COMP8811
DATA ANALYTICS AND INTELLIGENCE
SEMESTER 2 2024
Assignment 2

Sales Forecasting

Jau-Chian Pan 1564921

Table of Contents

Abstract 1
Introduction and Problem Identified 2
Dataset 3
ABI Technique Review 3
Time Series Analysis 3
Advantage /Benefits 4
Disadvantages/ Limitations 4
Time Series Models 4
AutoRegressive (AR) 4
Moving Average (MA) 5
Seasonal AutoRegressive Integrated Moving Average 5
Solution Design and Development 6
Evaluation 8
Monthly Sales Trends Across Stores 9
Sales Comparison on Holidays vs Non-Holidays 10
Time Series Decomposition 10
SARIMA Model Evaluation 11
Dashboard 12
Conclusion and Future Development 13
Reference 14

Abstract

This project focuses on the need for accurate sales predictions within the retail decision-making area using advanced business intelligence tools. The Seasonal AutoRegressive Integrated Moving Average (SARIMA) model can be used to sort sales data and capture trends and seasonal changes to provide more accurate predictions.

The suggested solution was designing a dashboard that combines the required model and makes sales predictions visually accessible for decision-makers. Evaluating the dataset and the model to accurately capture sales trends.

Test results show that the model provides accurate predictions and identifies seasonal changes. However, further fine-tuning is required to get real-time analysis working properly, with other improvements that can be made, such as nonstop data collection and the merging of external economic factors to make the model more effective in real-world scenarios.

Introduction and Problem Identified

With the incoming holiday season, retailers face the challenge of coordinating supply chains and ensuring that products are readily available [1]. Research indicates that sales figures peak towards the end of the year because of increasing consumer spending on gifts and holiday preparations [2]. Retailers experience a significant surge in this demand, and this purchasing behaviour tends to increase for items such as toys, electronics, and holiday decorations [3]. How do retailers face this seasonal demand?

The study shows that effective inventory management is crucial for retailers to balance stock levels and meet customer needs [4]. Overestimating can lead to overstocking and increased storage costs, while underestimating can lead to out-of-stock items, missed sales opportunities, and products may remain stable or decline in demand [5];

Moreover, when customers are unable to get their desired items, it can lead to complaints or unhappy customers, which may impact long-term loyalty and brand perception. So consistency in stock forecasting and accurate sales predictions based on observable trends are essential for inventory management [6]. Having accurate sales predictions not only helps to optimize inventory levels and ensure product availability so that customers get what they need to improve their shopping experience [7].

This report discusses how we can use an Advanced Business Intelligence(ABI) tool to predict sales and provide retailers with an effective solution.

Time-series analysis(TSA) is discussed its basic theory, explains its applicability to addressing sales prediction challenges, and explores its benefits and limitations. The model Seasonal Autoregressive Integrated Moving Average (SARIMA) is also discussed, as well as its theory and characteristics. The model's performance is examined and evaluated, and the result is discussed in detail.

To deliver the complete plan to our end users. This report proposes a solution from a brief introduction, a whole development process into a flowchart and software tools, including model building, front-end and backend. An interface demo is displayed, its functionalities are explored and how this system can bring benefits to retail to face their issues.

Dataset

Walmart is one of the largest retail corporations in the world, with over 10,000 stores across several countries that sell a wide range of products, such as groceries, electronics, clothing, and home goods. They have several years' worth of data that can be studied for seasonal spikes and can be used to gather useful information.

Historical sales data has been studied to help in developing an effective SMP. The dataset used for this project, which contains detailed sales information from Walmart, was downloaded from Kaggle.

The dataset includes 6,435 rows and eight columns, covering attributes such as store numbers, dates, weekly sales, holiday flags (signifying if a week had a holiday), Consumer Price Index (CPI), and unemployment rates; it also captures weekly sales data from 45 unique stores between February 2010 and October 2012. This solid foundation for analysing seasonal sales trends and predicting future sales allows Walmart to make more informed operational decisions.

ABI Technique Review

Time Series Analysis

TSA is a sequence of similar items that occur chronologically, with the data representing various information types, such as stock prices and sales figures. The prime goal of this analysis is to identify patterns and trends in the data to make accurate predictions[8].

Time series data is unique because it captures time-related data, making the order of observations crucial, with identical items being numerical records (observations or measurements) showing the changing values of a quantity. [9]. It shows trends, seasonality, cyclic behaviour, and randomness.

- **Trends** refer to the data's direction and its potential future movement.
- **Seasonality** indicates repeating patterns regularly, often influenced by holidays or weather.
- **Cyclic behaviour** involves irregular, repeated variations with no static time frame, usually influenced by economic factors.
- **Randomness** captures unpredictable changes with no consistent pattern.

