Original Research Article

A Comprehensive Study of Sales Predictions Using Time Series Analysis: Case Study- Walmart Sales

Abstract:

This article presents a comprehensive study of sales predictions using time series analysis, focusing on a case study of Walmart sales data. Leveraging a dataset from Kaggle comprising weekly sales data from various Walmart stores along with additional features such as holiday flags, temperature, fuel prices, Consumer Price Index (CPI), and unemployment rates, this study explores the effectiveness of time series analysis in forecasting future sales trends. Various time series analysis techniques including AutoRegressive Integrated Moving Average (ARIMA), Seasonal AutoRegressive Integrated Moving Average (SARIMA), Prophet, Exponential Smoothing, and Gaussian Processes are applied to model and forecast Walmart sales data. 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.

Commented [A1]: Indicate the aim of the study in the abstract

Keywords: Sales prediction, Time series analysis, Walmart, ARIMA, SARIMA, Prophet, Exponential Smoothing, Gaussian Processes, Forecasting, Model evaluation.

Commented [A2]: Use the following outlines:
1.Introduction
2. Materials and Methods
3. Results and Discussion
4. Conclusion and Recommendations

1. Introduction

In many fields, predictive modeling has become an effective tool for predicting future trends and helping decision-makers make data-driven choices [1]. In particular, time series analysis is essential for forecasting future values based on previously collected data points arranged chronologically [2]. 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.

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 [3]. 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 [4]. It includes a range of approaches, including autocorrelation analysis, decomposition, smoothing techniques, and forecasting strategies like seasonal ARIMA and ARIMA (AutoRegressive Integrated Moving Average) [6]. 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.

2. Background

Commented [A3]: This must be part of the introduction.
Please merge it with the introduction.

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 [4]. 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 [5] 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 [6], 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.

Importance of Time Series Analysis in Sales Forecasting

Commented [A4]: This must be a subsection of the introduction.

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 to identify trends, seasonal variations, and cyclical patterns in sales data [4]. These insights enable businesses to anticipate demand fluctuations and adjust their strategies accordingly [7].

Forecasting Accuracy:

By analyzing historical sales data using time series techniques, businesses can develop accurate forecasting models [6]. These models take into account past sales performance and external factors, leading to more reliable predictions [8].

Resource Allocation and Inventory Management:

Sales forecasts derived from time series analysis help businesses allocate resources efficiently and manage inventory levels effectively [9]. This optimization minimizes stockouts and excess inventory, enhancing operational efficiency [10].

Marketing and Promotion Strategies:

Understanding sales patterns over time enables businesses to optimize their marketing and promotion strategies [11]. Time series analysis helps identify peak sales periods and consumer behavior trends, guiding targeted marketing efforts [12].

Budgeting and Financial Planning:

Accurate sales forecasts derived from time series analysis facilitate budgeting and financial planning processes [13]. Businesses can use these forecasts to set realistic revenue targets, allocate budgets effectively, and make informed investment decisions [14].

Commented [A5]: These must serve as transition tools (sentences) in the methods subsection

3. Dataset and Variable Definitions

Commented [A6]: This must be part of the subsection of Materials.

The retail industry is characterized by dynamic sales patterns influenced by various internal and external factors. Understanding these dynamics is crucial for optimizing business strategies and enhancing performance. In this paper, we provide an overview of a comprehensive dataset obtained from Walmart, covering a period from February 5, 2010, to October 26, 2012. This dataset offers valuable insights into Walmart's sales dynamics, allowing for in-depth analysis of sales trends, seasonal variations, and the impact of external factors on sales performance.

Dataset Description:

The dataset comprises weekly sales data from Walmart stores, with each observation containing information on store-specific sales, holiday flags, and additional factors such as temperature, fuel price, Consumer Price Index (CPI), and unemployment rate. The inclusion of store-level data enables the exploration of sales performance across different store locations, while the temporal dimension provided by the date variable facilitates the analysis of sales trends over time.

Key Variables:

Date: The date variable serves as the time index for our time series analysis. We can use it to organize our data into chronological order and identify patterns and trends over time.

Weekly_Sales: This is our target variable, representing the sales we want to forecast. We'll use historical weekly sales data to train our time series models and predict future sales.

Holiday_Flag: Incorporating this binary variable into our model allows us to account for the impact of holidays on sales. We can create dummy variables for holiday weeks and non-holiday weeks to capture any seasonal effects on sales.

