Exploring the Application of Data Analytics to Minimize Waste in the Fast Fashion Industry

Valentina Martucci

A Thesis Submitted in Partial Fulfilment
of the requirements for the
Degree of
Master of Science in Data Analytics

cct College Dublin
Computing • IT • Business

September 2024

Supervisor: Taufique Ahmed

Abstract

This study explores the application of data analytics and machine learning techniques to optimize inventory management and reduce waste in the fast fashion industry. The research utilized secondary data from the Visuelle 2.0 dataset, incorporating sales information, restock data, and product images to create a multimodal model for product classification. A convolutional neural network (CNN) was integrated with sales data to classify products into high, medium, and low-selling categories, a novel approach in fast fashion forecasting that combines visual and numerical data. Multiple machine learning models, including Neural Networks, Random Forest, and time series models like ARIMA, were evaluated for restock prediction, demonstrating that advanced models outperform traditional forecasting methods in predicting sales and restocks. Hyperparameter tuning was applied to optimize the models, and the final multimodal CNN model achieved an accuracy of 95.35%. The results demonstrate the potential of combining sales and visual data to improve inventory control and reduce waste, contributing to more sustainable practices in the fast fashion industry. Future work will focus on refining the multimodal approach, improving real-time prediction, and further exploring the integration of additional data sources, such as customer behaviors, to enhance predictive accuracy.

Table of Contents

1. Introduction

The quick turnover of trends, low prices, and quick production cycles of the fast fashion industry have completely changed how consumers interact with fashion. Thanks to this business strategy, which was pioneered by retailers like Zara and H&M, high-fashion trends are now more widely available to consumers than ever before. Fast fashion's popularity, though, has a high social and environmental price. Due to overproduction, significant waste, and a large carbon footprint caused by the industry's emphasis on volume and speed, the fashion industry is one of the main sources of pollution in the world.

1.1 Background

In order to keep up with the constantly shifting needs of their customers, fast fashion retailers create new collections several times in a season using an inventory turnover strategy. Even while this strategy has financial benefits, it is not sustainable by nature. Overproduction is frequently the result of pressure to maintain high inventory levels and react fast to market developments. Large amounts of unsold items are the result of excess inventory, which finally ends up as garbage. The fashion industry is one of the world's biggest industrial pollutants, producing an estimated 92 million tons of garbage year (Niinimäki et al., 2020).

In addition to producing garbage, the fast fashion industry uses a lot of natural resources. Rapid manufacturing procedures in this industry use a lot of resources, including raw materials, electricity, and water. For example, up to 2,700 liters of water—enough to sustain one person's drinking for 2.5 years—may be used during the production of a single cotton t-shirt (WWF, 2013). Furthermore, the usage of petroleum-derived synthetic fibers like polyester further exacerbates environmental deterioration by adding to the amount of non-biodegradable waste and microplastic contamination in the ocean.

Fast fashion has important social repercussions in addition to its environmental effects. Since the majority of production takes place in developing countries, the industry's need for cheap labor has resulted in poor working conditions, low pay, and even abuses of human rights. These problems have sparked a global push for the fashion industry to adopt more sustainable methods, which has prompted businesses to look for creative ways to strike a balance between profit and social and environmental responsibility.

1.2 Problem Statement & Research Question

Fast fashion is facing increasing attention to environmental concerns, yet inventory control and waste reduction remain major obstacles for the sector. Conventional methods of managing inventories frequently result in overstocking or stockouts because they are unable to accurately forecast customer demand. While stockouts can result in lost sales opportunities and decreased customer satisfaction, overstocking causes extra inventory that cannot be sold and contributes to waste. Fashion merchants' financial performance is impacted by these inefficiencies, which also have a negative impact on the environment.

The difficulty is in designing a supply chain that is flexible enough to adjust quickly to shifting customer demands without sacrificing sustainability. This is where data analytics becomes an effective instrument. Fast fashion companies may minimize waste, optimize inventory levels, and obtain deeper insights into consumer behavior by utilizing modern data analytics approaches. By using data analytics to forecast demand, identify trends, and manage inventories in real time, production may be more closely matched to real market demands.

This study aims to address the following research question:

Can data analytics be leveraged to optimize inventory management and reduce waste in the fast fashion industry?

This question seeks to explore the potential of data-driven solutions to address the industry's sustainability challenges.

1.3 Research Objectives

The primary hypothesis of this research is that the application of data analytics can significantly enhance inventory management and reduce waste within the fast fashion industry. To test this hypothesis, this study aims to achieve the following objectives:

1. **Investigate the Current State of Inventory Management and Waste Reduction Practices in the Fast Fashion Industry**: This objective involves conducting a comprehensive review of existing literature and practices to understand the current landscape of inventory management and waste reduction in the fast fashion sector. The goal is to identify the strategies being employed, the challenges faced, and the gaps that exist in managing inventory and reducing waste.

2. **Explore the Potential of Data Analytics to Improve Inventory Management and Reduce Waste**: This objective seeks to delve into the capabilities of data analytics within the context of the fast fashion industry. It involves examining case studies where data analytics has been successfully implemented, analyzing the techniques used, and understanding how these techniques can be applied to improve inventory management and reduce waste.
3. **Develop and Test a Data-Driven Model for a Fast Fashion Company**: The final objective is the practical application of the insights gained from the previous objectives. This involves developing a data-driven inventory management model tailored to the needs of a fast fashion company. The model will utilize various data inputs, such as sales data, seasonal trends, and consumer behavior, to forecast demand accurately and optimize inventory levels. The effectiveness of the model will be evaluated based on its ability to reduce waste and improve overall inventory management.

1.4 Significance of the Study

This research is important because it tackles a crucial requirement for sustainability in a sector of the economy that is frequently criticized for its effects on the environment and society. The goal of this study is to offer practical insights that can assist fast fashion companies in cutting waste, streamlining their supply chains, and implementing more sustainable practices by investigating the use of data analytics in inventory management. Furthermore, the findings of this research may also have wider implications outside of the fashion business, providing insightful guidance for other industries dealing with comparable issues with waste reduction and inventory management.

In a world where consumer awareness of sustainability is growing, companies that fail to adapt may risk losing market share to more agile and environmentally conscious competitors. Thus, this research not only contributes to the academic understanding of data analytics applications in fashion but also provides practical solutions that can drive change in the industry.

1.5 Scope and Limitations

The scope of this study is limited to the fast fashion business, with a specific focus on how data analytics can be utilized to eliminate waste in inventory management. In order to create an inventory management model, the study will investigate a number of data analytics approaches, such as clustering, predictive modeling, and machine learning algorithms. While the research will provide valuable insights into the potential of data analytics, it is important to acknowledge its limitations. The primary sources of data for the study will be case studies and secondary data, which might not fully represent the subtleties of real-world applications. To guarantee that the model created in this study is both generalizable and scalable, it may also require additional testing in various scenarios and improvement.

1.5 Outline of the Thesis

This thesis is organized into several chapters, each building on the findings and discussions of the previous ones. Chapter 2 presents a comprehensive literature review, providing an overview of current inventory management practices and the role of data analytics in the fast fashion industry. Chapter 3 outlines the research methodology, detailing the qualitative and quantitative methods used to gather and analyze data. Chapter 4 presents the results of the study, including the findings from case studies, and the data-driven inventory management model. Chapter 5 discusses these findings in relation to the research objectives, highlighting their implications for the fast fashion industry. Finally, Chapter 6 concludes the thesis by summarizing the main findings, discussing the limitations of the study, and suggesting areas for future research.

In summary, the fast fashion sector is faced with the dual difficulties of satisfying consumer demand and decreasing its environmental impact. It is at a crossroads. Businesses may revolutionize their operations, improve inventory control, and contribute to a more sustainable future by utilizing the power of data analytics. In order to give fast fashion firms a path for adopting data-driven strategies that are in line with their corporate goals and environmental responsibility, this research attempts to investigate this potential.

2. Literature Review

The fast fashion industry is characterized by quick inventory turnover and high levels of consumer demand. However, this business model often leads to significant waste and unsustainable practices. Recently, data analytics has emerged as a potential solution to these challenges. This literature review will explore the current state of research on the use of data analytics in inventory management within the fast fashion industry.

2.1 Current Practices in Fast Fashion Industry

The fast fashion industry is indeed facing increasing scrutiny due to its overproduction practices, which have led to significant environmental and social consequences.

The industry's rapid expansion has resulted in resource waste, overproduction, and environmental deterioration (Sethi, 2021 – McNeill, and Moore, 2015). This overproduction has contributed to the generation of almost 92 million tons of waste annually, making the fashion industry one of the top three polluters globally (Stringer, Mortimer & Payne, 2020).

Several studies have highlighted the issues associated with inventory management in the fast fashion industry. For example, (Nguyen, Le, and Ho, 2020) discuss how the problem of overproduction has led to a global environmental injustice, particularly affecting low and middle-income countries, where much of the waste ends up in second-hand clothing markets.

(Niinimäki, Peters, Dahlbo, Perry, Rissanen, and Gwilt, 2020) focus on the environmental impact of waste generated by the industry: the authors assessed the environmental impacts of the fashion value chain, focusing on water usage, chemical pollution, carbon dioxide (CO 2) emissions and textile waste. The researchers found that the fast-fashion industry produces more than 92 million tons of waste per year and uses 79 trillion liters of water.

(Long and Nasiry, 2019) examines the environmental impact of the fast fashion business model, focusing on its influence on product quality, variety, and inventory management decisions. They use a model to study how quick production and flexible design in fast fashion lead companies to make more styles but with lower quality, which increases waste. The authors find that fast fashion companies often reduce product quality to keep up with changing trends, which leads to more unsold clothes and environmental damage. They also examine different environmental policies, like stricter rules on waste and taxes on production, and show that while these can reduce waste, they might also lower the quality of products, which could end up harming the environment even more. The study highlights the difficult balance between fast fashion's success and its negative impact on sustainability.

These studies provide a clear picture of the challenges that need to be addressed in this industry.

2.2 Historical Context of Data Analytics in Retail

The use of data analytics in retail has grown exponentially in recent decades. Early applications in retail research were based on basic statistical techniques for sales forecasting and inventory management. As the volume of data increased, so did the complexity and capabilities of analytical tools. (Waller and Fawcett, 2013) discuss the transformative potential of predictive analytics and big data in supply chain design and management and highlight how these tools can improve decision-making processes in retail environments. Similarly, (Chen, Chiang, and Storey, 2012) provide a comprehensive overview of business intelligence and analytics, highlighting their evolution and the increasing role they play in driving business impact across various industries, including retail.