The model is represented like this:

$$Y_t = T_t \times S_t \times C_t \times R_t$$

- Yt is the **observed value at time t.**
- **Tt is the trend, showing the direction of the data**.

- St is seasonal and represents regular, repeating patterns.
- Ct is cyclical behaviour, irregular, repeating patterns.
- Rt is random, reflecting unpredictable changes with no pattern.

Advantage /Benefits

1. It can predict future values by analysing based on historical data [10].
2. It can deliver insights without needing complex models or large datasets, which saves time and money while still providing reliable predictions [11].
3. It helps to identify trends and seasonal changes in data.
4. It supports improved decision-making in retail by providing more accurate sales predictions.

Disadvantages/ Limitations

The advantages of using a TSA model are numerous, along with the disadvantages:

1. It is heavily dependent on past data, which can cause results that do not capture the upcoming trends properly, which is obvious when using new products or promotions [12]. This disadvantage can lead to less accurate predictions.
2. Choosing the wrong model makes it more likely to give poor predictions. In contrast, simple models can ignore vital trends in complex data.
3. It is known to not be good with sudden changes such as a market shift, economic conditions or when a competitor changes actions [13]. Choosing a model ill-suited for these sudden changes can lead to important variations in the data that can be missed or ignored by the poor model.
4. Models that combine with machine learning can help adapt to changes, but as a result, they can introduce noise if not properly tuned. Noise is random changes or unimportant information in the data that does not follow any pattern. This introduction makes it harder for real trends to be picked up and learnt from because the model is more likely to learn the wrong lesson from the random changes, making any predictions less reliable [14].

With the advantages and disadvantages listed above, this information will help fine-tune and understand TSA. It will make data analysis easier, along with finding trends, patterns and seasonal changes, making TSA a fine fit for a retail environment.

Time Series Models

The SARIMA model is used in this assignment, but it is important to introduce key components: AutoRegressive (AR) and Moving Average (MA).

AutoRegressive (AR)

AR model predicts current values by analysing previous observations. This approach enables it to identify and leverage patterns over time based on historical data [15].

Imagine a mail carrier delivering packages consistently over the past year. He wants to know how many packages he will provide next month. The AR model focuses on

months with past values that have a stronger influence because of higher autoregressive coefficients. These months can be seen as more predictive of future delivery months. In contrast, data from other months with lower coefficients have a smaller impact and may be affected by random fluctuations. The model minimises unnecessary noise and maintains prediction accuracy.

Moving Average (MA)

The MA Model predicts future values based on past error terms. Unlike other models that rely solely on past observations, the MA model also considers random changes as errors that occurred in previous time steps. The model learns the difference between actual and predicted values to make better predictions.

For instance, a two-dollar shop wants to predict daily sales using an MA model. One day, sales might surge unexpectedly due to a Halloween promotion because the store is not ready for such an influx of sales. This spike is captured as an error term and is noted as such. In the next prediction, this error term influences the prediction, helping the model account for similar random variations and reflect their potential impact on future sales.

Seasonal AutoRegressive Integrated Moving Average

After understanding the AR and MA models, we now introduce the SARIMA model, an extension of ARIMA specifically designed to handle seasonality in data. It combines AR and MA models with differencing techniques but lacks an explicit mechanism to address seasonality.

SARIMA effectively adds seasonal components to capture these repeating patterns, such as annual cycles. It uses past values and error terms and incorporates seasonal changes to better model the underlying data structures. These components help the model capture short-term trends and seasonal patterns and make predictions more accurate when seasonality is important.

For example, in a hospital, patient admission data often shows peaks during flu season in winter and allergy season in spring. SARIMA can effectively capture these seasonal spikes, allowing hospital administrators to plan resources, such as staffing and medical supplies, to meet the increased demand.

Solution Design and Development

This project aims to use past data to predict future sales and visualise the results. The solution involves creating a user-friendly dashboard that integrates with the existing web system, allowing users to access and understand these predictions easily. There are two main sections of tools used in this solution:

Data Analysis Tools:

Python is used as a programming language because of its wide range of useful libraries. Pandas read data, check for missing values, and explore data. Matplotlib and Seaborn are used to create charts to visualise the data. Stats models are used to implement the SARIMA model for prediction.