Temperature: Temperature can influence consumer behavior and purchasing patterns, especially for seasonal products like clothing or outdoor equipment. Including temperature data in our model allows us to assess the relationship between temperature fluctuations and sales performance.

Fuel_Price: Changes in fuel prices can affect transportation costs and consumer spending habits. Integrating fuel price data into our model enables us to analyze how fuel price fluctuations impact sales, particularly for items requiring transportation.

CPI (Consumer Price Index): CPI serves as a measure of inflation or deflation and reflects changes in the cost of living. Incorporating CPI data into our model helps us account for the effects of inflation on consumer purchasing power and adjust our sales forecasts accordingly.

Unemployment: Unemployment rates can influence consumer confidence and discretionary spending. Including unemployment data in our model allows us to assess the relationship between unemployment levels and sales performance, helping us anticipate changes in consumer behavior.

Commented [A7]: These are not needed in the paper, the time series models used are not multiple linear regression models.
Remove them from the paper.

By incorporating these variables into our time series models, we can build more accurate forecasts and gain deeper insights into the factors driving Walmart's sales performance over time. Each variable provides valuable information that, when analyzed together, can help us make informed decisions and develop effective strategies to optimize sales and business outcomes.

The dataset presents numerous opportunities for analysis, including:

- Exploring sales trends and seasonal variations over the study period.
- Assessing the impact of holidays and external factors on sales performance.
- Developing predictive models to forecast future sales trends and optimize business strategies.

Limitations:

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.

4. Exploratory Data Analysis (EDA)

Commented [A8]: Must be part of the method subsection

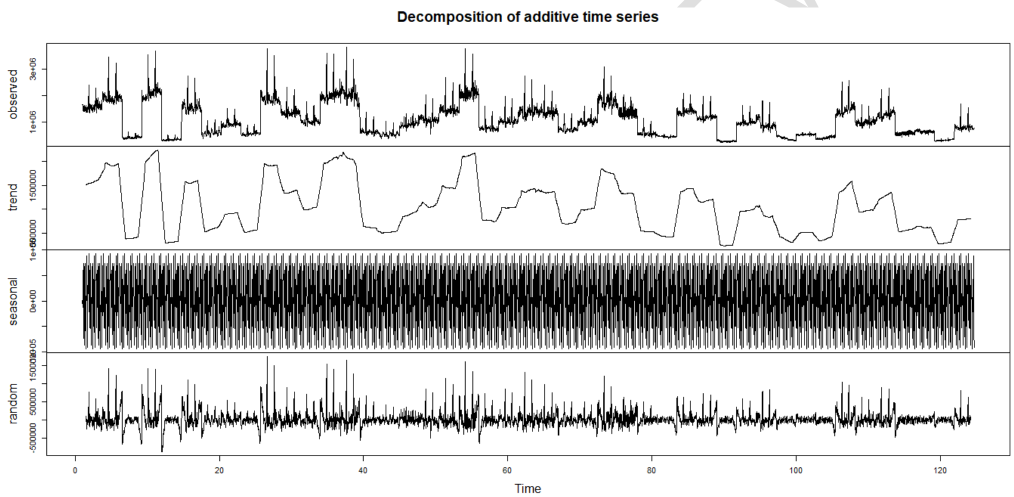
Exploratory Data Analysis (EDA) is a fundamental step in understanding the Walmart sales dataset and preparing it for time series modeling. The following steps were undertaken to explore and preprocess the data:

4.1. Data Import and Sorting:

- The Walmart sales data was imported into RStudio 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.

4.2. 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.



Graph 1: Decomposition plot for Walmart Sales

The trend component of the decomposed time series plot (Graph 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.

4.3. Outlier Detection and Treatment

4.3.1. Box Plots and Scatter Plots

Commented [A9]: Separate method development and results.
Develop the methods without using the statistics and figures from the results.

