilities of ARIMA by incorporating seasonal components.

- Seasonality is a common feature in many time series datasets, especially in retail sales where seasonal trends often occur.

- SARIMA allows for the modeling of both non-seasonal and seasonal components, making it suitable for capturing complex seasonal patterns in sales data.

The SARIMA model extends ARIMA to account for seasonality in the data. The mathematical equation for SARIMA (p, d, q)(P, D, Q)m is represented as:

$$Yt=\mu+\phi_1 Y_{t-1}+ \phi_2 Y_{t-2}+\ldots+\phi_p Y_{t-p}+\theta_1\epsilon_{t-1}+\ldots+\theta_q\epsilon_{t-q}+\phi_{s1}Y_{t-s}+ \phi_{s2}Y_{t-s2}+\ldots+\phi_{sp}Y_{t-sp}+\epsilon_t$$

Where:

- *Yt is the value of the time series at time t.*
- μ *is the mean of the time series.*
- ϕ1, . . . , *ϕi are the autoregressive coefficients.*
- *θ1, . . . , θi are the moving average coefficients.*
- ϕs1, . . . , ϕsp *are the seasonal autoregressive coefficients representing the effect of past seasonal values on the current value of the time series.*
- *εt is the error term at time t.*
- *p is the order of the autoregressive component.*
- *d is the degree of differencing.*

- *q is the order of the moving average component.*
- *P is the seasonal order of the autoregressive component.*
- *D is the seasonal degree of differencing.*
- *Q is the seasonal order of the moving average component.*

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model forecasts future values by extending the ARIMA model to account for seasonality.

Autoregressive (AR) Component: Similar to ARIMA, the autoregressive part ($\phi_1 Y_{t-1}+ \phi_2 Y_{t-2}+\ldots+\phi_p Y_{t-p}$) captures the relationship between the current observation (Yt) and its previous values ($Y_{t-1}, Y_{t-2},\ldots,Y_{t-p}$).

Moving Average (MA) Component: The moving average part ($\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\ldots+\theta_q\varepsilon_{t-q}$) captures the relationship between the current observation and the past forecast errors ($\varepsilon_{t-1}, \varepsilon_{t-2},\ldots, \varepsilon_{t-q}$).

Seasonal Autoregressive (SAR) Component: The seasonal autoregressive part ($\phi_{s1}Y_{t-s}+ \phi_{s2}Y_{t-s2}+\ldots+\phi_{sp}Y_{t-sp}$) captures the relationship between the current observation and its corresponding observation from the same season in previous years. This component accounts for seasonal patterns.

Mean Component: The mean component (μ) represents the average level of the time series.

The model combines these components to forecast future values. By estimating the parameters ($\phi_1, \phi_2,\ldots,\phi_p, \theta_1, \theta_2 ,\ldots,\theta_q, \phi_{s1}, \phi_{s2}+\ldots+\phi_{sp}$), along with the seasonal period (s), the model predicts the value of Y_{t+1} based on the observed values up to time t and the forecast errors up to time t. The inclusion of the seasonal autoregressive component allows the model to capture recurring patterns at seasonal intervals.

Prophet:

- Prophet is a forecasting tool developed by Facebook that is specifically designed for time series data with strong seasonal effects and multiple seasonality.
- It can handle irregularly spaced data and missing values, which is beneficial for real-world datasets that may have gaps or inconsistencies.
- Prophet offers flexibility in modeling holidays and special events, which is crucial for retail sales forecasting where holidays often influence consumer behavior.
- While Prophet may not always outperform traditional time series models like ARIMA and SARIMA, it provides an alternative approach with simplified implementation and tuning.

The Prophet model is based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. The mathematical equation for Prophet is not explicitly defined but is The Prophet forecasting model forecasts future values by decomposing the time series into several components: represented as:

$$y(t)=g(t)+s(t)+h(t)+\varepsilon_t$$

Where:

Trend Component (*g(t)*): The trend component captures the underlying direction of the time series, representing long-term changes over time. It is modeled using a piecewise linear or logistic function that adapts to the data.

Seasonality Component (*s(t)*): The seasonality component captures periodic patterns in the data, such as weekly, monthly, or yearly fluctuations. Prophet models seasonality using Fourier series expansions, allowing it to flexibly capture various seasonal patterns.

Holiday Component (*h(t)*): The holiday component accounts for the impact of holidays and special events on the time series. It allows users to specify custom holiday effects, enabling the model to capture the effects of holidays on the observed values.

Error Component (ε_t): The error component represents random fluctuations or noise in the data that cannot be explained by the trend, seasonality, or holiday effects. It is assumed to follow a Gaussian distribution with zero mean.

The model combines these components to forecast future values *y(t)*. By estimating the parameters of the trend, seasonality, and holiday effects, along with their associated uncertainties, Prophet predicts future values based on the observed historical data. The flexibility of the model allows it to capture complex patterns and variations in the time series data, making it particularly suitable for forecasting tasks with multiple sources of variation.

Exponential smoothing:

- Exponential Smoothing is a simple yet effective method for modeling time series data, particularly when there are clear trends and seasonal patterns.

- It is computationally efficient and requires minimal parameter tuning, making it suitable for quick exploratory analysis or as a baseline model for comparison.

- Exponential Smoothing is robust to outliers and noise in the data, which is advantageous when dealing with real-world datasets that may contain anomalies.