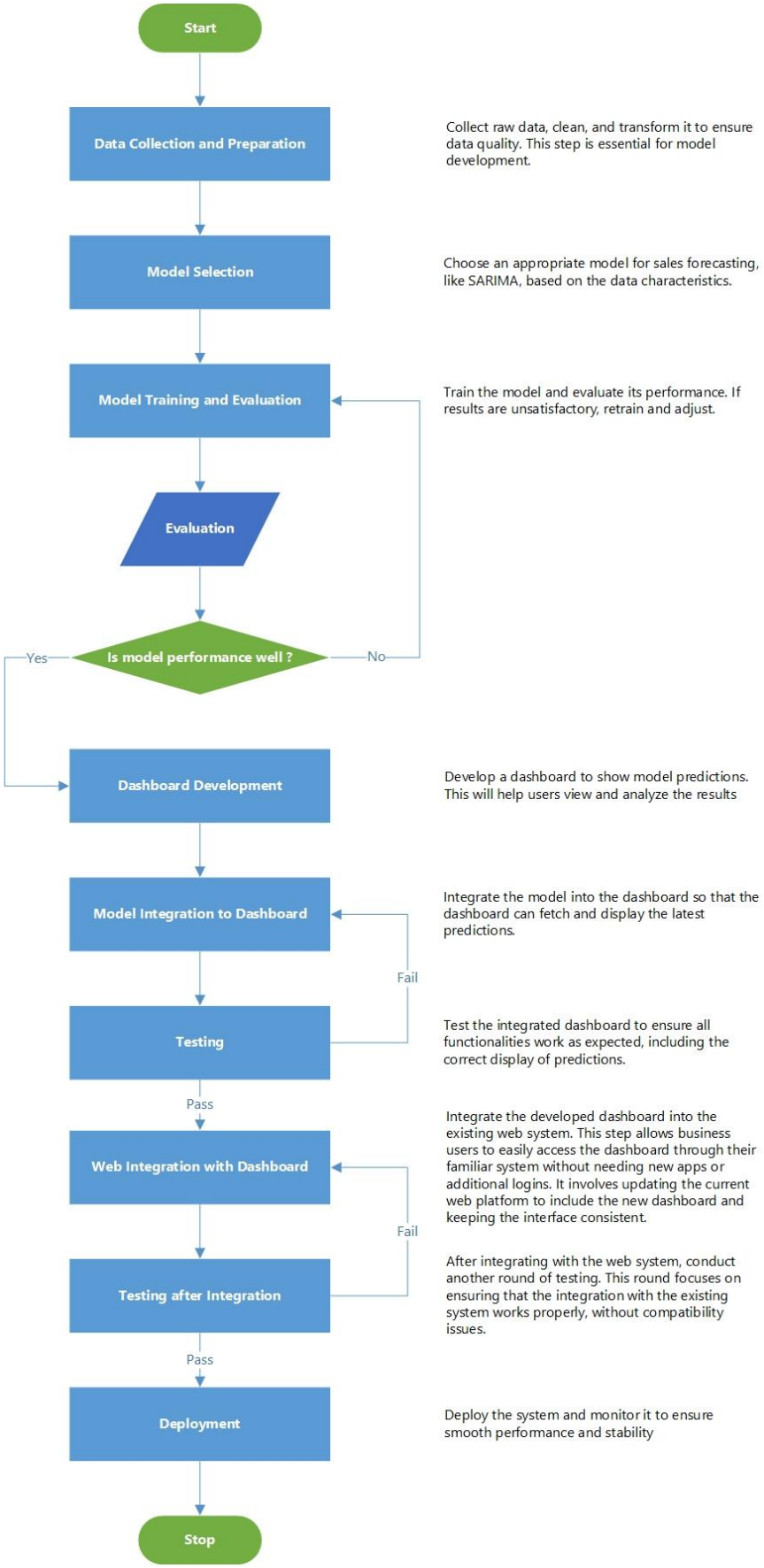
Dashboard Tools:

A dashboard is designed and included with the company's existing website. The front end is developed using HTML, CSS, and JavaScript, while Django controls the back end. The Django REST Framework (DRF) is used to create an API to provide data for the dashboard. This setup lets the dashboard get real-time predictions. It is then easily combined with the existing website, simplifying the process for business managers to view sales forecasts by clicking the "Get Prediction" button. When searching, an AJAX request is sent to Django, which processes the request and displays the correct data on the dashboard.