The paper by (Raji et al., 2024) provides a detailed review of real-time data analytics in the retail industry, focusing on both USA and global practices. The authors discuss the historical evolution of retail operations from traditional, intuition-based methods to the adoption of advanced, data-driven strategies. In the USA, major retailers like Walmart and Amazon have employed technologies such as RFID, IoT, and machine learning to enhance customer personalization, optimize inventory management, and implement dynamic pricing strategies. This shift has allowed retailers to better understand consumer behavior, tailor marketing campaigns, and respond to market demands in real-time. Globally, similar practices have been adopted, though challenges such as data privacy, integration complexities, and the need for skilled professionals persist. Overall, the paper underscores the transformative impact of real-time data analytics on retail, enabling more efficient supply chains and personalized customer experiences (Raji et al., 2024).

In sectors like grocery and electronics, the implementation of data analytics has led to substantial improvements in efficiency and profitability. For example, the use of predictive analytics in the grocery sector has optimized perishable goods inventory management, reducing waste and increasing sales (Bell, Gallino, & Moreno, 2018).

These successes provide a valuable precedent for fast fashion retailers, who face similar challenges in managing rapidly changing inventories.

2.3 Potential of Data Analytics in Fast Fashion

In the fast fashion industry, the application of advanced analytics is still relatively new, but its potential for improving inventory management and reducing waste is significant.

(Davenport and Dyché, 2013) explore how big data is being utilized by large companies, including those in retail, to optimize their operations and improve efficiency, setting a precedent for fast fashion retailers to follow.

(Giri, Thomassey, and Zeng, 2019) provide an overview of customer analytics in the fashion retail industry in the era of big data. They discuss how customer analytics can create value in the fashion retail industry and examine strategies and methodologies to analyze consumer data. They also highlight the challenge of retailers in the ability to turn customer data into intelligent actionable insights, according to PwC and SAP retailer survey (Verhoef, Kooge, and Walk, 2016). Their research suggests that employing and investing in these methods and technologies can lead to improved revenues, increased sales, higher customer retention rates, and sustainability in uncertain markets.

The paper titled “The roles of data analytics in the fashion industry” (Oh, 2020) discusses the importance and potential of data analytics in the fashion industry. The author begins by acknowledging the abundance of data in the current fashion business environment. The paper emphasizes the increasing recognition of data’s importance among fashion professionals to improve sales and margins. The author suggests that advancements in data analytics and machine learning have led to a greater appreciation of the value of utilizing Artificial Intelligence (AI)-based software or applications to create efficient fashion design, merchandising, and marketing strategies. It is also highlighted that while many retailers like Amazon have been aggressively finding ways to apply advanced data analytics to improve performance in various areas, traditional fashion brands and retailers tend to rely on experts’ gut instinct rather than data-driven decision making using advanced data analytics. However, a growing interest and investment in embracing big data and data analytics has been observed in the fashion industry.

Data Analytics is a powerful and essential capability for fashion firms to be competitive. The quantity, quality, and diversity of available data continue to grow, creating new and significant opportunities for businesses to use data to improve their decisions with respect to both internal resources, as well as external relationships with suppliers and customers. In this regard, (Sztandera, 2020) brings an interesting perspective on the importance of data analytics education in college fashion curricula, underlying how this discipline is spreading across the industry.

Furthermore, (Silva, Hassani and Madsen, 2020) provide an overview of how Big Data techniques can be applied to the fast fashion industry, with activities like trend forecasting, reducing wastage via returns and excess inventory, analysing and enhancing consumer experience, engagement and marketing campaigns, better quality control and less counterfeits and shortening supply chains to be the most prominent applications. They also found evidence indicating that brands such as Zara, Burberry, LVMH, Swarovski, H&M, Lesara, ASOS, Adidas, Hugo Boss, Macy’s, Montblanc, Tory Burch, GAP and Ralph Lauren have all started using advanced analytics to their advantage.

Meanwhile, (Bradlow, Gangwar, Kopalle and Voleti, 2017) present a case study on how not only the volume of data, but also its quality, along with domain knowledge and statistical techniques, has high importance in a retailing context.

2.4 Data Analytics application in Fast Fashion

The application of data analytics in inventory management within the fast fashion industry is crucial for optimizing supply chain operations, meeting consumer demands but also reducing waste.

In terms of inventory management, fast fashion retailers face the challenge of demand forecasting for continuously changing fashion trends and designs (Chen and Lu, 2021). The demand for fast fashion products is highly volatile and influenced by the latest fashion trends, requiring retailers to implement efficient inventory management and demand forecasting models. The paper “Demand Forecasting for Multichannel Fashion Retailers by Integrating Clustering and Machine Learning Algorithms” (Chen and Lu, 2021) discusses the application of data analytics in demand forecasting for multichannel fashion retailers. The authors propose a model that integrates k-means clustering with extreme learning machines (ELMs) and support vector regression (SVR) for demand forecasting. The research results showed that both the KM-ELM and KM-SVR models are superior to the simple ELM and SVR models. They have higher prediction accuracy, indicating that the integration of clustering analysis can help improve predictions. (Henzel, Wawrowski, Kubina, Sikora, and Wróbe, 2022) also presents several approaches to demand forecasting in the fashion industry, including the naïve method, a custom neighbor approach, a parametric linear mixed model, and an ensemble approach. The authors found that the ensemble method provided the best result for this scenario.

Understanding consumer trends and identifying successful patterns in the production of clothes it’s also critical to optimize the inventory and reduce waste; (Bani-Hani, Al-Obeidat, Benkhelifa and Adedugbe, 2020) present a framework for online social network volatile data analysis in the context of the fast fashion industry. The authors argue that consumer satisfaction data is very volatile for some products due to a short requirement period. Therefore, identifying satisfaction in products is important as it allows businesses to alter production plans based on the level of consumer satisfaction for a product.

(Surabani and Rodriguez, 2023) present a case study of a company that successfully implemented data analytics in their processes. By analyzing customer data, they identified distinct customer segments based on demographics, purchasing behavior, and preferences leading to a boost in sales as well as improved inventory management by analyzing historical sales data, inventory turnover rates and seasonal trends. On the other hand, the authors also discussed the main disadvantages of using data analytics for decision making such as high costs for implementation (infrastructure, maintenance and human resources).

Another great example of a practical application of data analytics technique to improve the inventory management of a fast fashion retailer can be found in (Caro and Gallien, 2010). The authors worked in collaboration with Zara, a Spain-based retailer, to address the problem of distributing a limited amount of inventory across all the stores in a fast-fashion retail network. They formulated and analyzed a stochastic model predicting the sales of an article in a single store during a replenishment period as a function of demand forecasts, the inventory of each size initially available, and the store inventory management policy. They then formulated a mixed-integer program embedding a piecewise-linear approximation of the first model applied to every store in the network. This allowed them to compute store shipment quantities maximizing overall predicted sales, subject to inventory availability and other constraints. The implementation of this optimization model by Zara to support its inventory distribution process, and the ensuing controlled field experiment performed to assess the impact of that model relative to the prior procedure used to determine weekly shipment quantities, resulted in increased sales by 3% to 4%, which is equivalent to $275 M in additional revenues for 2007, reduced transshipments, and increased the proportion of time that Zara's products spend on display within their life cycle.

Time series techniques have been employed in several studies for demand forecasting: (Gardner and Diaz-Saiz, 2002) report a case study on forecasting and inventory planning of a distributor of “product parts.” They find that the right classification of the seasonal time series and the use of proper decomposition procedure can help enhance forecasting results and inventory planning significantly. (Aviv, 2003) explores the supply chain inventory management problem with the ARIMA-based time series demand forecasting model.

Long-term forecasting in fast fashion helps determine order quantities and predict initial sales. Since every item is new with no sales history, forecasting relies on comparing new products to similar past products and using their sales data to make predictions. Based on that, (Brahmdeep and Thomassey, 2016) uses a combination of clustering and classification procedures; more in details, the author proposes a three-step methodology for forecasting sales of new products:

1. **Clustering:** Historical sales data is analyzed using k-means clustering to group products with similar sales patterns (life curves). The average sales curve for each cluster is then extracted as a "prototype of sales."
2. **Decision Trees:** Predict the sales pattern group using to the identified sales prototypes.
3. **Forecasting:** The trained decision tree classifier assigns new products to a specific sales prototype based on their characteristics. This assigned prototype serves as the sales forecast for the new product.

Some AI models are also developed for the complex issue of sales forecasting of new items without or with a limited amount of historical data. These models are based on clustering and classification techniques using decision trees (Thomassey and Fiordaliso, 2006), ANN (Thomassey and Happiette, 2007), or grey model (GM) combined with ANN (Choi et al., 2012).

To summarize, the literature indicates that while the fast fashion industry faces significant challenges in inventory management and waste reduction, data analytics offers promising solutions. Adopting sustainable practices and developing innovative demand forecasting models are crucial for the long-term success of fast fashion retailers. However, the application of data analytics in this context is still relatively new, and further research is needed to fully understand its potential and limitations. Most of the evaluated papers focus on consumers and big data techniques; this research project aims to contribute to this growing field of study by exploring the use of data analytics in inventory management within the fast fashion industry.

3. Methodology

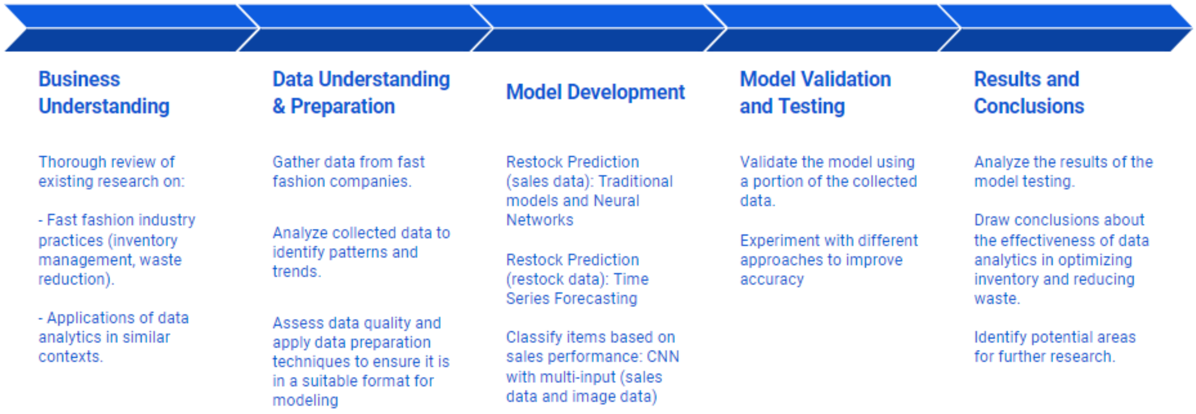
This chapter details the methodology adopted to explore the application of data analytics in optimizing inventory management and reducing waste in the fast fashion industry. The study follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, a widely recognized framework for data mining projects. This methodology consists of six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Hotz, 2024). Each phase of CRISP-DM was carefully executed to ensure that the research objectives were met, and that the analysis was robust and comprehensive. The chapter also discusses the hardware resources and tools used, particularly in training the models using GPU acceleration.