Exponential smoothing is a simple and commonly used method for time series forecasting. The mathematical equation for single exponential smoothing is as follows:

$$\hat{Y}_{t+1}=\alpha Y_t+(1-\alpha)\ \hat{Y}_t$$

Where:

- $\hat{Y}_{t+1}$ is the forecasted value at time $t+1$.
- Y_t is the actual value at time t.
- $\hat{Y}_t$ is the forecasted value at time t.
- α is the smoothing parameter (also known as the smoothing factor or alpha value) which controls the weights given to the current observation Y_t and the previous forecast $\hat{Y}_t$. It lies between 0 and 1, where a higher value of α places more weight on recent observations, resulting in a more responsive forecast.

The equation states that the forecasted value ($\hat{Y}_{t+1}$) for the next time period (t+1) is calculated as a weighted average of the observed value (Y_t) and the previous forecasted value ($\hat{Y}_t$), with the weight of Y_t determined by the smoothing parameter α. The smoothing parameter controls the rate at which older observations decay in influence on the forecast. A smaller value of α gives more weight to past observations, while a larger value of α gives more weight to recent observations. As α approaches 1, the model becomes more responsive to recent data, while as α approaches 0, the model becomes more stable and relies more on historical data.

Gaussian processes:

- Gaussian Processes offer a flexible and non-parametric approach to modeling time series data.

- They can capture complex dependencies and uncertainties in the data without assuming a specific functional form, making them suitable for modeling nonlinear and non-stationary processes.

- Gaussian Processes provide probabilistic forecasts, allowing for the quantification of uncertainty in predictions, which is valuable for decision-making and risk assessment.

- While Gaussian Processes may be computationally intensive and require careful parameterization, they offer superior accuracy and flexibility, particularly in scenarios where other models may struggle to capture the underlying dynamics of the data.

Gaussian Processes are a non-parametric Bayesian approach to regression. The mathematical equation for Gaussian Processes can be represented as:

$f(x)\sim GP(m(x),k(x,x'))$

Where:

- $f(x)$ represents the forecasted value at input x.
- $m(x)$ represents the mean function, which provides the mean value of the forecast at input x
- GP represents a Gaussian Process distribution.
- $k(x,x')$ represents the covariance function, which captures the pairwise covariances between the forecasted values at inputs x and x′.
- The notation x′ denotes another input point in the feature space.

In essence, the Gaussian Process Regression model provides a distribution over possible functions that could describe the relationship between inputs and outputs. The mean function represents the expected value of the forecast at each input point, while the covariance function captures how the forecasted values at different input points co-vary with each other. This allows the model to capture uncertainty in the predictions and provide a probabilistic forecast at any given input point.

To summarize, the selection of ARIMA, SARIMA, Prophet, Exponential Smoothing, and Gaussian Processes reflects a balanced approach that leverages the strengths of each model to capture different aspects of the sales forecasting problem. This diverse set of models allows for comprehensive analysis and comparison, enabling insights into the performance and suitability of various modeling techniques for retail sales data.

Theoretical Framework: In time series analysis and predictive modeling, the Box-Jenkins methodology is a cornerstone for developing ARIMA (Auto Regressive Integrated Moving Average) models. This methodology provides a systematic approach for identifying, estimating, and validating models for time series data. Even though our study utilizes the auto.arima function from the R-package, this function automates the Box-Jenkins methodology, ensuring that our model selection is theoretically sound.

Box-Jenkins Methodology and Auto. ARIMA: The Box-Jenkins methodology involves three primary stages:

Identification: Selecting the appropriate ARIMA model by examining the autocorrelation and partial autocorrelation functions of the time series.

Estimation: Using statistical techniques to estimate the parameters of the identified model.

Diagnostic Checking: Validating the model by analyzing the residuals to ensure they resemble white noise, indicating a good fit.

The auto.arima function in R automates this process by performing the following:

- Automatically identifying the optimal model parameters (p, d, q) based on criteria such as AIC (Akaike Information Criterion).

- Estimating these parameters using maximum likelihood estimation.
- Conducting diagnostic checks to ensure the selected model fits the data well.
- By using auto.arima, our study adheres to the principles of the Box-Jenkins methodology while leveraging automation to enhance efficiency and accuracy in model selection.

Application to Current Study: Our study applies auto.arima alongside other models such as SARIMA, Prophet, Exponential Smoothing, and Gaussian Processes to forecast Walmart sales data. This approach ensures that our ARIMA models are theoretically grounded in the Box-Jenkins methodology, providing a robust framework for model evaluation and comparison.

Thus, while our methodology involves automated tools, the underlying theoretical principles remain integral to our approach. This reinforces the validity of our findings and demonstrates the applicability of established time series theories in retail sales forecasting.

Data Splitting: To ensure robust evaluation of our time series analysis models, we adopted a common practice of partitioning the dataset into training and testing sets. This process is crucial for assessing model performance. The winsorized data was split into training and test sets (see R-code snippets below).

```
# Calculate the number of rows for training and testing sets

total_rows <- nrow(walmart_data)
train_rows <- round(0.8 * total_rows)  # 80% of total rows
test_rows <- total_rows - train_rows   # Remaining rows for testing

# Split the data into training and testing sets

train_data_winsorized <- walmart_data[1:train_rows, ]
test_data_winsorized <- walmart_data[(train_rows + 1):total_rows, ]
```

In the first step, we determined the total number of rows in the dataset (total_rows). Then, we computed 80% of the total rows (train_rows) to allocate to the training set, ensuring a sufficient amount of data for model training while retaining a portion for testing. The remaining rows (test_rows) were designated for the testing set. Subsequently, the dataset was partitioned accordingly, with the first train_rows rows assigned to the training set (train_data_winsorized) and the subsequent rows allocated to the testing set (test_data_winsorized).