User Experience and Benefit:

1. Reducing Cost: Integrating a dashboard with the existing system can save significant costs by avoiding the need to develop a new system.
2. Time Savings: Creating a whole new system can be time-consuming; our solution can avoid this process.
3. Minimizing Computing Resource Usage: The dashboard operates as a lightweight plugin, minimizing any additional load on the system and preserving computing resources.
4. Reducing dependency on human resources: An automated forecasting system can reduce the traditional need for managers to rely on data analysts to provide reports.
5. Reducing Training Costs: Users are already familiar with the existing system, so instead of learning an entirely new one, training sessions can be shorter, leading to reduced costs.
6. Visualisation of Predictions: Showing predictions in easy-to-read graphs makes it simpler for people without technical knowledge to understand complex data. This approach not only makes the data clearer but also helps decision-makers make quick and confident choices based on the visual information.
7. Easy Reporting: Users can easily download the prediction as an Excel file. This feature is helpful when managers need to present reports to higher-level decision-makers within the company. Since most people are familiar with Excel, this format allows for easy use, further analysis, and integration into additional reports.

The flowchart shown represents the process:



Evaluation

Exploratory Data Analysis

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6435 entries, 0 to 6434
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Store         6435 non-null   int64
 1   Date          6435 non-null   datetime64[ns]
 2   Weekly_Sales  6435 non-null   float64
 3   Holiday_Flag  6435 non-null   int64
 4   Temperature   6435 non-null   float64
 5   Fuel_Price    6435 non-null   float64
 6   CPI           6435 non-null   float64
 7   Unemployment  6435 non-null   float64
 8   Month         6435 non-null   period[M]
dtypes: datetime64[ns](1), float64(5), int64(2), period[M](1)
memory usage: 452.6 KB
None
```

The image shows the dataset's information, which is a total of 6,435 rows and nine columns. The columns include Store, Date, Weekly_Sales, Holiday_Flag, Temperature, Fuel_Price, CPI, Unemployment, Month, and their data types. All values in the dataset are present.

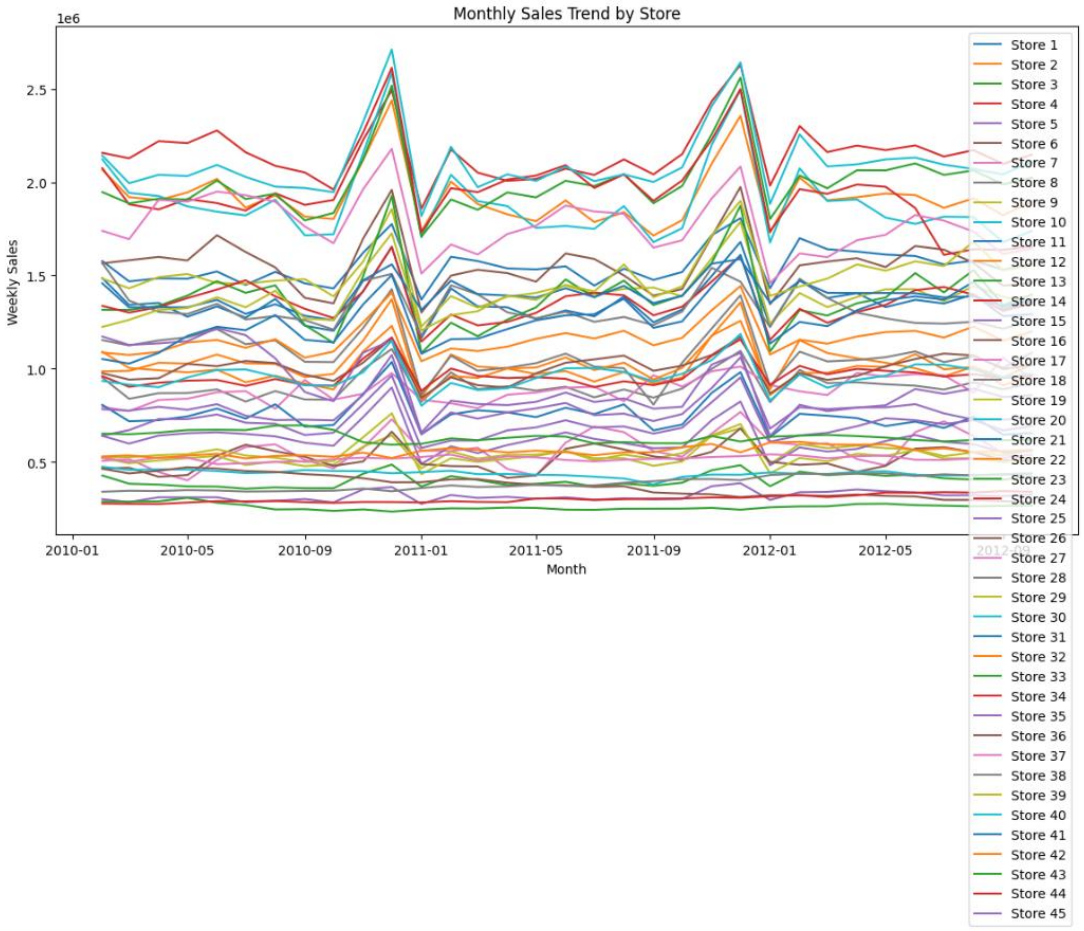
```
   Store       Date  Weekly_Sales  Holiday_Flag  Temperature  Fuel_Price  \
0      1 2010-02-05    1643690.90             0        42.31       2.572
1      1 2010-02-12    1641957.44             1        38.51       2.548
2      1 2010-02-19    1611968.17             0        39.93       2.514
3      1 2010-02-26    1409727.59             0        46.63       2.561
4      1 2010-03-05    1554806.68             0        46.50       2.625