3.1 CRISP-DM Framework

The CRISP-DM framework provides a structured approach to conducting data mining and analytics projects. The six phases of the CRISP-DM methodology, as applied in this research, are as follows:

1. **Business Understanding**: This phase focused on understanding the objectives and requirements from a business perspective and converting this knowledge into a data mining problem definition. The primary objective was to explore how data analytics could be leveraged to enhance inventory management and reduce waste in the fast fashion industry. Literature review has been critical for this step.

2. **Data Understanding**: This phase involved collecting the data and familiarizing ourselves with its properties. It included data exploration and the initial assessment of data quality, such as identifying missing values, inconsistencies, and duplicates. Initial insights have been visualized.
3. **Data Preparation**: Data preparation is often one of the most time-consuming phases in the CRISP-DM process. In this research, data preparation involved cleaning the datasets, handling missing values, normalizing data, and feature engineering. This phase ensured that the data was in a suitable format for modeling.
4. **Modeling**: In the modeling phase, various machine learning algorithms were applied to the prepared data. This phase included selecting appropriate modeling techniques, generating test designs, building models, and assessing model quality.
5. **Evaluation**: The evaluation phase assessed the models to ensure they adequately addressed the business objectives. This included evaluating model performance using different metrics and validating that the models generalize well to new data.
6. **Deployment**: The findings and models developed could be used for strategic decision-making in fast fashion inventory management. This dissertation and its supporting documents (data and code files) represent the deployment of the work.



Business Understanding

Thorough review of existing research on:

- Fast fashion industry practices (inventory management, waste reduction).

- Applications of data analytics in similar contexts.

Data Understanding & Preparation

Gather data from fast fashion companies.

Analyze collected data to identify patterns and trends.

Assess data quality and apply data preparation techniques to ensure it is in a suitable format for modeling

Model Development

Restock Prediction (sales data): Traditional models and Neural Networks

Restock Prediction (restock data): Time Series Forecasting

Classify items based on sales performance: CNN with multi-input (sales data and image data)

Model Validation and Testing

Validate the model using a portion of the collected data.

Experiment with different approaches to improve accuracy

Results and Conclusions

Analyze the results of the model testing.

Draw conclusions about the effectiveness of data analytics in optimizing inventory and reducing waste.

Identify potential areas for further research.

Figure 3.1 - Project Workflow

3.2 Data Sources

The primary data sources for this research were sales and restocks datasets provided by a fast fashion company, along with a dataset containing product images (Github.io, 2017). Additionally, a dataset containing images of the products was used to explore the potential of integrating visual data into predictive models. These datasets included:

- **Sales Dataset**: Information on product sales over time, including product IDs, sales dates, quantities sold, and other attributes. (Github.io, 2017)
- **Restocks Dataset**: Data on product restocking events, detailing when products were restocked, the quantities, and associated product IDs. (Github.io, 2017)
- **Image Dataset**: Product images used to explore the integration of visual data into predictive models. (Github.io, 2017)

3.3 Data Understanding and Preparation

Data understanding and preparation are critical steps in the CRISP-DM process, as they ensure that the datasets are clean, structured, and ready for modeling. The process involved several stages, each of which is described below:

3.3.1 Data Cleaning and preparation

During the data understanding phase, several issues were identified, including missing values, duplicates, and inconsistent data formats. The following methods were used for data cleaning and normalization:

- **Handling Missing Values**: Luckily both the sales and restocks dataset had no missing values for the regression task. Instead, missing values have been handled during the time series phase, as data processing techniques resulted in the creation of some missing values. More details on the imputation techniques are discussed in the relevant section.
- **Categorical Data**: Due to the nature of the data, and specifically the sales dataset containing item's characteristics, many features were identified as categorical data. In order to prepare the data for the modeling phase, those features have been encoded into numerical values.
- **Data Deduplication**: Duplicate entries were identified and removed based on unique identifiers like product ID (external_code) and sales date. This step was necessary to prevent duplicate records from skewing the results.

3.3.2 Feature Engineering

Feature engineering involved creating new features to enhance the predictive power of the dataset:

- **Temporal Features**: Features such as day of the week, month, and season were generated from the sales date to capture temporal patterns in sales. This helped model the impact of different times of the year or week on sales.

- **Product-Level Aggregation**: Data were aggregated by product ID (external_code) to analyze sales patterns at the product level. Although this aggregation reduced data granularity, it was necessary to simplify the modeling process and focus on general trends.
- **Seasonal Features:** The fast fashion is an industry affected by seasonal trends. To make sure the models were able to capture seasonal patterns, new features such as the collection (summer – autumn – winter - spring) were added to the sales dataset.

3.4 Models Evaluation

The second phase of the methodology involved the development and evaluation of various machine learning models to predict restocks and analyze sales trends. This phase was split into two main components: predicting restocks using the sales dataset and conducting time series analysis with the restocks dataset. This phase was conducted using GPU acceleration, specifically with CUDA and Keras-Tuner, to optimize model training times and performance.

3.4.1 Predicting Restocks Using Sales Data

Multiple machine learning models were tested to predict restock quantities based on historical sales data. The chosen model included a mix of some techniques evaluated in the literature review, as well as some new addictions. The models evaluated include:

- **Linear Regression**: A baseline model used to establish a reference point for predictive accuracy. Although simple and interpretable, the linear regression model was inadequate for capturing complex relationships in the data.
- **Random Forest**: An ensemble model that provided significant improvements over linear regression. It effectively handled non-linear relationships and interactions between features, resulting in better predictive accuracy.
- **Gradient Boosting Machines (GBM)**: GBM models like XGBoost were tested due to their ability to handle complex data structures and provide robust predictions. These models iteratively build trees to minimize prediction errors, making them highly effective for this task.
- **Neural Networks**: A simple neural network model was implemented to explore deep learning techniques. Using GPU acceleration with CUDA, the neural network was trained with multiple hidden layers and activation functions like ReLU.

Each model was evaluated based on the following metrics (amsten, 2020):

- **Mean Absolute Error (MAE)** - It is the average difference between the initial values and the predictions. In essence, it outlines our predictions based on the actual results. There is one drawback, though, which is that it doesn't indicate whether we are over- or under-predicting our data based. It can be represented mathematically in this way:

$$\text{MAE} = \frac{1}{N}\sum_{j=1}^{N} |y_j - \ddot{y}_j|$$

- **Mean Absolute Percentage Error (MAPE)** - The percentage equivalent of mean absolute error (MAE).
- **Mean Squared Error (MSE)** - Perhaps the most popular metric used for regression problems. It essentially finds the average of the squared difference between the target value and the value predicted by the regression model.

$$\text{MSE} = \frac{1}{N}\sum_{j=1}^{N} (y_j - \ddot{y}_j)^2$$

- **Root Mean Squared Error (RMSE)** - It is he square root of the average squared differences between actual and predicted values. It keeps the benefits of Mean Squared Error (MSE) but reduces how strongly larger errors are penalized by using the square root.
- **R-squared (R^2)** - This is a post-metric, meaning it's derived from other metrics. Its purpose is to answer how much of the total variation in the target variable (Y) is explained by the variation in the regression line (X). As shown in the formula below, it is calculated using the sum of squared errors. If the sum of Squared Error of the regression line is small => R^2 will be close to 1 (Ideal), meaning the regression was able to capture 100% of the variance in the target variable.

$$R^2 = 1 - \frac{MSE\ (predictions\ against\ the\ actual\ values)}{MSE\ (mean\ prediction\ against\ the\ actual\ values)}$$

3.4.2 Time Series Analysis Using Restocks Data

The restocks dataset, which contained more data points, was used for time series analysis. This analysis aimed to understand the temporal patterns in restocking events and improve the accuracy of future predictions. Key steps included:

- **Data Decomposition**: The restocks data were decomposed into trend, seasonal, and residual components using techniques like Seasonal Decomposition of Time Series (STL). This helped in understanding the long-term trends and seasonal variations in restocking behavior.
- **Model Selection**: Several time series models, including ARIMA (AutoRegressive Integrated Moving Average), Exponential Smoothing, and XGboost, were tested to model the restocks data. ARIMA was chosen for its ability to handle non-stationary data, while SARIMA was selected for its flexibility in capturing seasonality and holidays.
- **Evaluation**: The models were evaluated using metrics such as Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) to assess their accuracy in forecasting future restocks. The time series models provided insights into the cyclical nature of restocking events, although their performance was limited by the reduced data shape after aggregation. More details on the metrics used are given in the next section.

3.5 Multimodal Analysis

The final phase of the research focused on developing a novel approach to classify product images based on sales performance, integrating visual data into the predictive modeling process. This represents a significant innovation compared to existing literature, which typically relies on numerical sales data alone.

3.5.1 Data Integration

The sales dataset was merged with the image dataset based on the product ID (external_code). The image dataset was used to train a model that classifies products into low, medium, and high-selling categories based on their visual characteristics. The sales data were used to label these images according to their sales performance. This approach allows the model to learn patterns and features from the images that correlate with sales performance, providing insights that go beyond traditional numerical data analysis.

3.5.2 Development of the Multimodal Model

For this project, a Convolutional Neural Network (CNN) combined with a tabular model was developed to predict fast fashion product categories based on sales performance and product image features. The aim was to leverage multimodal data—combining visual and numerical features—using neural networks to classify products into sales categories (high, medium, low).

- **CNN Architecture**: The model architecture combined sales features (numerical data) and product images (visual data) using a CNN for image processing and dense layers for sales features. Key components include an image input and a sales data input. Dropout layers were applied to prevent overfitting.
- **Training and Optimization**: The CNN was trained using GPU acceleration with CUDA (NVIDIA Developer, 2012) to manage the high computational load associated with image processing and deep learning. Keras-Tuner (keras.io, n.d.) was employed to optimize the hyperparameters of the CNN, such as the number of layers, learning rate, and filter sizes. This optimization process was crucial for achieving the best possible performance from the model.

3.5.3 Model Training and Evaluation

The model was trained on both sales data and image data. The training process involved multiple architectures:

3. **Baseline Model**: A simple architecture with minimal dense layers and a small number of neurons.
4. **Deeper Model**: A model with additional dense layers and neurons to capture more complex relationships.
5. **Wide Model**: The selected architecture from the hyperparameter tuning process, optimized to balance accuracy and model complexity.