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

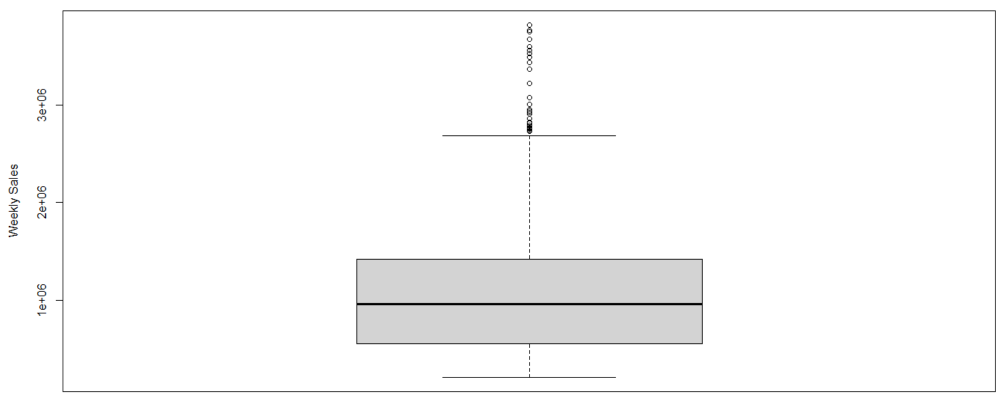


Figure 1: Boxplot showing potential outliers in the Weekly Sales.

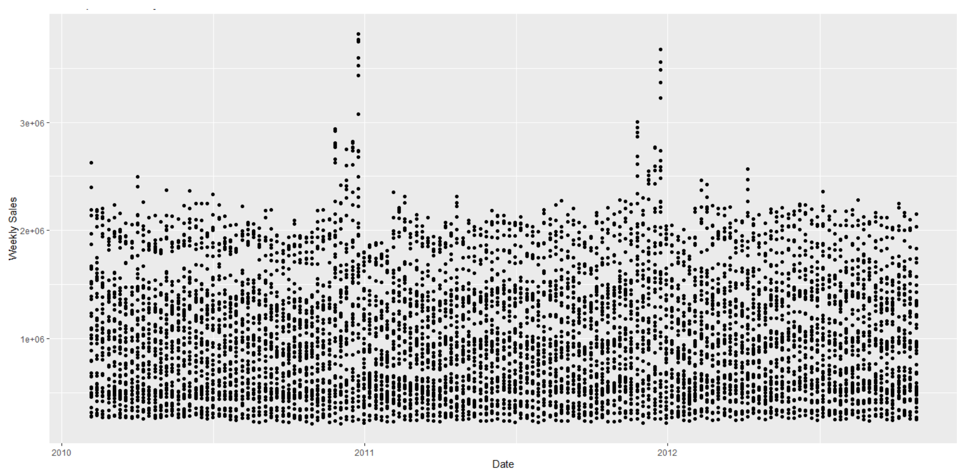


Figure 2: Scatterplot Showing Potential Outliers in the Weekly Sales.

Based on the analysis of boxplots and scatterplots, there is evidence suggesting the presence of outliers in the Weekly_Sales column. The boxplot (Figure 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 (Figure 3) illustrates several data points that appear to deviate substantially from the overall trend, suggesting potential anomalies in the sales data.

Commented [A10]: This must be in the Results and Discussion section.

4.3.2. Quantile-based outlier detection

Commented [A11]: This must be in the Results and Discussion section.

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

4.3.3. 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 Figure 3).

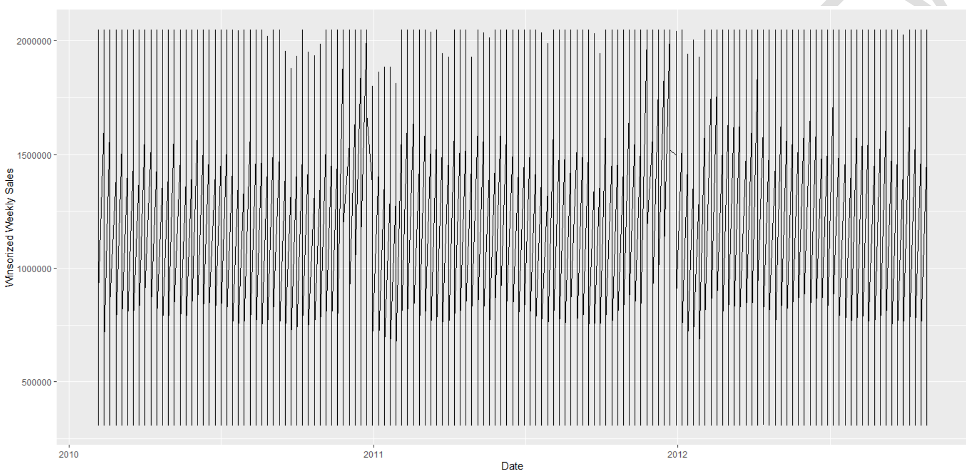


Figure 3: Plot of Winsorized Weekly Sales Over Time

The winsorized dataset was also visualized using a boxplot (See Figure 4). The boxplot displayed the distribution of the winsorized sales data, indicating the median, interquartile range (IQR). No outliers were observed after winsorization.