          CPI  Unemployment    Month
0  211.096358         8.106  2010-02
1  211.242170         8.106  2010-02
2  211.289143         8.106  2010-02
3  211.319643         8.106  2010-02
4  211.350143         8.106  2010-03
      Store       Date  Weekly_Sales  Holiday_Flag  Temperature  Fuel_Price  \
6430     45 2012-09-28     713173.95             0        64.88       3.997
6431     45 2012-10-05     733455.07             0        64.89       3.985
6432     45 2012-10-12     734464.36             0        54.47       4.000
6433     45 2012-10-19     718125.53             0        56.47       3.969
6434     45 2012-10-26     760281.43             0        58.85       3.882

            CPI  Unemployment    Month
6430  192.013558         8.684  2012-09
6431  192.170412         8.667  2012-10
6432  192.327265         8.667  2012-10
6433  192.330854         8.667  2012-10
6434  192.308899         8.667  2012-10
Unique stores: [ 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45]
There are 45 unique stores in the dataset.
```

The presence of 45 unique stores suggests that the dataset spans multiple locations, which allows for analysing and comparing performance across different store environments and conditions.

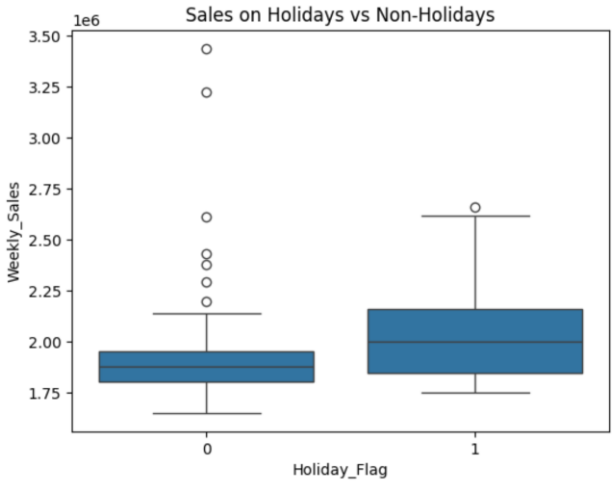
Monthly Sales Trends Across Stores



This chart illustrates the monthly sales trends for 45 stores from January 2010 to September 2012. Each line represents a store, and the Y-axis indicates weekly sales figures. Notably, there are significant sales spikes around the end of each year, likely correlating with holiday seasons, which reflect increased consumer spending during these periods.

There is considerable variation in sales performance across different stores, with some consistently achieving higher sales than others. This insight suggests that location or other factors significantly impact store success. These trends can assist management in optimising inventory levels, allocating resources effectively during high-demand periods, and tailoring marketing efforts to boost sales in underperforming locations.

Sales Comparison on Holidays vs Non-Holidays



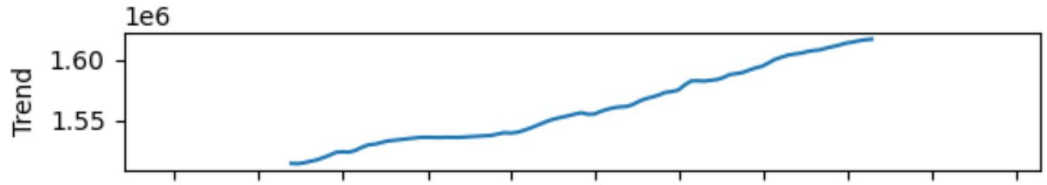
This boxplot compares weekly sales for holidays (Holiday Flag = 1) and non-holidays (Holiday Flag = 0). The median weekly holiday sales are slightly higher than non-holiday sales, indicating increased sales during holidays. The range of the environment for holidays also shows greater changeability, which suggests that sales change more during holidays, likely due to promotions and increased consumer activity.