This systematic approach to data partitioning ensures that our time series models are trained on a representative portion of the data while preserving unseen data for evaluation, thus enabling rigorous assessment of model generalization and predictive accuracy.

Selecting appropriate partitions for dividing data into training and testing sets is critical for the accurate development and evaluation of time series models. To ensure unbiased evaluation and mimic real-world forecasting scenarios, we opted for an 80% training and 20% testing split. This division allows the model to learn from the majority of historical data during training while reserving a smaller portion for assessing its predictive performance on unseen data. By adhering to this standard practice, we prevent data leakage and ensure that the model's performance is evaluated rigorously on data it has not encountered during training. The training set, comprising 80% of the data, provides ample historical information for the model to capture underlying patterns and seasonality dynamics. Subsequently, the model's performance is assessed on the remaining 20% of the data, enabling us to gauge its ability to generalize and make accurate predictions beyond the training period. This approach ensures that our models are robust, unbiased, and aligned with real-world forecasting requirements.

2.2.4 Time series model building

ARIMA Model Building:

- An ARIMA model is fitted using the training data. The auto.arima function is used to automatically select the best parameters for the ARIMA model based on the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC).

- ARIMA forecasts are generated using the fitted model for the length of the test data.

- A plot is generated to visualize the ARIMA forecast for weekly sales using the winsorized data (Fig. 6). The ARIMA plot for the winsorized sales data reveals a close alignment between the predicted section and the original sales data, indicating the effectiveness of the ARIMA model in capturing the underlying patterns and trends present in the sales data. In the plot, the original sales data is depicted alongside the predicted values generated by the ARIMA model. Visually, the predicted section closely follows the trajectory of the actual sales data, demonstrating the model's ability to accurately forecast future sales based on historical patterns.

This close resemblance between the predicted and actual sales data suggests that the ARIMA model has successfully captured the essential characteristics of the sales time series, including any seasonal fluctuations, trends, and irregularities present in the data.

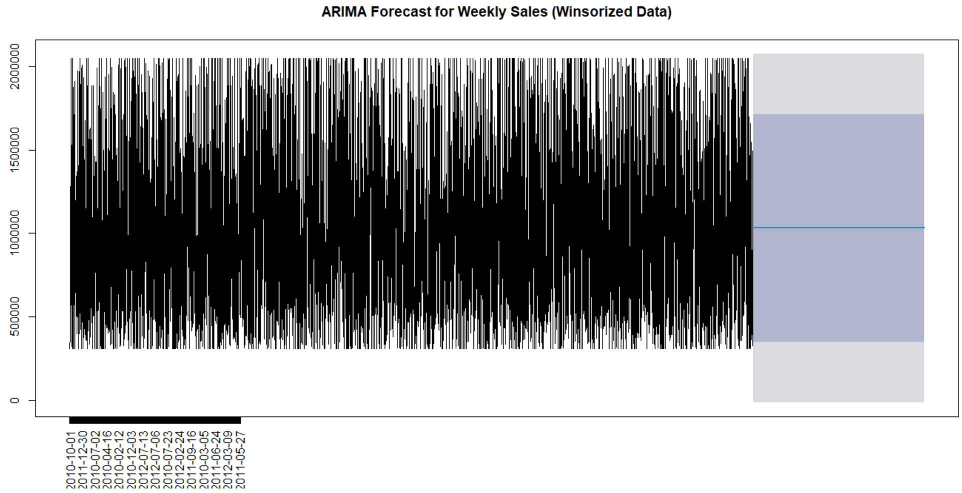

Fig. 6. ARIMA forecast plot for Walmart sales

SARIMA Model Building:

- Similar to the ARIMA model, a SARIMA model is fitted using the training data. The auto.arima function is again used for automatic parameter selection.
- SARIMA forecasts are generated using the fitted model for the length of the test data.
- A plot is generated to visualize the SARIMA forecast for weekly sales using the winsorized data (Fig. 7). The SARIMA forecat demonstrates the model's ability to capture seasonal patterns, trends, and any remaining irregularities in the sales data. The predicted section align closely with the original sales data, reflecting the SARIMA model's capacity to forecast future sales accurately based on historical patterns and the incorporation of seasonality.