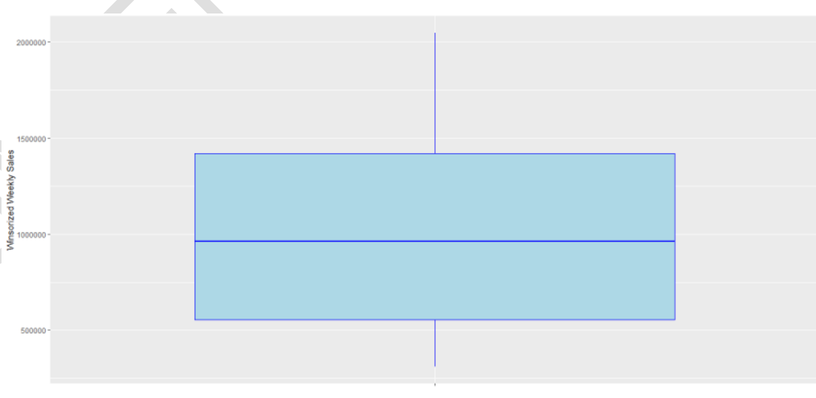


Figure 4: Boxplot of Winsorized Weekly Sales

Commented [A12]: This must be in the Results and Discussion section.

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.

5. Time Series Model Building

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

5.1. 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.

SARIMA (Seasonal AutoRegressive 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.

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.

EXPONENTIAL SMOOTHING:

Commented [A13]: This must be in the Methods subsection.

Commented [A14]: Develop the equations of the model

Commented [A15]: Develop the equations of the model

Commented [A16]: Develop the equations of the model

Commented [A17]: Develop the equations of the model

- 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.

GAUSSIAN PROCESSES:

Commented [A18]: Develop the equations of the model

- 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.

In summary, 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.

5.2. 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.

5.3. Time Series Model Building

5.3.1. 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 (Figure 5).

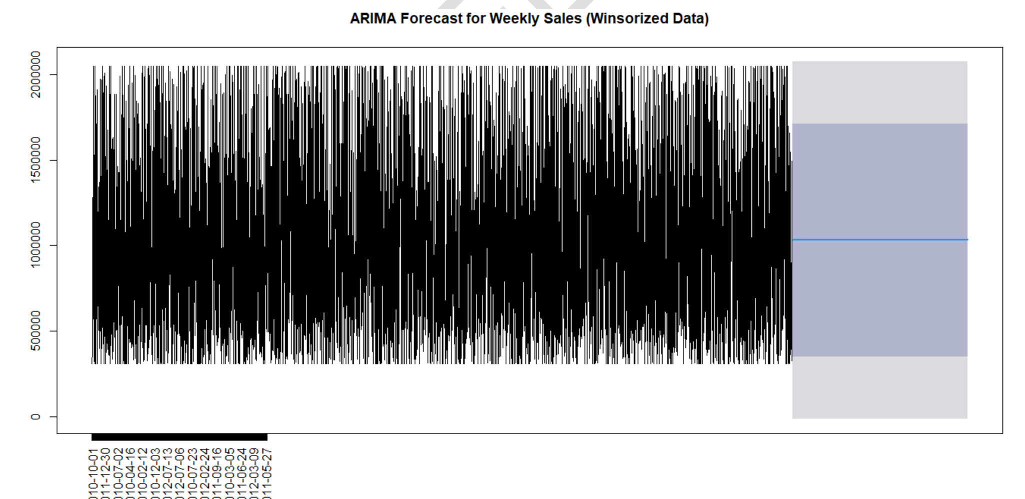


Figure 5: ARIMA Forecast Plot for Walmart Sales

Commented [A19]: This must be in the Results and Discussion section.