Each model was trained for 10 epochs using a batch size of 32 and a validation split of 20%. The optimizer used was Adam, and the loss function was sparse categorical crossentropy.

The performance of the models was evaluated using the following classification metrics (Bajaj, 2021):

- **Precision**: The proportion of true positive predictions out of all predicted positive instances, measuring the model's ability to avoid false positives.
- **Recall:** The proportion of true positive predictions out of all actual positive instances, reflecting the model's ability to identify all relevant instances.
- **F1 score:** The harmonic mean of precision and recall, providing a balance between the two metrics to evaluate the model's overall performance.
- **Accuracy:** The proportion of correct predictions out of all predictions, indicating the overall correctness of the model across all classes.

3.6 Challenges and Limitations

Several challenges and limitations were encountered during the research:

3.6.1 Data Aggregation and Quality Issues

Aggregating data by product ID reduced data granularity, affecting the model’s ability to capture fine-grained variations. This challenge highlighted the trade-off between data simplicity and model complexity. Extensive data cleaning and preprocessing were required to address missing values and inconsistencies, which introduced potential biases.

3.6.2 Model Complexity and Hardware Resources

Training advanced models like neural networks and multimodal models required significant computational resources. The use of CUDA for GPU acceleration and Keras-Tuner for hyperparameter optimization was essential in managing these computational demands. However, the need for specialized hardware and software may limit the scalability and applicability of these models in other contexts.

3.6.3 Integration of Image Data

The integration of image data for sales classification posed unique challenges, including the need for substantial computational power and the complexity of training deep learning models like CNNs. Despite these challenges, the results demonstrated the value of incorporating image data into sales predictions, providing a novel insight that could be explored further in future research.

The methodology outlined in this chapter provides a comprehensive framework for applying data analytics in the fast fashion industry. By following the CRISP-DM framework and leveraging GPU acceleration, the study developed robust models for predicting restocks, analyzing sales trends, and classifying products based on sales performance and visual attributes. The introduction of image-based classification represents a novel contribution to the field, offering new avenues for enhancing inventory management and reducing waste in fast fashion.

4. Ethics and Data Validity

Ethical issues are crucial in any research effort, especially when working with large datasets that can contain sensitive information. This chapter focuses on the usage of secondary data from the Visuelle 2.0 dataset (Skenderi et al., 2022) while addressing the ethical issues, data validity, and licensing compliance pertinent to the fast fashion project. The project uses data from this publicly available dataset, although it is crucial to note that, while customer data was anonymised in the original source, it was not used in this study.

4.1 Primary Research and Ethical Considerations

Although primary research involving human participants is not part of this study, ethical principles remain central to the responsible use of secondary data. The Visuelle 2.0 dataset provides rich, multi-modal data, including product-level information, sales data, stock levels, prices, images, and exogenous factors such as weather reports and Google Trends data (Github.io, 2017). While the dataset also contains anonymized customer purchase histories from over 667,000 users, this specific subset of data has not been employed in this study. By consciously excluding customer data, the project avoids potential privacy concerns, ensuring compliance with ethical standards surrounding personal data use and protection.

Even though customer data was not used, the project still adheres to the ethical guidelines laid out by the General Data Protection Regulation (GDPR) and other relevant data protection frameworks. These frameworks ensure that any personal or anonymized data must be treated with respect, protecting the identities and rights of individuals.

4.2 Data Validity and Use of Secondary Data

The data used in this project comes from the Visuelle 2.0 dataset, which serves as the secondary data source. Secondary data refers to data that has been previously collected and published for purposes other than this specific research. In this case, the Visuelle 2.0 dataset was created by HumaticsLAB and includes comprehensive information about product sales, inventory levels, product characteristics, and exogenous factors like weather and Google Trends data, making it an ideal resource for fast fashion sales forecasting and analysis.

Since this study does not involve the collection of new, primary data, the project relies on this rich, well-validated secondary dataset. The dataset has been curated to ensure high-quality, multi-modal information, and its use adheres to licensing conditions. Although the Visuelle 2.0 dataset contains anonymized customer data, this portion of the data was not employed in this research, further ensuring that ethical considerations regarding personal data were respected.

In terms of data validity, the focus of this project lies in the performance and accuracy of the machine learning models built from the secondary data. The validity of the research is demonstrated through the models' performance in predicting fast fashion sales and restocking needs. As highlighted in the results, the model's accuracy, represented through key metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), shows that the data has been effectively utilized to produce reliable and actionable insights. This approach to data validity aligns with best practices in data science, where model performance is a key indicator of the effectiveness and validity of the data used.

4.3 Licensing and Ethical Compliance

The Visuelle 2.0 dataset is licensed publicly under a framework that permits academic and research use. According to the dataset’s license, accessible via HumaticsLAB (HumaticsLAB, 2022), proper attribution to the creators is required, and any redistribution of the data must adhere to the same licensing terms. For this project, the dataset is used strictly for research purposes, ensuring compliance with both the license's non-commercial use restrictions and its attribution requirements.

The use of the dataset without the customer data avoids potential issues related to data protection laws, as no personally identifiable information (PII) is accessed or processed. This approach maintains the ethical integrity of the research while still allowing for robust analysis using the available product-level data.

In summary, this project maintains a strong ethical foundation by respecting privacy concerns, ensuring data validity, and adhering to the Visuelle 2.0 dataset's licensing terms. The decision to exclude anonymized customer data further strengthens the ethical integrity of the study, allowing for focused analysis on product forecasting while safeguarding individual privacy.

5. Results

This chapter presents the results of the data analysis and model development processes described in the methodology. Each step taken during the analysis is carefully detailed, including data exploration, model evaluation, and the integration of multimodal data. The results provide insights into the effectiveness of different machine learning models for predicting restocks and classifying items based on sales and visual data. This chapter also discusses the implications of the findings and justifies the chosen methodologies.

5.1 Data Exploration and Preparation Results

The initial phase of the research focused on exploring and preparing the sales and restocks datasets. This step was crucial to ensure data quality and suitability for modeling.

5.1.1 Data Cleaning and Preparation

During this phase, several issues were identified, including missing values, duplicates, and inconsistent data formats. EDA and statistical analysis have been also performed on the data to better understand and predict modeling behaviour. Details of this is explained below for each of the analysis:

- **Sales dataset:** The dataset contains various columns, each representing different attributes of products and their sales over a 12-week lifecycle. There are no missing values in the dataset, which is excellent for modeling as it reduces the need for imputation or dropping rows. The restock values range from 1 to 389, with a mean of approximately 22.4 and a standard deviation of around 17.0, indicating a right-skewed distribution with some outliers. This has been evident when plotting the histogram of the restock variable shows a right-skewed distribution shown in Figure 5.1, confirming that most products have relatively low restock values, with a few products having very high restocks. This suggests that a transformation (e.g., logarithmic) could be beneficial when modeling, particularly if using a linear regression model, to normalize this distribution.
- **Restock dataset:** The dataset contains 949,766 entries; all columns are integer and there are no missing values. The columns contain purely restocks information (external_code, retail, week, year, qty). The observations are based on weekly figures unlike the sales dataset where we have the release_date for each product and the sales for the first 12-week period. There is missing data for years 2016 and 2020, 2021 (that can be explained with the covid19 outbreak when retails where closed).

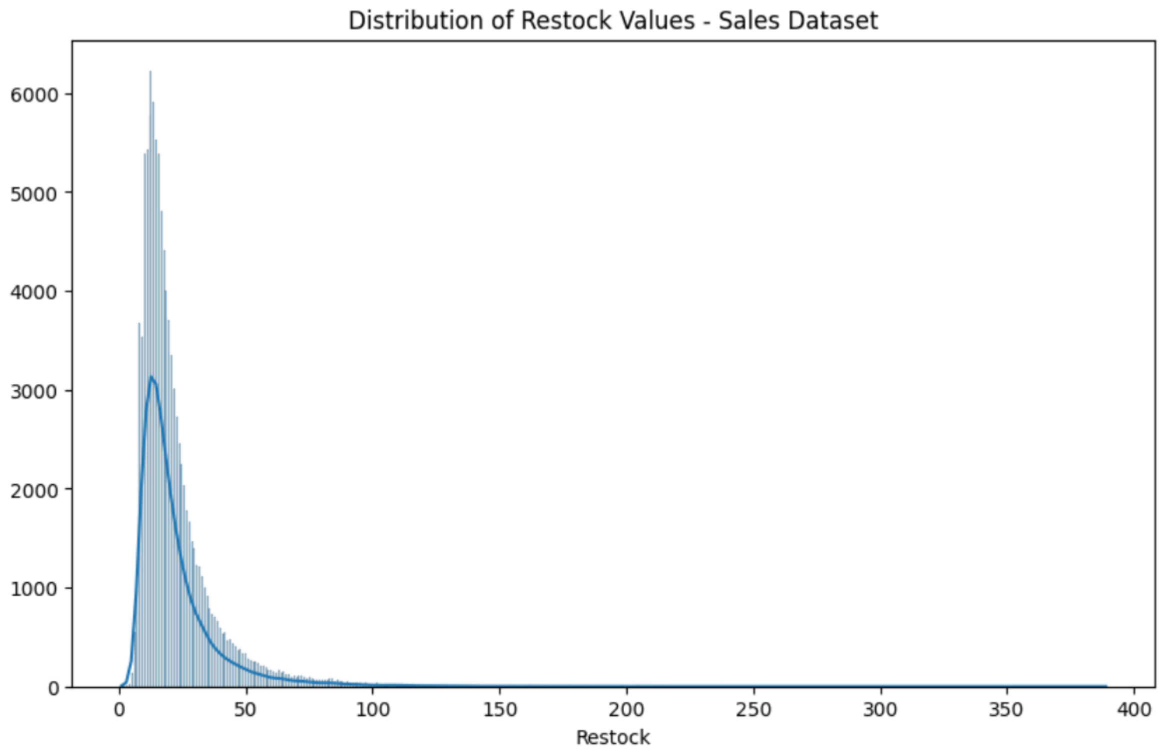


Figure 5.1 – Distribution of restock values in sales dataset

Some features were plotted to better understand the business problem and any trend in the data.

The bar plot in Figure 5.2 shows the total sales distribution across the 12-week lifecycle. It indicates the following:

- Sales tend to be higher in the initial weeks and gradually decline over time.
- There are noticeable drops in sales after week 1 and week 6, suggesting potential points of inventory depletion or reduced demand.