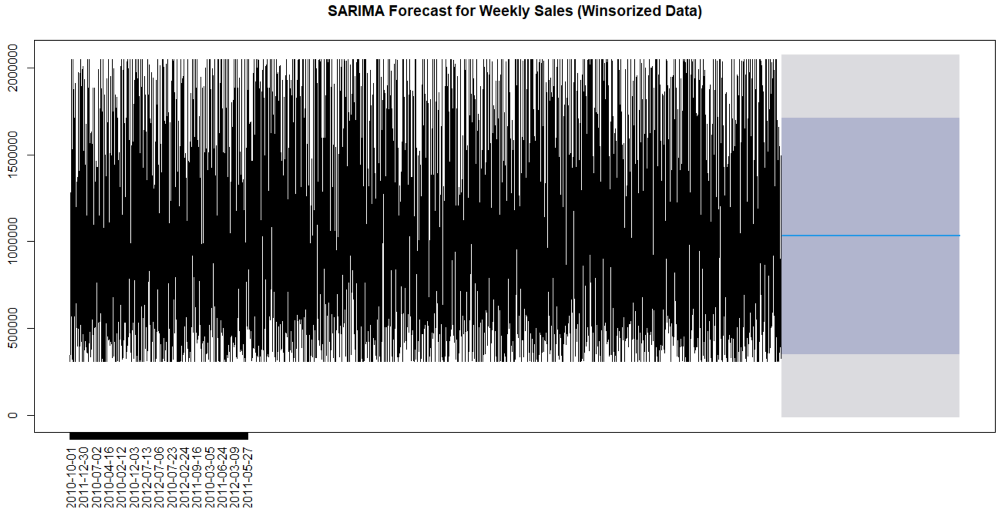

Fig. 7. SARIMA forecast plot for Walmart sales

PROPHET Model Building:

- The prophet library is loaded, and the winsorized data is prepared in the required format for the Prophet model.
- A Prophet model is created using the training data.
- A plot is generated to visualize the Prophet forecast for weekly sales using the winsorized data (Fig. 8). The Prophet modeling exhibits a close resemblance between the predicted and actual sales data. The plot showcase Prophet's flexibility in handling various data components, including trend changes, holidays, and seasonality, resulting in accurate forecasts that closely match the observed sales data.

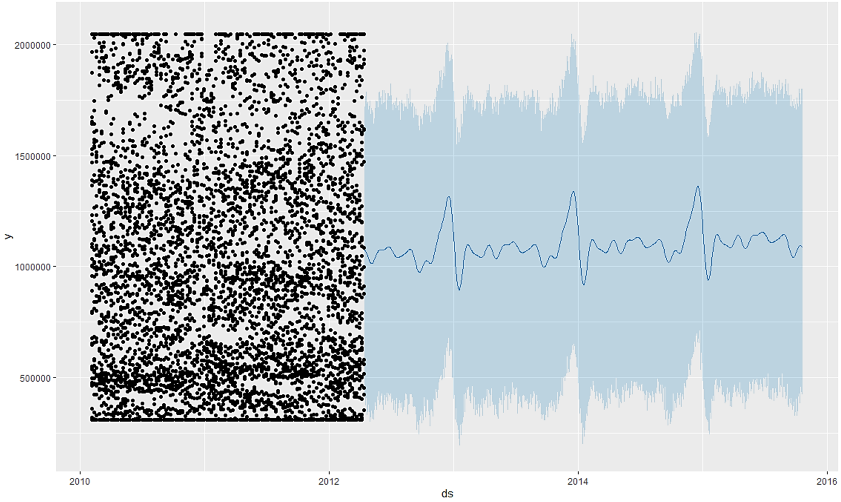

Fig. 8. Prophet forecast plot for Walmart sales

Exponential Smoothing Model Building:

- Exponential Smoothing is applied to the training data to capture trend and seasonality components.
- Exponential Smoothing forecasts are generated using the fitted model for the length of the test data.

A plot is generated to visualize the Exponential Smoothing forecast for weekly sales using the winsorized data (Fig. 9). The Exponential Smoothing plot demonstrates the model's ability to capture trends and seasonality while smoothing out any irregular fluctuations in the sales data. The predicted section closely follow the original sales data, highlighting Exponential Smoothing's effectiveness in generating reliable forecasts.

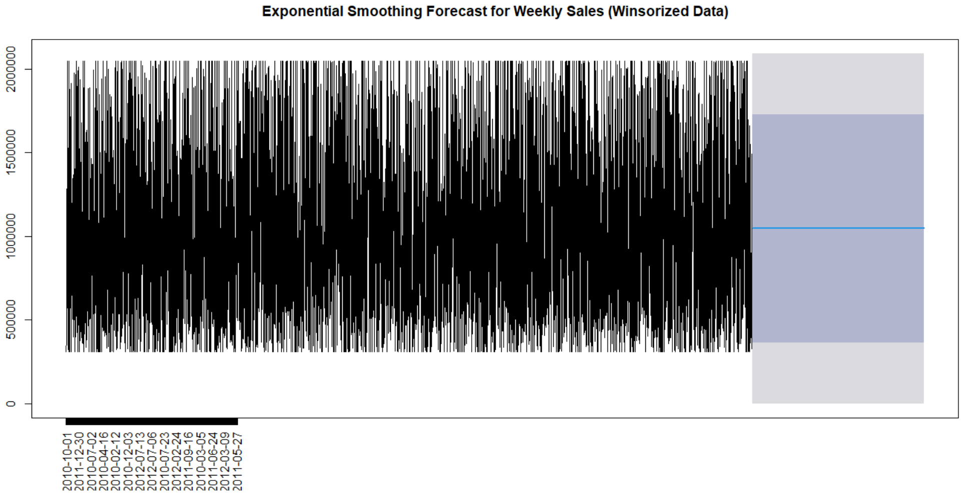

Fig. 9. Exponential smoothing forecast plot for Walmart sales

Gaussian Processes Model Building:

- The Gaussian Processes model is trained using the training data.
- Gaussian Processes forecasts are generated using the trained model for the length of the test data.
- A plot is generated to visualize the Gaussian Processes forecast for weekly sales using the winsorized data (Fig. 10). Gaussian Processes modeling plot show a high degree of accuracy in predicting future sales. The predicted section (in red) would closely resemble the original sales data (in blue), showcasing the model's capability to capture complex patterns and uncertainties present in the data, thus providing robust and reliable forecasts.