5.3.2. 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 (Figure 6)

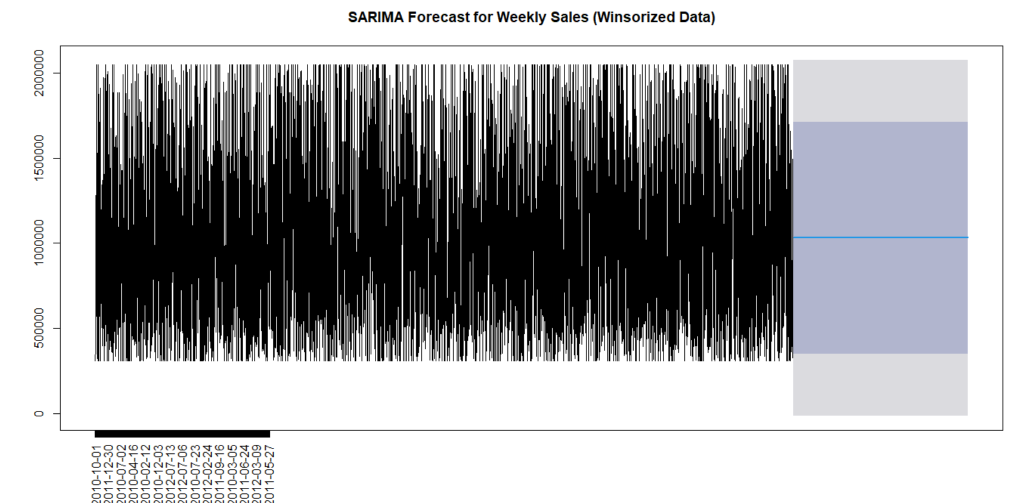


Figure 6: SARIMA Forecast Plot for Walmart Sales

Commented [A20]: This must be in the Results and Discussion section.

5.3.3. 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 (Figure 7)

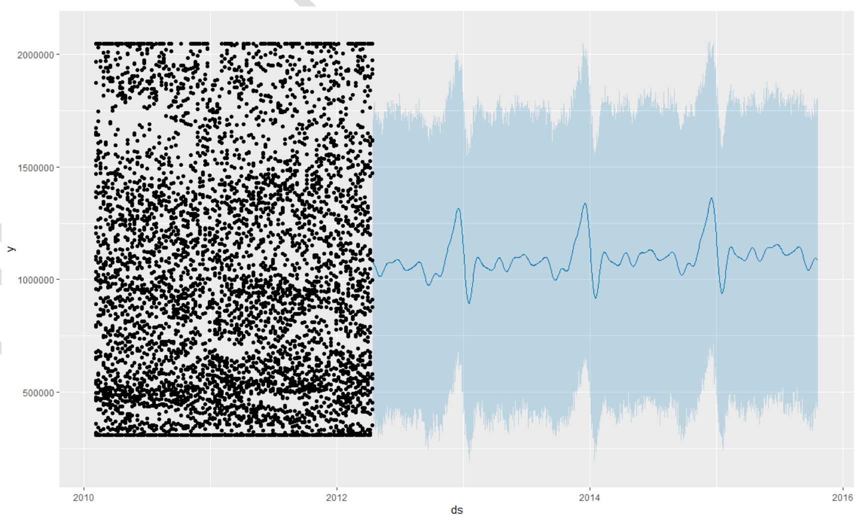


Figure 7: Prophet Forecast Plot for Walmart Sales

Commented [A21]: This must be in the Results and Discussion section.

5.3.4. 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 (Figure 8)