There are several outliers for holidays and non-holidays, with sales reaching up to 3.5 million. These outliers represent exceptional sales weeks, possibly influenced by specific promotional events or other external factors. Overall, sales tend to be higher and more variable during holidays, highlighting the need for careful planning to meet increased demand during these periods.

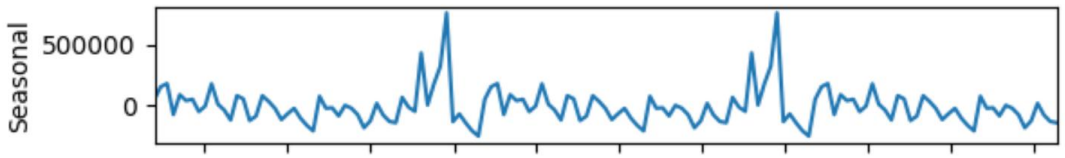
Time Series Decomposition

The images below show the decomposition of a time series into four components: Original Series, Trend, Seasonal, and Residual. Each part provides specific insights into the sales data.

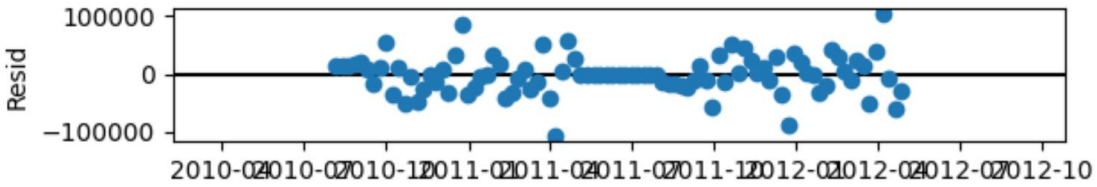
The graph displays the original weekly sales data. You can see some significant spikes at specific times, such as around the end of the year, likely due to holiday promotions or events.



The graph shows the trend component of the time series, indicating the overall movement over time. In this case, a gradual upward trend suggests that sales have been generally increasing.

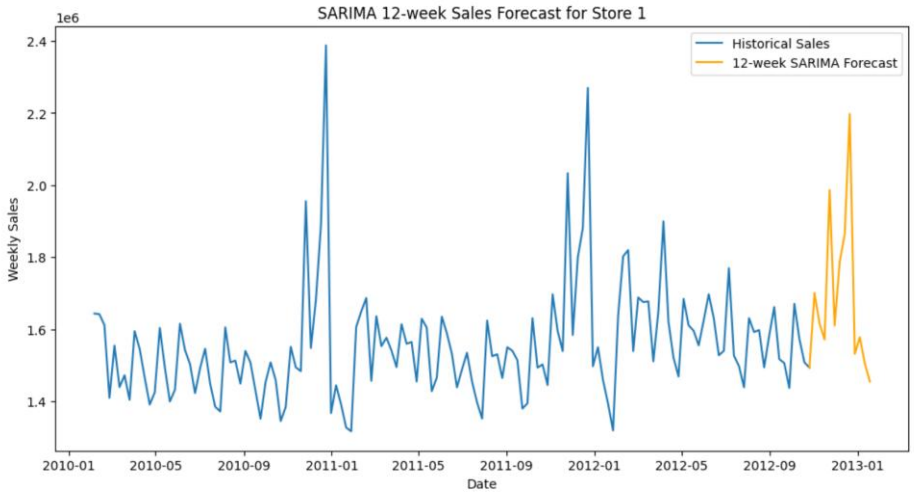


The graph illustrates the seasonal component, representing repeating patterns in the data. You can observe recurring spikes, particularly towards the end of each year, which could be linked to holiday sales or seasonal demand.



The graph shows the residuals, representing the remaining data after removing the trend and seasonal components. Ideally, the residuals should appear as random noise without any distinct patterns. Here, you can see some variation, with a few points showing larger deviations, which may indicate anomalies or unexpected market fluctuations.

SARIMA Model Evaluation



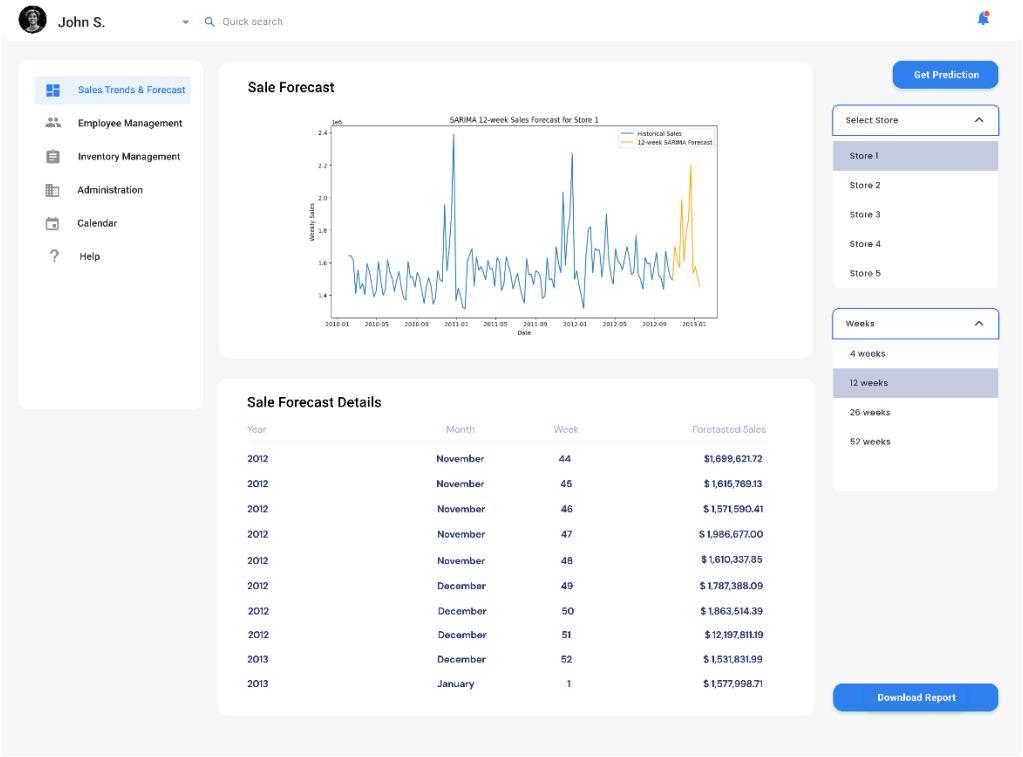
This chart shows the 12-week sales forecast for Store 1 using the SARIMA model. The blue line represents the historical weekly sales for Store 1 from early 2010 to late 2012. The historical data shows large seasonal peaks, especially noticeable around the end of 2010 and 2011. The orange line represents the forecast starting in late 2012. This forecast projects a similar pattern, with increasing sales during specific periods, indicating expected seasonal fluctuations. The model has effectively captured the trend and seasonality present in the historical data.