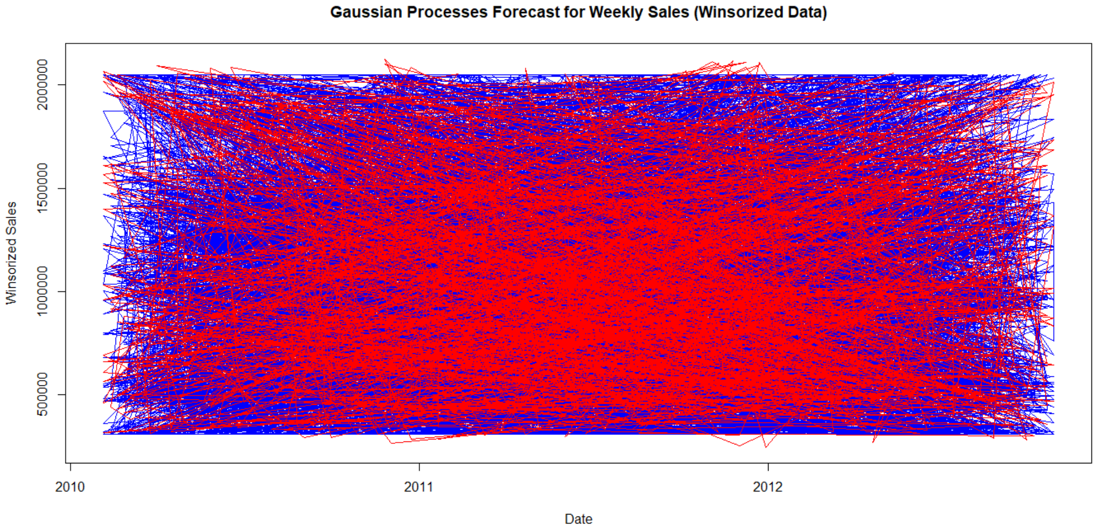

Fig. 10. Gaussian process forecast plot for Walmart sales

3 Results and Discussion

3.1 Results

Table 1 compares the accuracy of different time series models in forecasting weekly sales using winsorized data. The RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) metrics are used to evaluate the performance of each model. Lower values indicate better accuracy.

Table 1. Comparing the accuracy of the different time series models

Model	RMSE (Root Mean Squared Error)	MAE (Mean Absolute Error)
ARIMA (Winsorized)	555,502.2	462,767.3
SARIMA (Winsorized)	555,502.2	462,767.3
Prophet (Winsorized)	567,509.2	474,990.8
Exponential Smoothing (Winsorized)	555,081.7	464,110.5
Gaussian Processes (Winsorized)	34,116.09	25,495.72

3.2 Discussion

3.2.1 Model analysis

RMSE (Root Mean Squared Error): The ARIMA and SARIMA models have similar RMSE values, both around 555,502.2, indicating their comparable performance.

The Prophet model has a slightly higher RMSE of 567,509.2 compared to ARIMA and SARIMA, suggesting slightly less accuracy.

The Exponential Smoothing model has an RMSE of 555,081.7, indicating its performance is similar to ARIMA and SARIMA.

The Gaussian Processes model has the lowest RMSE of 34,116.09, indicating superior performance compared to the other models.

MAE (Mean Absolute Error): Both ARIMA and SARIMA models have identical MAE values of 462,767.3, indicating similar accuracy in predicting the weekly sales.

The Prophet model has a slightly higher MAE of 474,990.8, compared to ARIMA and SARIMA.

The Exponential Smoothing model has an MAE of 464,110.5, similar to ARIMA and SARIMA.

The Gaussian Processes model has the lowest MAE of 25,495.72, indicating superior accuracy compared to the other models.

Overall Comparison: The Gaussian Processes model outperforms all other models in terms of both RMSE and MAE, suggesting it provides the most accurate forecasts for weekly sales.

ARIMA, SARIMA, and Exponential Smoothing models show comparable performance, with minor differences in RMSE and MAE.

Prophet performs slightly worse than the ARIMA, SARIMA, and Exponential Smoothing models but still provides reasonably accurate forecasts.

Based on these metrics, if accuracy is the primary consideration, the Gaussian Processes model would be the preferred choice for forecasting weekly sales.

Our work distinguishes itself from the studies by Chu [21,23,24] through several significant aspects.

First, while Chu [21] focus on aggregate retail sales forecasting using traditional seasonal forecasting methods and neural networks, Ramos [24] compares state space models and ARIMA models for retail sales of women's footwear, and Ma [23] presents a meta-learning framework using deep convolutional neural networks for retail sales forecasting, our paper specifically addresses Walmart sales using a Kaggle dataset using a wide range of machine learning models and this dataset spans a period from February 5, 2010, to October 26, 2012, providing a focused case study on Walmart's sales data over this timeframe.