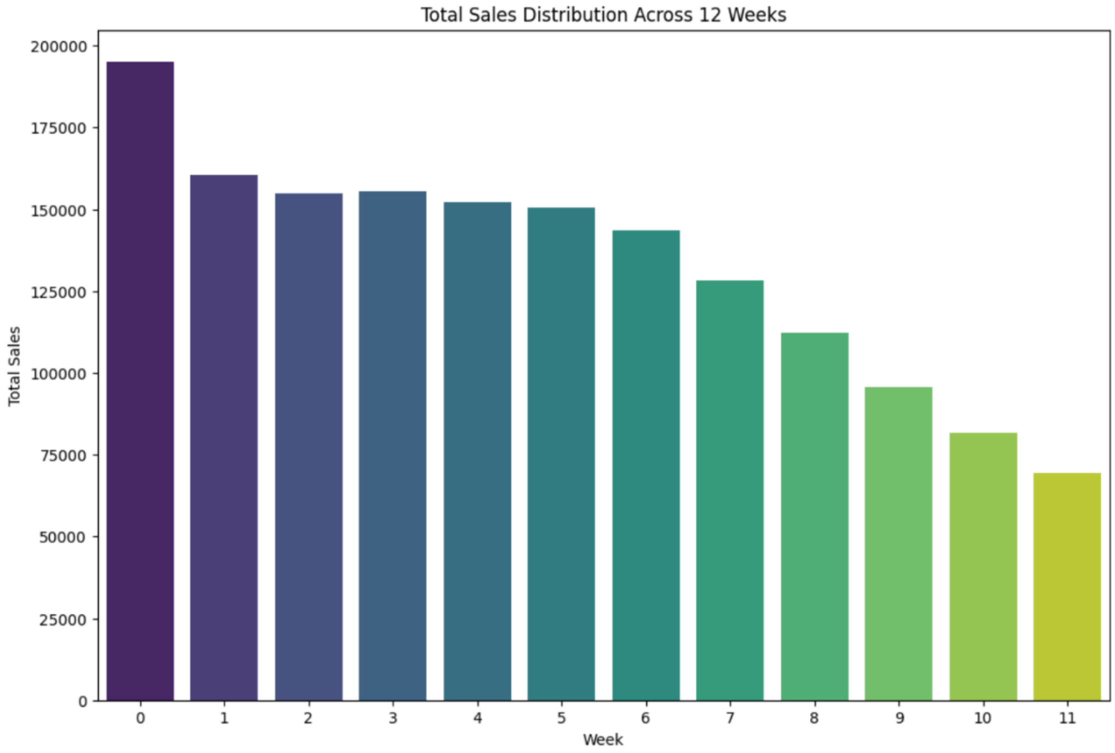


Figure 5.2 - Sales trend across initial 12 weeks

Lastly, correlation analysis has been performed to better understand the relationship between the features of sales dataset. The plotted correlation matrix shown in Figure 5.3 is revealing the following additional insights on the sales data:

- The restock variable has some positive correlations with the early weeks' sales (weeks 0, 1, and 2), which makes sense since high initial sales might trigger more restocking.
- As the weeks progress, the correlation between weekly sales and restock decreases, likely due to the natural decline in sales over time.
- There are strong correlations between the sales in consecutive weeks, indicating a potential trend.

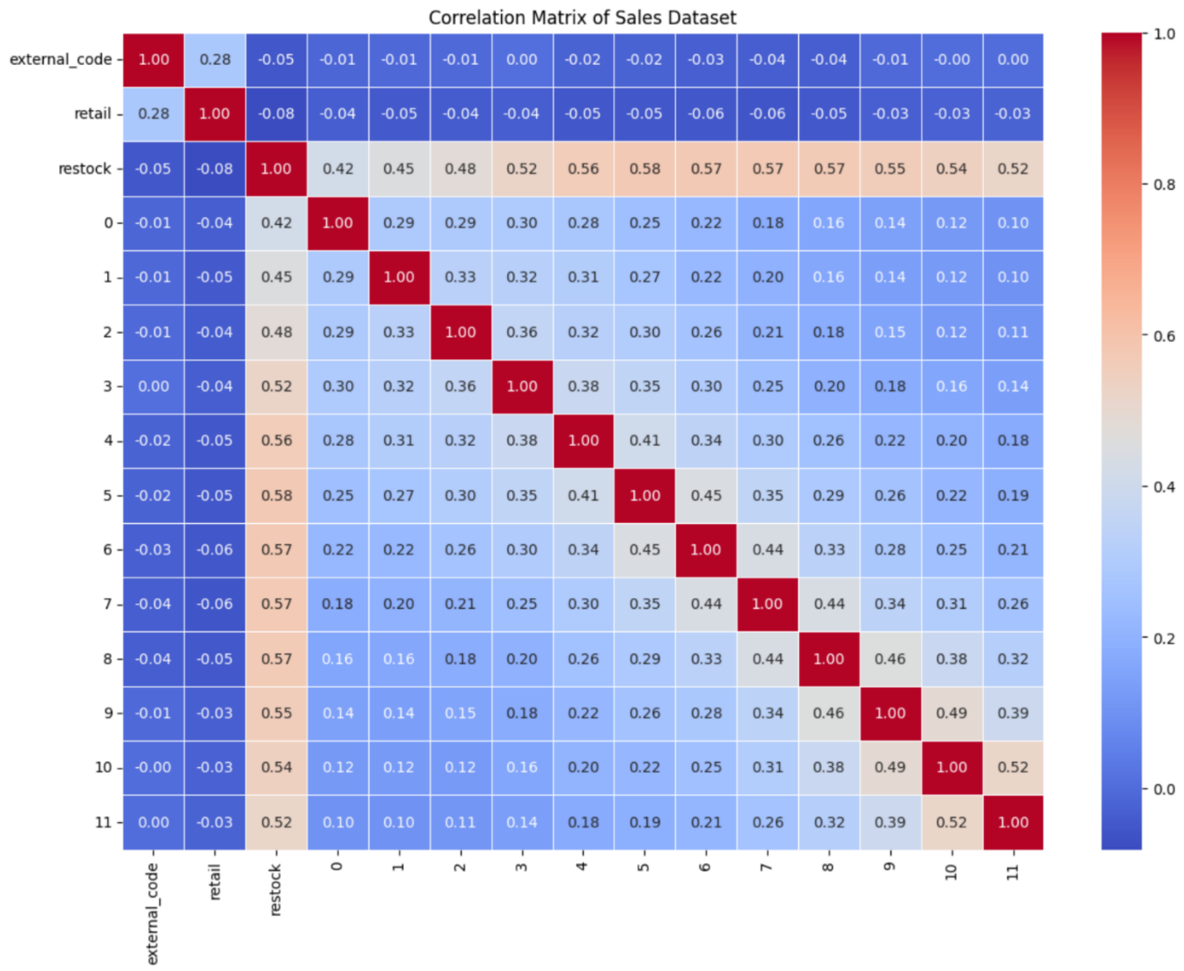


Figure 5.3 - Correlation between sales features

An analysis of the restocks dataset has then been done, and the line plot in Figure 5.4 shows the weekly restocking patterns across different years, revealing a few key insights:

- **Seasonality:** There's a clear pattern of peaks and troughs, which could indicate seasonal demand variations. This is crucial for forecasting when to increase or decrease stock levels.
- **Yearly Trends:** Each year has its unique trend, but some common weeks (like around the middle of the year) consistently show higher restocking levels across different years. This might correlate with fashion seasons or sales periods.
- **Potential Anomalies:** Some years show spikes at unusual times, which might need further investigation to understand if they were due to promotional events, changes in supply chain strategy, or other factors.

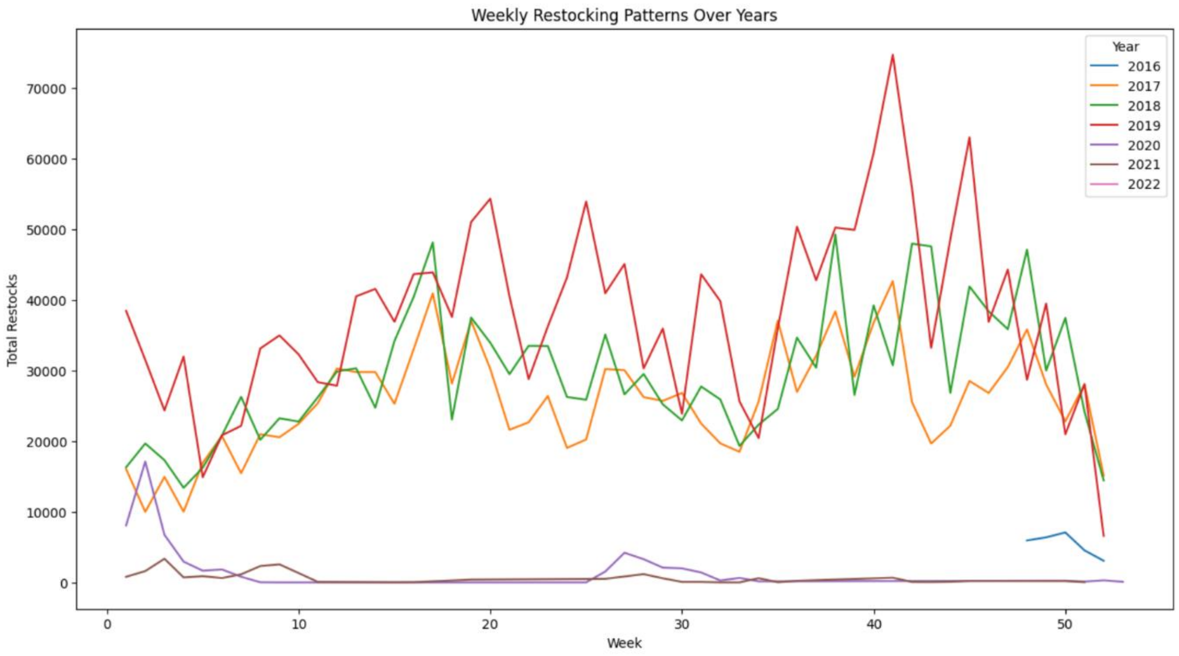


Figure 5.4 - Restocks trends

Several data manipulation techniques have been used to align the dates and features of the two datasets to create a new one, but the resulting dataset showed very little correlation between the only feature from restocks (‘qty’) and the sales features. Because of this, the two datasets have been used independently: sales.csv has been used to explore forecasting techniques, while restocks.csv has been used to explore time series forecasting techniques due to its large number of datapoints.

To standardize the scale of numerical features, such as sales quantity and restock quantity, StandardScaler() was applied. Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data (e.g. Gaussian with 0 mean and unit variance) (Scikit-Learn, 2019).

5.1.2 Feature Engineering

Feature engineering was performed to enhance the predictive power of the dataset. Key features were created to capture temporal patterns and product characteristics:

- **Temporal Features**: New features such as winter/summer/spring release were generated from the sales date. These features aimed to capture the temporal variability in sales, recognizing that consumer purchasing behavior often varies by time of year. For example, sales might be higher during specific seasons like summer or winter.