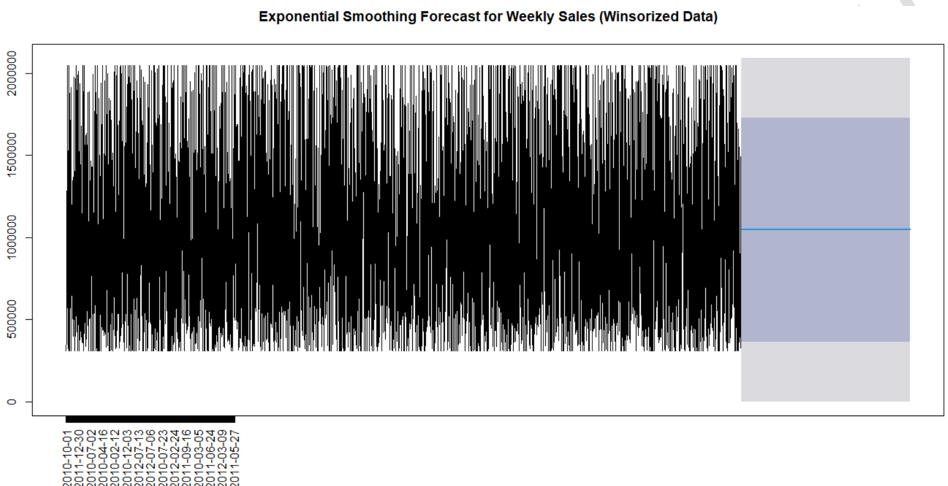


Figure 8: Exponential Smoothing Forecast Plot for Walmart Sales

5.3.5. 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 (Figure 9)

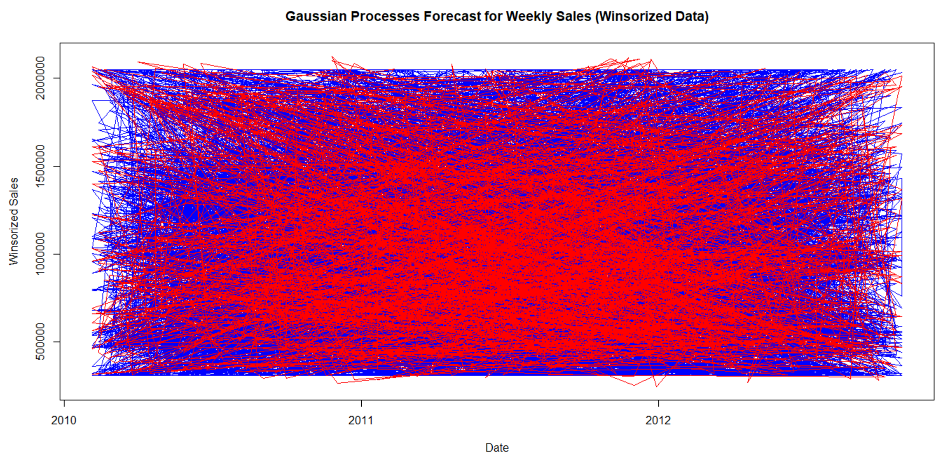


Figure 9: Gaussian Process Forecast Plot for Walmart Sales

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

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.

5.4. Model Analysis

5.4.1. 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.

Commented [A22]: This must be in the Results and Discussion section.

Commented [A23]: Define the model selection process in the methods section.
Put the gotten values of the model evaluation metric in the Results and Discussion section

Commented [A24]: RMSE and MAE indicate the same best model. So, use one of them in the paper, not both.

5.4.2. 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.

5.4.3. 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:

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.

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.

Commented [A25]: RMSE and MAE indicate the same best model. So, use one of them in the paper, not both.

Commented [A26]: This must be in the Results and Discussion section.

Commented [A27]: This must be in the Results and Discussion section.

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.

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.

6. Conclusion:

Commented [A28]: This must be in the Conclusion and Recommendations section.

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.

7. Future Perspective and Recommendations:

Commented [A29]: This must be in the Conclusion and Recommendations section.

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.

COMPETING INTERESTS DISCLAIMER:

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. (2013). An introduction to statistical learning. Springer.

[2] Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2015). Time series analysis: Forecasting and control. John Wiley & Sons.

[3] Kaggle. (2024). Walmart sales data. Retrieved from *https://www.kaggle.com/code/yasserh/walmart-sales-prediction-best-ml-algorithms/input*

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

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

[6] Shumway, R. H., & Stoffer, D. S. (2017). Time series analysis and its applications: With R examples. Springer.

[7] Nicolau, J. L. (2019). Tourism demand forecasting: A review of empirical research. Annals of Tourism Research, 78, 107833.

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

[9] Armstrong, J. S., & Collopy, F. (2001). Principles of forecasting: A handbook for researchers and practitioners. Springer Science & Business Media.

[10] Chatfield, C. (2000). Time series forecasting. CRC Press.

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

[12] Waters, C. D. J. (2009). Principles of marketing. Routledge.

[13] Cox, D. R. (2012). Principles of statistical inference. Cambridge University Press.

[14] Chase, C. W., Jacobs, F. R., & Aquilano, N. J. (2019). Operations management for competitive advantage. McGraw-Hill Education.