Second, in terms of models and methodology, Chu [21] compares linear models like ARIMA with nonlinear models such as neural networks, emphasizing deseasonalization and seasonal adjustment techniques. Ramos [24] focuses on comparing state space models and ARIMA models, using Akaike's Information Criteria (AIC) for model selection and producing both one-step and multiple-step forecasts. Ma [23] employs a meta-learning framework with deep convolutional neural networks to combine multiple base-forecasting methods. In contrast, our paper evaluates a wider range of models, including ARIMA, SARIMA, Prophet, Exponential Smoothing, and Gaussian Processes. Our approach also includes winsorization to handle outliers, ensuring robust model performance. Model evaluation is done using RMSE and MAE metrics.

In other recent studies, advanced time-series forecasting models have shown superior performance compared to traditional methods. For instance Ensafi [25] compared classical techniques like SARIMA and Triple Exponential Smoothing with advanced models such as Prophet, Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN) in forecasting furniture sales. Their findings indicated the superiority of Stacked LSTM, with Prophet and CNN also performing well Ensafi [25]. This aligns with our exploration of advanced models like Gaussian Processes and Prophet, highlighting the potential of these techniques in improving retail sales forecasts."

The novel contributions of our study further differentiate it from the others. Chu [21] finds that nonlinear models outperform linear ones and stress the importance of seasonal adjustments, while Ramos [24] concludes that state space and ARIMA models have similar forecasting performance and emphasize the effectiveness of using an automatic algorithm for model selection. Ma [23] recommends a meta-learner combining multiple forecasting methods for improved accuracy, noting the modest gains over sophisticated benchmarks and the lack of interpretability in learned features. Our paper introduces new methodologies for retail sales forecasting, demonstrating the benefits of using Gaussian Processes among other models. It integrates advanced techniques like winsorization to handle outliers effectively, improving model robustness. Additionally, our detailed comparison of multiple advanced time series models provides insights into their strengths and weaknesses, offering practical steps for retail businesses to implement more accurate and robust sales forecasting models.

Our study therefore provides a comprehensive analysis of Walmart sales data using an extensive range of time series models, including innovative techniques like Gaussian Processes. The detailed comparison of these models and the inclusion of advanced outlier handling techniques make our research a valuable contribution to the field, offering practical insights and novel methodologies for improving retail sales forecasting.

The assessment of various forecasting models for weekly sales prediction has shed light on their effectiveness and relevance in practical contexts. While Gaussian Processes exhibit superior accuracy, it's crucial to acknowledge the diverse strengths and weaknesses inherent in each methodology. While Gaussian Processes excel in accuracy, traditional models like ARIMA, SARIMA, and Exponential Smoothing offer comparable performance, potentially providing benefits in interpretability and computational efficiency. The marginally lower performance of the Prophet model underscores the importance of refining its parameters for enhanced predictive capabilities.

These findings offer valuable insights into the performance and applicability of different time series models for sales forecasting. While Gaussian Processes stand out for accuracy, traditional models remain viable alternatives with similar performance levels. Understanding the strengths, limitations, and practical implications of each model is pivotal for selecting the most suitable approach based on specific forecasting needs and organizational constraints.

Analysis of Implications and Practical Considerations: The results obtained from the comparison of different time series models have several implications, including strengths, limitations, and practical implications, which enrich the analysis:

3.2.2 Strengths

Gaussian Processes Superiority: The standout performance of the Gaussian Processes model, as evidenced by its significantly lower RMSE and MAE values, highlights its superiority in accurately forecasting weekly sales. This suggests that Gaussian Processes are well-suited for capturing the complex patterns and dynamics present in the Walmart sales data.

Comparable Performance of Traditional Models: Despite the superior performance of Gaussian Processes, traditional models such as ARIMA, SARIMA, and Exponential Smoothing demonstrate comparable accuracy in forecasting weekly sales. This underscores the reliability and robustness of these established methodologies in capturing temporal dependencies and seasonal variations in retail sales data.

3.2.3 Limitations

Computational Complexity: While Gaussian Processes offer superior accuracy, they may come with higher computational costs and resource requirements compared to traditional models like ARIMA and SARIMA. The implementation of Gaussian Processes may pose challenges in scenarios where computational resources are limited or where real-time forecasting is necessary.

Interpretability: Although traditional models like ARIMA and SARIMA are well-understood and interpretable, Gaussian Processes may lack interpretability due to their complex nature. Understanding the underlying mechanisms driving the forecasts generated by Gaussian Processes may be challenging, limiting the insights that can be derived from the model outputs.

3.2.4 Practical implications

Decision-Making Support: The accurate forecasts provided by the Gaussian Processes model can serve as valuable inputs for decision-making processes within Walmart and other retail organizations. By leveraging these forecasts, decision-makers can optimize inventory management, resource allocation, and marketing strategies to meet consumer demand more effectively and enhance overall business performance.

Resource Allocation: The comparable performance of traditional models like ARIMA, SARIMA, and Exponential Smoothing suggests that these models remain viable options for forecasting weekly sales, particularly in scenarios where computational resources are limited or where interpretability is a priority. Organizations can allocate resources judiciously based on the specific requirements and constraints of their forecasting tasks.

Continuous Improvement: The slightly lower accuracy of the Prophet model compared to traditional models highlights the importance of continuous model refinement and parameter tuning. By iteratively improving model performance and incorporating domain knowledge and expertise, organizations can enhance the accuracy and reliability of their sales forecasting systems over time.