```
Forecast for the next 4 weeks:
2020-12-06    390.0
2020-12-13    405.0
2020-12-20    420.0
2020-12-27    435.0
Freq: W-SUN, Name: predicted_mean, dtype: float64
Mean Squared Error (MSE): 0.0
Root Mean Squared Error (RMSE): 0.0
```

The model's Mean Squared Error is 0.0, and the Root Mean Squared Error is also 0.0, indicating that the forecast perfectly matches the test data for this period. However, such perfect results are rarely seen in real-world scenarios. The dataset contains only 52 weeks of sales data, which suits the current situation. It is necessary to continuously collect data to improve the model for real-world scenarios. A longer period will help reveal long-term trends and seasonality, allowing the model to learn the complexity of the data more effectively and make better future predictions.

Dashboard

The dashboard integrates a new "Sales Trends & Forecast" section into the existing system. The new section provides a visual and interactive approach to sales forecasting.



The dashboard can be separated into three parts:

Sales Forecast Graph: This section shows a graph that combines historical sales data with forecasted trends for the selected store, allowing users to visualise both past sales patterns and projected changes over time.

Sales Forecast Details Table: A table beneath the graph provides forecasted sales information organised by year, month, and week. It gives specific numerical forecasts to complement the visual trends shown in the graph.

Filters and Get Prediction Button: On the right side of the interface are the dropdown filters, which are able to select one of 42 available stores and a time range option to select the number of weeks, 4, 12, 26, or 52. The "Get Prediction" button lets users submit the selected filter and generate a graph, along with a report. The filters let users focus on specific stores and use an adjustable prediction period to use as they see fit.

Download Report: Saves the forecasting report into a new Excel file on the user's desktop. The file contains columns for 'Year,' 'Month,' 'Week,' and 'Sales Prediction Number.'

This dashboard is easy to use to view sales forecasts. It has simple tools to pick stores and time ranges, and a table with detailed forecast data. The design is clean and helps users make quick decisions.

Conclusion and Future Development

In conclusion, after evaluating the model and dataset, the result shows that the model captures seasonality and trends well. However, it may need to be more adaptable to sudden changes in consumer behaviour or other factors. The solution is still under development.

Key steps have been completed, including the model building and the dashboard. Steps still need to be completed, such as further testing, combining the dashboard with the web platform, and model integration with the dashboard.

When these steps are complete, the retailer can check past sales and sales predictions through the dashboard anytime. This completed dashboard will help retailers with the upcoming holiday season and hopefully streamline resource and inventory management.

Collecting data is important as this will continue to improve the model's performance. This additional data will increase the range of captured factors affecting sales and make better predictions.