- **Product-Level Aggregation**: Data were aggregated by product ID (external_code) to analyze sales patterns at the product level. This aggregation allowed for a more focused analysis on individual products, although it also resulted in a significant reduction in data points. The aggregated dataset was used to develop models that predict restocks based on product-level sales history. The trade-off between data granularity and simplicity was considered, as aggregation reduced the dataset size but simplified the modeling process.
- **Categorical Encoding**: Categorical variables such as product category, color, and fabric type were encoded using one-hot encoding. This method created binary columns for each category, which is particularly useful for models like linear regression and neural networks that require numerical input. This method has been chosen over label encoding for the following reasons (Brownlee, 2017):
    - Avoids Ordinality Assumptions: Unlike label encoding (where categories are assigned numerical values like 1, 2, 3), one-hot encoding does not assume any inherent order or ranking among categories. This is important because, in many cases, the categories do not have a natural order (e.g., colors or fabrics).
    - Interpretability: Each category is represented as a distinct binary column, making it easier to interpret the importance of each category when looking at model coefficients or feature importance.

5.2 Model Development and Evaluation

The second phase of the research involved developing and evaluating machine learning models to predict restocks and perform time series analysis. This section details the results for each model tested and the rationale behind their selection.

5.2.1 Predicting Restocks Using Sales Data

Various machine learning models were analyzed and tested on the dataset, with different preprocessing steps applied to improve model performance. The models evaluated include Linear Regression, Random Forest, Decision Tree, Support Vector Machine (SVM), and a Basic Neural Network. These models were tested on both scaled and non-scaled data. The sales data (y) exhibited significant skewness, which could negatively impact model performance. To address this, a log transformation was applied to the target variable, reducing the impact of outliers and making the distribution of the target variable more normal.

The results were evaluated using standard performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2).

Model Evaluation

The following is a detailed analysis of the results, which are visualized in the performance comparison plot in Figure X.

- **Linear Regression**: The performance of linear regression was significantly impacted when using the log-transformed and scaled data, resulting in extremely high MSE and RMSE values. This suggests that linear regression is highly sensitive to the target variable transformation and performs poorly with skewed data. Linear regression performed slightly better on the original and scaled datasets, but overall, it underperformed compared to more complex models.
- **Random Forest**: The Random Forest model demonstrated consistent performance across all datasets, achieving lower MSE, RMSE, MAE, and MAPE values compared to linear regression. It also showed a higher R^2 value, indicating that it captured the variability in the data more effectively. The model performed best on the log-transformed and scaled dataset, suggesting that Random Forest benefits from normalization and log transformation.
- **Decision Tree**: The Decision Tree model performed reasonably well on the scaled and original datasets, but it struggled with the log-transformed dataset, as indicated by its higher MSE and lower R^2. The model's performance lagged behind Random Forest, possibly due to the fact that individual decision trees are prone to overfitting, whereas Random Forest mitigates this issue through ensembling.
- **Support Vector Machine (SVM)**: The SVM model's performance varied significantly depending on the dataset used. On the log-transformed and scaled dataset, it achieved lower MSE and MAE, indicating improved prediction accuracy. However, SVM required scaling to perform optimally, and its results were less competitive on the original dataset. The R^2 value for the log-transformed and scaled dataset was higher, demonstrating that SVM could effectively model the complex relationships in the data when the right preprocessing was applied.
- **Basic Neural Network**: The ANN model, which was applied using a basic architecture, achieved competitive results across all datasets. The model showed significant

improvements in performance on the scaled and log-transformed datasets, particularly in terms of lower MAE and MSE. The higher R² values for the ANN on the transformed datasets suggest that the model was able to generalize better when the data was normalized and log-transformed.

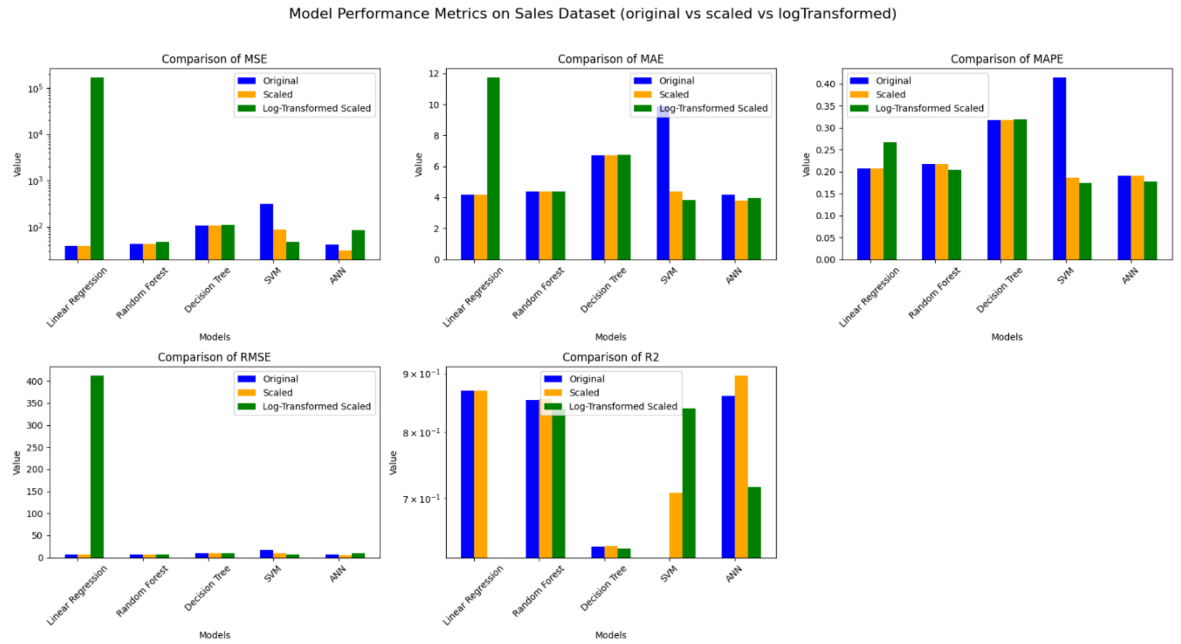


Figure 5.5 - Restock predictions using sales dataset, model comparison

Results and Observations

- **Impact of Scaling**: Scaling the input features improved the performance of models like SVM and the neural network, which rely on the scale of data. For tree-based models like Random Forest and Decision Tree, scaling did not have a significant impact on performance.
- **Log Transformation**: Log-transforming the target variable improved model performance across the board by reducing the skewness in the sales data. This was particularly beneficial for models like Linear Regression and SVM, which assume normally distributed data.
- **Model Comparison**: Among the models tested, Random Forest consistently outperformed the others, providing the best balance between interpretability and predictive power. The neural network also performed well but required more tuning and computational resources compared to Random Forest.

Hyperparameter Tuning and Next Steps

Based on the initial results, the next step involved fine-tuning the SVM and ANN models to improve their performance further. Given the computational complexity of SVM, the tuning process posed challenges, particularly with the GPU. To address this, Support Vector Regressor (SVR) from

ThunderSVM was employed (Abdullah, 2020), which required significant GPU resources. The SVR model was tuned using various kernel functions, such as RBF and polynomial kernels, to improve its ability to capture non-linear relationships in the data. Despite the computational challenges, the final tuned SVR model showed promising results, improving its performance on the log-transformed and scaled dataset.

For the Artificial Neural Network (ANN), hyperparameter tuning was performed using Keras Tuner. The tuning focused on optimizing parameters such as the number of neurons, learning rate, batch size, and dropout rate. The Keras Tuner allowed for automated searching through a wide range of hyperparameters, resulting in a significantly improved model. The final tuned ANN model demonstrated strong performance, particularly in minimizing MSE and MAE while achieving higher R^2 values, suggesting better generalization.

Results and conclusions

After evaluating both models, we can conclude the following:

- SVR outperformed ANN in terms of MSE, RMSE, and MAE. It demonstrated slightly better predictive accuracy and error minimization, making it the preferred model for predicting restocks in this particular dataset.
- ANN still provided competitive results, particularly in terms of R^2 and MAPE. Despite its slightly higher error metrics, ANN's performance was comparable to SVR, making it a viable alternative, especially when using neural networks is preferable for other reasons (such as scalability or interpretability with more complex data).

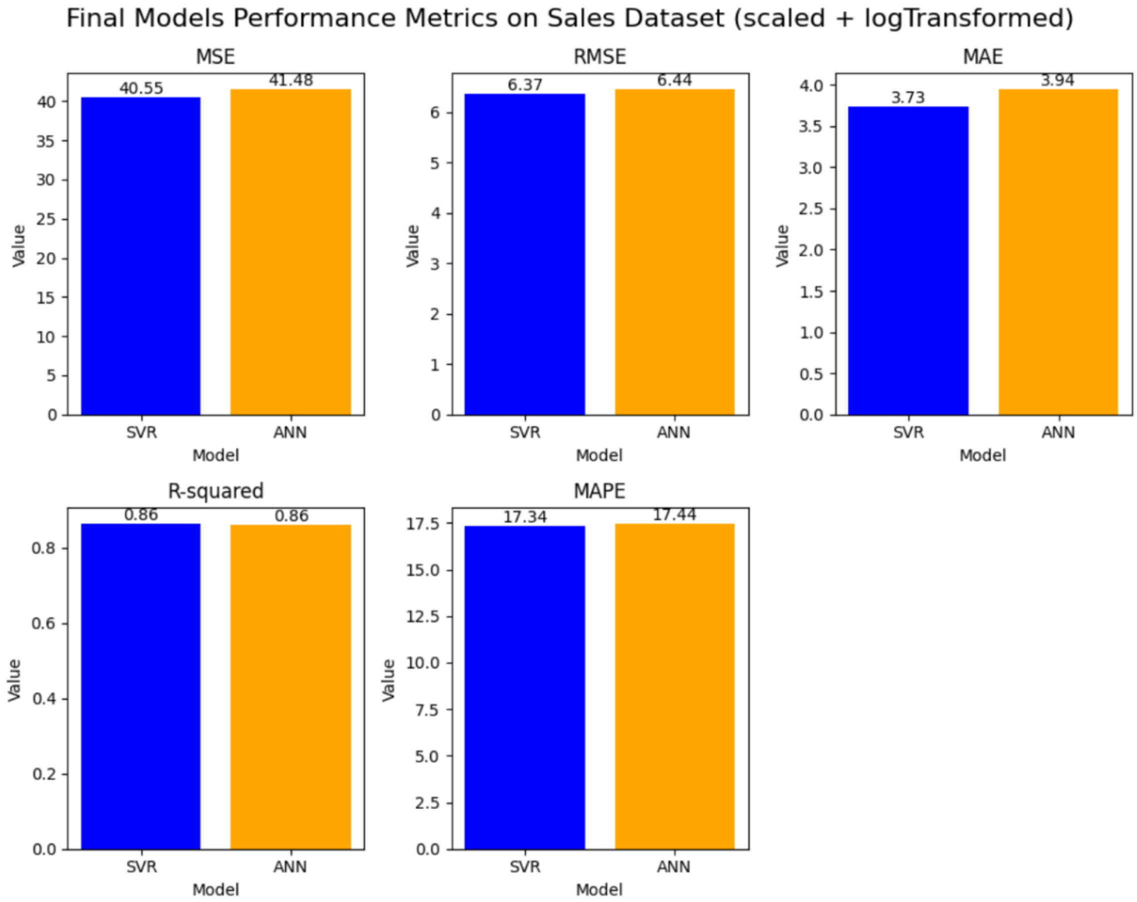


Figure 5.4 - Restocks prediction using sales dataset, final models' comparison

When plotting the actual vs predicted values (Figure 5.5), both SVR and ANN show a similar trend: there is strong clustering around the actual values the deviation increases for larger restock amounts, meaning that both models struggle with larger restock predictions.