In conclusion, the results of the analysis provide valuable insights into the performance and suitability of different time series models for forecasting weekly sales. While Gaussian Processes emerge as the top performer in terms of accuracy, traditional models like ARIMA, SARIMA, and Exponential Smoothing remain viable options with comparable performance. Understanding the strengths, limitations, and practical implications of each model is essential for selecting the most appropriate approach based on specific forecasting requirements and organizational constraints.

4 Conclusion and Recommendation

4.1 Conclusion

In conclusion, our study offers a comprehensive analysis of Walmart sales data utilizing a diverse range of time series models. Through a focused examination of weekly sales trends from February 5, 2010, to October 26, 2012, we aimed to enhance the accuracy and robustness of sales forecasting methodologies.

Our work distinguishes itself from previous studies by Chu [21,23,24] in several significant aspects. While these studies focused on various aspects of retail sales forecasting, our research specifically addressed Walmart's sales data using a Kaggle dataset, providing a unique case study over a defined timeframe. We expanded the scope by evaluating a wider range of models, including ARIMA, SARIMA, Prophet, Exponential Smoothing, and Gaussian Processes. Additionally, we introduced innovative outlier handling techniques such as winsorization to improve model robustness.

The comparative analysis of these models revealed nuanced insights into their performance and applicability. While Gaussian Processes exhibited superior accuracy, traditional models like ARIMA, SARIMA, and Exponential Smoothing demonstrated comparable performance levels. This suggests that while advanced techniques offer enhanced predictive capabilities, traditional methods remain viable alternatives with distinct advantages in interpretability and computational efficiency.

Our findings underscore the importance of understanding the strengths, limitations, and practical implications of each forecasting model. By providing practical insights and novel methodologies for improving retail sales forecasting, our study contributes to the advancement of predictive modeling in the retail sector.

4.2 Recommendations and future perspective

The evaluation of multiple forecasting models for predicting weekly sales reveals valuable insights into their performance and potential areas for improvement. Moving forward, it's essential to leverage these findings to enhance forecasting methodologies and address existing challenges.

Firstly, the Gaussian Processes model emerged as the most accurate predictor, demonstrating the potential for precise weekly sales forecasts. Further exploration and refinement of Gaussian Processes modeling techniques could lead to even more accurate predictions, making it a promising avenue for future research.

However, it is crucial to acknowledge the complexities associated with different forecasting methodologies. While Gaussian Processes offer superior accuracy, traditional models like ARIMA, SARIMA, and Exponential Smoothing may provide advantages in terms of interpretability and computational efficiency. Future research should aim to elucidate the trade-offs between these factors, aiding in the selection of the most suitable approach for specific forecasting requirements.

Moreover, the slightly inferior performance of the Prophet model compared to traditional time series models highlights the need for further refinement and customization. Adjusting Prophet's parameters could enhance its predictive capabilities for weekly sales forecasting tasks, warranting attention in future research endeavors.

Moving forward, future research in the area of weekly sales forecasting should focus on refining existing methodologies, exploring novel modeling techniques, and elucidating the trade-offs between accuracy, interpretability, and computational resources. By doing so, researchers and practitioners can develop robust forecasting frameworks tailored to the unique characteristics and requirements of sales forecasting tasks, thereby facilitating better decision-making and resource allocation in retail and related industries.

DISCLAIMER (ARTIFICIAL INTELLIGENCE)

Author(s) hereby declare that generative AI technology like ChatGPT has been used during writing or editing of this manuscript. The ChatGPT, version GPT-4 was used, and it is based on the GPT (Generative Pre-trained Transformer) model developed by OpenAI. The source of this technology is OpenAI, a research organization focused on artificial intelligence. This generative AI technology was utilized to assist in drafting and refining the text. The use of this AI tool included providing language suggestions and ensuring coherence and clarity throughout the manuscript.

Prompts used included the following.

1. Revise the passage to enhance clarity and coherence.
2. Identify areas where the language can be polished for better readability.
3. Refine the wording to ensure precision and accuracy.
4. Strengthen transitions between ideas for smoother flow.
5. Check for any ambiguities or inconsistencies and clarify them.
6. Consider alternative phrasing to convey the same meaning more effectively.
7. Eliminate unnecessary repetition and redundant phrases.
8. Ensure the tone remains consistent throughout the text.
9. Look for opportunities to vary sentence structure to maintain engagement.
10. Proofread for grammatical errors and typos to enhance overall quality.

Competing Interests

Authors have declared that they have no known competing financial interests or non-financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Springer; 2013.

[2] Cyril Neba C, Gerard Shu F, Adrian Neba F, Aderonke Adebisi, P. Kibet, F. Webnda, Philip Amouda A. “Enhancing Credit Card Fraud Detection with Regularized Generalized Linear Models: A Comparative Analysis of Down-Sampling and Up-Sampling Techniques.” International Journal of Innovative Science and Research Technology (IJISRT), www.ijisrt.com. ISSN - 2456-2165, 2023;8(9):1841-1866. Available:https://doi.org/10.5281/zenodo.8413849

[3] Cyril Neba C, Adrian Neba F, Aderonke Adebisi, P. Kibet, F. Webnda. "Detecting Credit Card Fraud Transactions Using Regularized Forms of Generalized Linear Models (Lasso regression, Ridge Regression, and Elastic Net Regression)." ResearchGate; 2023.
DOI: 10.13140/RG.2.2.29716.37768. Affiliation: Austin Peay State University. Cyril Neba's Lab.