Reference

[1] E. Obermair, A. Holzapfel, and H. Kuhn, "Operational planning for public holidays in grocery retailing - managing the grocery retail rush," *Oper Manag Res*, vol. 16, no. 2, pp. 931–948, Jun. 2023, doi: 10.1007/s12063-022-00342-z.

[2] "Christmas Shopping Statistics Statistics: Market Data Report 2024." Accessed: Nov. 11, 2024. [Online]. Available: https://worldmetrics.org/christmas-shopping-statistics/

[3] A. Eira, "57 Essential Christmas Shopping Statistics: 2024 Market Share Analysis & Data," Financesonline.com. Accessed: Nov. 11, 2024. [Online]. Available: https://financesonline.com/christmas-shopping-statistics/

[4] I. I. Shajema, "EFFECT OF INVENTORY CONTROL PRACTICES ON PERFORMANCE OF RETAIL CHAIN STORES IN NAIROBI COUNTY, KENYA," *Journal of International Business, Innovation and Strategic Management*, vol. 2, no. 1, Art. no. 1, May 2018.

[5] F. Inventory, "Guide to Minimizing Inventory Overstocks and Understocks," Finale Inventory. Accessed: Nov. 11, 2024. [Online]. Available: https://www.finaleinventory.com/inventory-management/guide-to-minimizing-inventory-overstocks-and-understocks

[6] D. Waters and D. Waters, *Inventory Control and Management*. Hoboken, UNITED KINGDOM: John Wiley & Sons, Ltd., 2002. Accessed: Nov. 11, 2024. [Online]. Available: http://ebookcentral.proquest.com/lib/unitec/detail.action?docID=219750

[7] S. Valayakkad Manikandan, "Data-Driven Retail: Leveraging Forecasting Models To Enhance Customer Experience And Operational Efficiency," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 06, pp. 2582–5208, Jul. 2024, doi: 10.56726/IRJMETS60655.

[8] "Practical Time Series Analysis : Step by Step Guide Filled with Real World Practical Examples." Accessed: Nov. 11, 2024. [Online]. Available: https://web-p-ebscohost-com.libproxy.unitec.ac.nz/ehost/ebookviewer/ebook/ZTAwMHh3d19fMTYwNzg1MF9fQU41?sid=b7c5eb5d-c4fc-402b-9776-3202f1b0c6e7@redis&vid=0&format=EB&lpid=lp_6&rid=0

[9] G. Tunnicliffe Wilson, M. Reale, and J. Haywood, *Models for Dependent Time Series*. Milton, UNITED KINGDOM: CRC Press LLC, 2015. Accessed: Oct. 31, 2024. [Online]. Available: http://ebookcentral.proquest.com/lib/unitec/detail.action?docID=2122534

[10] D. C. Montgomery, "Introduction to Time Series Analysis and Forecasting".

[11] "Time Series Analysis: Definition, Types & Examples | Sigma Computing." Accessed: Oct. 31, 2024. [Online]. Available: https://www.sigmacomputing.com/resources/learn/what-is-time-series-analysis

[12] M. Mas-Machuca, M. Sainz, and C. Martinez-Costa, "A review of forecasting models for new products," *IC*, vol. 10, no. 1, pp. 1–25, Feb. 2014, doi: 10.3926/ic.482.

[13] M. G. Dekimpe and D. M. Hanssens, "Time-series models in marketing:: Past, present and future," *International Journal of Research in Marketing*, vol. 17, no. 2, pp. 183–193, Sep. 2000, doi: 10.1016/S0167-8116(00)00014-8.

[14] A. Parmezan, V. Alves de Souza, and G. Batista, "Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model," *Information Sciences*, Jan. 2019, doi: 10.1016/j.ins.2019.01.076.

[15] P. J. Brockwell, R. A. Davis Jr., and R. A. Davis, *Time Series: Theory and Methods*. New York, NY, UNITED STATES: Springer, 2009. Accessed: Oct. 31, 2024. [Online]. Available: http://ebookcentral.proquest.com/lib/unitec/detail.action?docID=3070644