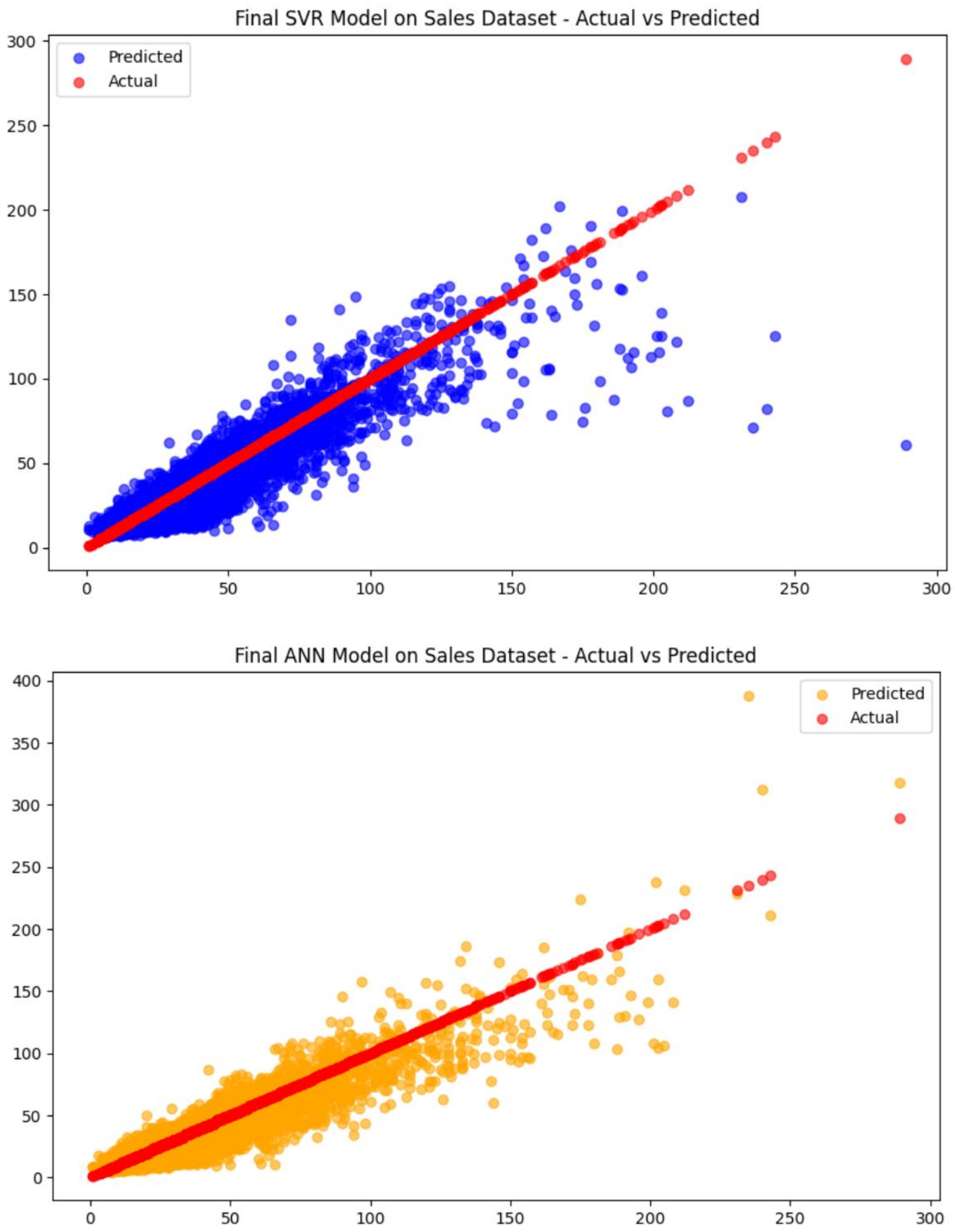


Figure 5.5 - Restocks prediction using sales dataset, comparing SVR and ANN predictions

Based on the performance metrics and error analysis, the SVR model is chosen as the best model for this use case, as it provides better overall accuracy and minimizes errors slightly more effectively than the ANN model. Both models, however, offer strong predictive capabilities and were fine-tuned effectively through hyperparameter tuning.

5.2.2 Time Series Analysis Using Restocks Data

Time series analysis was conducted using the restocks dataset to identify temporal patterns and improve forecasting accuracy.

The objective was to model restocking behavior over time and forecast future restock needs. However, several challenges were encountered due to data limitations, particularly the reduction in data size after aggregation and missing data during the COVID-19 period.

Data Aggregation by Product ID

The first step in the time series analysis was to aggregate the restock data by product ID (external_code). This aggregation combined the restock quantities for each product over time. While this aggregation was necessary for organizing the data and ensuring product-level consistency, it had a significant impact on the data's granularity. The aggregation substantially reduced the size of the dataset, which was less than ideal for time series analysis. A smaller dataset limits the ability to capture long-term trends and seasonal patterns accurately, making it challenging for the models to generate reliable forecasts.

After aggregating the data, the time series of restock quantities was plotted. This initial plot revealed significant gaps in the data, particularly during the COVID-19 period, where data for multiple years was missing. These missing data points were problematic, as they disrupted the continuity of the time series, making it difficult for models to learn patterns and trends effectively.

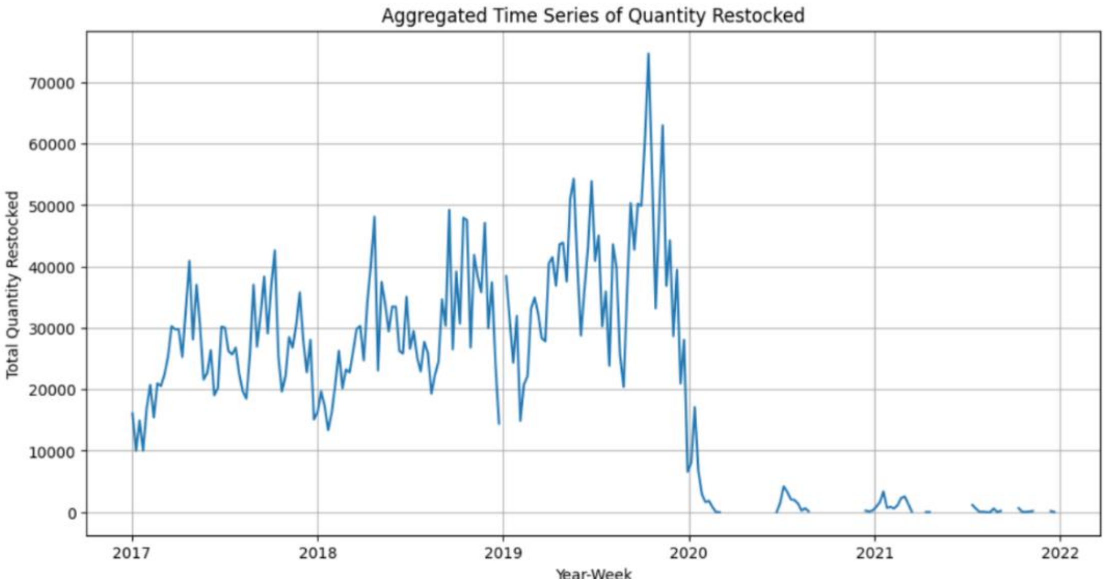


Figure 5.6 - Restocks time series

First Approach: Removing COVID-19 Years

In the first approach, the missing data during the COVID-19 years (2020-2021) was addressed by removing these years entirely from the dataset. The objective was to mitigate the impact of missing data on the forecasting models. However, this reduction in data significantly affected the time series, leading to suboptimal performance of the models. The steps followed in this approach are detailed below.

Step 1: Transforming the Data into a Time Sequence

After performing initial explorations and handling missing data, the dataset was transformed into a time sequence to prepare it for time series modeling. This involved creating lag features for the restock quantities (qty). Those features were generated using time steps of 1, 2, and 5 days. This step is essential for time series analysis as it allows the model to use historical data to predict future values.

Step 2: Checking for Tend and Seasonality

Once the COVID-19 years were removed, the next step was to check for any existing trends or seasonality in the reduced dataset. This was done by decomposing the time series into its components: trend, seasonality, and residuals. Some patterns were observed, as expect for the nature of the fast fashion business.

Step 3: Preparing Data for Modeling

Once the lag features were generated, the next step was to split the data into training and testing sets.

- Feature Selection: The features used for modeling were the lagged restock quantities: qty_lag_1, qty_lag_3, and qty_lag_5. These features served as input variables (X) for the time series models, while the target variable (y) was the current restock quantity (qty).
- Train-Test Split: The data was split into training and testing sets using an 80/20 ratio. This ensures that the model is trained on the majority of the data but is evaluated on a separate portion of the data to assess its generalization ability. The train_test_split function was used with shuffling disabled to maintain the chronological order of the time series.

Step 4: Models Evaluation

After preparing the data, three time series models were evaluated: SARIMA, ARIMA, and Exponential Smoothing. These models were compared based on their ability to forecast future restock quantities. The model with the lowest RMSE was SARIMA, that was chosen as the best one for the next step: hyperparameter tuning. Moreover, SARIMA accounts for seasonality, so it may be preferable as seasonality is important in this dataset. SARIMA accounts for seasonality, so it may be preferable as seasonality is important in this dataset.