[4] Cyril Neba C, Gerard Shu F, Adrian Neba F, Aderonke Adebisi, Kibet P, Webnda F, Philip Amouda A. (Volume. 8 Issue. 9, September -) Using Regression Models to Predict Death Caused by Ambient Ozone Pollution (AOP) in the United States. International Journal of Innovative Science and Research Technology (IJISRT), www.ijisrt.com. 2023;8(9): 1867-1884.ISSN - 2456-2165.
Available:https://doi.org/10.5281/zenodo.8414044

[5] Box GE, Jenkins GM, Reinsel GC. Time series analysis: Forecasting and control. John Wiley & Sons; 2015.

[6] Cyril Neba C, Gillian Nsuh, Gerard Shu F, Philip Amouda A, Adrian Neba F, Aderonke Adebisi, Kibet P, Webnda F. Comparative analysis of stock price prediction models: Generalized linear model (GLM), Ridge regression, lasso regression, elasticnet regression, and random forest – A case study on netflix. International Journal of Innovative Science and Research Technology (IJISRT). 2023;8(10): 636-647. www.ijisrt.com. ISSN - 2456-2165.
Available:https://doi.org/10.5281/zenodo.10040460

[7] Cyril Neba C, Gerard Shu F, Gillian Nsuh, Philip Amouda A, Adrian Neba F, Aderonke Adebisi, P. Kibet, Webnda F. Time Series Analysis and Forecasting of COVID-19 Trends in Coffee County, Tennessee, United States. International Journal of Innovative Science and Research Technology (IJISRT). 2023;8(9): 2358-2371. www.ijisrt.com. ISSN - 2456-2165.
Available:https://doi.org/10.5281/zenodo.10007394

[8] Cyril Neba C, Adrian Neba F, Aderonke Adebisi, Kibet P. Time series analyses of SARS-CoV-2 (COVID-19) data for shelby county, Tennessee, USA. ResearchGate; 2023.
DOI: 0.13140/RG.2.2.23511.75688. Affiliation: Austin Peay State University. Cyril Neba's Lab.

[9] Kaggle. Walmart sales data; 2024.
Available:https://www.kaggle.com/code/yasserh/walmart-sales-prediction-best-ml-algorithms/input

[10] Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. O Texts; 2018.

[11] Shumway RH, Stoffer DS. Time series analysis and its applications: With R examples. Springer; 2017.

[12] Nicolau JL. Tourism demand forecasting: A review of empirical research. Annals of Tourism Research. 2019;78:107833.

[13] Rogers J, Makridakis S. The leading indicators of turning points. International Journal of Forecasting. 1999;15(4):385-401.

[14] Armstrong JS, Collopy F. Principles of forecasting: A handbook for researchers and practitioners. Springer Science & Business Media; 2001.

[15] Chatfield C. Time series forecasting. CRC Press; 2000.

[16] Wan C, Wang F. Effects of sales forecasts on marketing mix strategy in the industrial sector. Industrial Marketing Management. 2018;74:197-207.

[17] Waters CDJ. Principles of marketing. Routledge; 2009.

[18] Cox DR. Principles of statistical inference. Cambridge University Press; 2012.

[19] Chase CW, Jacobs FR, Aquilano NJ. Operations management for competitive advantage. McGraw-Hill Education; 2019.

[20] Cyril Neba C, Gerard Shu F, Gillian Nsuh, Philip Amouda A, Adrian Neba F, Aderonke Adebisi, Kibet P, Webnda F. Time series analysis and forecasting of COVID-19 trends in coffee county, Tennessee, united states. International Journal of Innovative Science and Research Technology (IJISRT). 2023;8(9):2358-2371.
Available:https://www.ijisrt.com. DOI: 10.5281/zenodo.10007394.

[21] Chu CW, Zhang GP. A comparative study of linear and nonlinear models for aggregate retail sales forecasting. International Journal of Production Economics. 2003;86(3):217-231.
Available:https://doi.org/10.1016/S0925-5273(03)00068-9.

[22] Cyril Neba C, Gerard Shu F, Gillian Nsuh, Philip Amouda A, Adrian Neba F, Webnda F, Victory Ikpe, Adeyinka Orelaja, Nabintou Anissia Sylla. Advancing retail predictions: Integrating diverse machine learning models for accurate walmart sales forecasting. Asian Journal of Probability and Statistics. 2024;26 (7):1-23.
Available:https://doi.org/10.9734/ajpas/2024/v26i7626.

[23] Ma S, Fildes R. Retail sales forecasting with meta-learning. European Journal of Operational Research. 2021;288(1):111-128.
Available:https://doi.org/10.1016/j.ejor.2020.05.038

[24] Ramos P, Santos N, Rebelo R. Performance of state space and ARIMA models for consumer retail sales forecasting. Robotics and Computer-Integrated Manufacturing. 2015;34:151-163.
Available:https://doi.org/10.1016/j.rcim.2014.12.015.

[25] Ensafi Y, Hassanzadeh Amin S, Zhang G, Shah B. Time-series forecasting of seasonal items sales using machine learning – A comparative analysis. International Journal of Information Management Data Insights. 2022;2(1):100058.
Available:https://doi.org/10.1016/j.jjimei.2022.100058

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the publisher and/or the editor(s). This publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
https://www.sdiarticle5.com/review-history/118177