The forecast plot in Figure 5.7 revealed discrepancies between predicted and actual restock quantities, further highlighting the difficulties encountered due to the small dataset and missing data during the COVID-19 period.

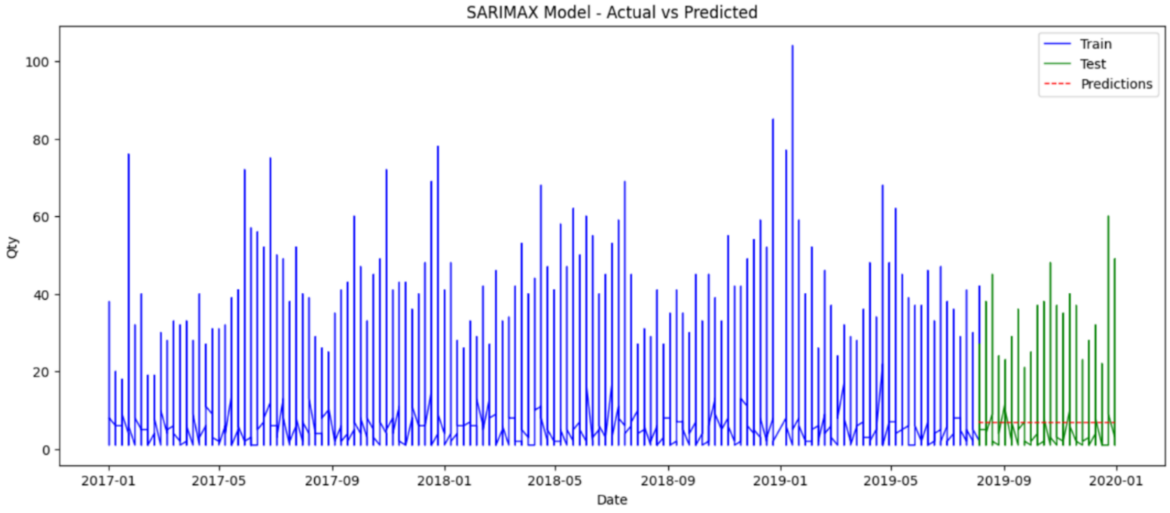


Figure 5.7 - SARIMAX predictions

Second Approach: Imputation of Missing Data

In the second approach, instead of removing the COVID-19 years, which significantly reduced the size of the dataset, the missing data during the pandemic years (2020-2021) was addressed through imputation techniques. The goal was to restore the continuity of the dataset and improve the models' ability to predict future restock quantities. After imputing the missing data, the SARIMA model and other models were re-evaluated to see if the imputed data could lead to more accurate predictions. Additionally, weighting was applied to the dataset (Amat, 2014) to account for the missing periods, which required using alternative libraries as standard ones (like statsmodels in Python) do not support weighted fitting for SARIMA models.

Since SARIMA could not be used with the weighted dataset, three alternative models—Linear Generalized Additive Model (GAM), XGBoost, and LightGBM—were selected for evaluation. These models were chosen for their ability to handle the complexity of the time series data and work effectively with weights.

After fitting and evaluating the three models—Linear GAM, XGBoost, and LightGBM—the results were visualized using performance metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R² score. These metrics were used to compare the models' forecasting abilities on the restocks dataset after imputation and weighting.

Despite the efforts to impute missing data and apply weights, the prediction accuracy of the three models—Linear GAM, XGBoost, and LightGBM—was not as high as expected. Figure X below illustrates the comparison of the models' performance:

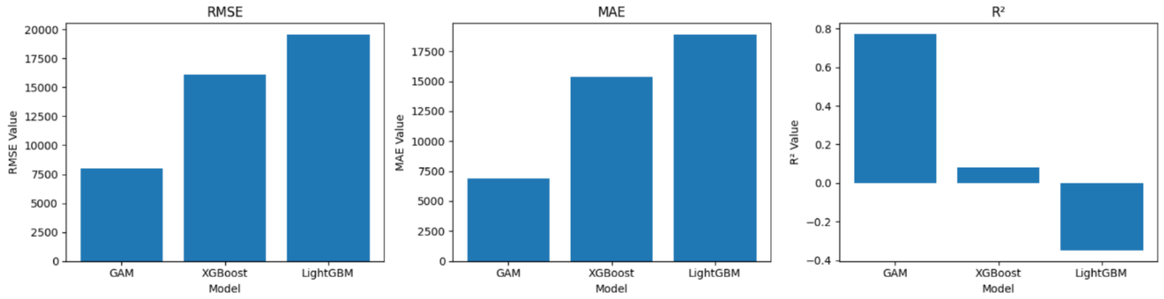


Figure 5.8 - Weighted models' performance metrics comparison

Linear GAM emerged as the best-performing model across all metrics (RMSE, MAE, and R²). It showed superior accuracy in predicting restock quantities, with smaller errors and a higher ability to explain the variance in the data. This is likely due to GAM's flexibility in capturing non-linear relationships, which seemed to be a good fit for the nature of the time series data.

Even after the missing data was imputed and weights were applied, and hyperparameters have been tuned, the model still struggled to provide accurate predictions. This was evident from the prediction plots, which didn't show a significant deviation from the original values of the Covid years. The combination of limited data points and imputation techniques did not yield reliable forecasts, likely due to the complexity and irregularity of restocking patterns during the COVID-19 period.

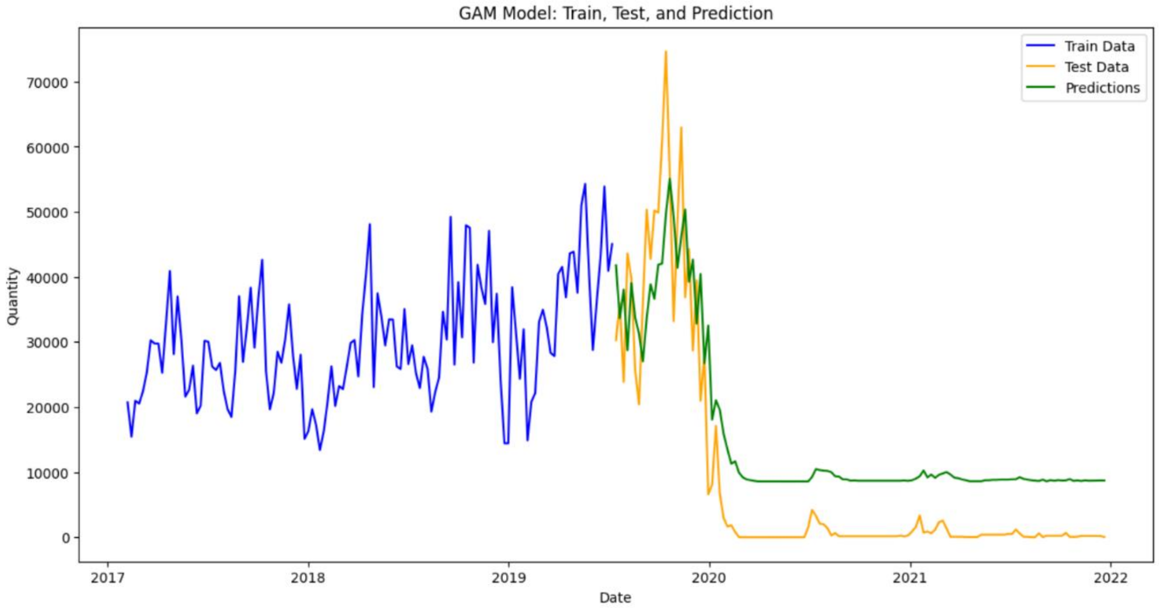


Figure 5.9 - Predictions of best weighted model: GAM

5.3 Multimodal Analysis

The third phase of the research involved developing a multimodal model that combined sales data with product images to classify items into low, medium, and high-selling categories. The goal was to use both numerical sales data and visual features extracted from product images to improve the accuracy of product sales classification.

5.3.1 Data Preprocessing and Feature Extraction

5.3.1.1 Sales Data Preprocessing

The sales data was preprocessed to ensure it was in a format suitable for integration with the image features. Key preprocessing steps included:

- **Data Aggregation**: The sales data was aggregated per product ID (external_code) to consolidate the restock and sales information into a product-level dataset.
- **Encoding Categorical Variables**: Categorical variables were encoded to prepare them for model training. This included variables such as product category, which were transformed into numerical representations using techniques like one-hot encoding.
- **Sales Category Creation**: Based on the aggregated sales data, products were classified into three categories: low, medium, and high-selling. This classification was essential for the

supervised learning task and served as the target variable for both the sales and image-based models.

5.3.1.2 Image Data Processing

For the image component, the ResNet50 model (TensorFlow, n.d.) was used to extract features from the product images. ResNet50, a pre-trained convolutional neural network, was chosen for its ability to capture complex visual features effectively (Garg, 2022). The steps for image processing included:

- **Loading Pre-trained ResNet50**: The ResNet50 model, pre-trained on the ImageNet dataset, was loaded with weights frozen to prevent further training on the current dataset.
- **Feature Extraction**: The output from the last convolutional layer of ResNet50 was used as the feature vector representing each image. These feature vectors captured high-level visual information about each product, such as color, texture, and shape.
- **Flattening Features**: The extracted features were flattened into a one-dimensional vector, making them suitable for integration with the sales data.

Luckily, no missing values in the image dataset were found and therefore there was no need to use data augmentation techniques, to create synthetic images based on existing ones to fill in gaps and improve model robustness.

The sales and image datasets were merged based on the product ID (external_code). This integration allowed for a comprehensive analysis that considered both historical sales patterns and visual attributes of the products.

5.3.2 Development of the Multimodal Model

The multimodal model was designed to incorporate both the processed sales data and image features (Point 8 Insights, 2022). The architecture of the models was based on two distinct inputs: one for the image features extracted using a pre-trained model (such as ResNet50) and another for the sales data. The goal was to allow the model to learn both visual and numerical patterns to improve the accuracy of sales category classification.

Three different architectures were tested, each varying in complexity and the number of layers and neurons. The following subsections describe these models in detail.

Model 1: Baseline Model (Simplest Architecture)

The first model implemented is a baseline architecture designed to be simple and lightweight, allowing for a fast initial exploration of the multimodal setup. It uses a small number of neurons and no dropout layers. This model serves as a control to compare the effects of increasing the model's depth and complexity.

- **Inputs**:
  - The model takes two inputs: one for image features and one for sales data.
  - image_input receives the features extracted from the ResNet50 model, while sales_input takes the preprocessed sales data.
- **Layers**:
  - The image input is passed through a dense layer with 128 neurons and a ReLU activation function.
  - The sales input is processed through a dense layer with 64 neurons and a ReLU activation function.
- **Concatenation**:
  - The outputs of these two dense layers are concatenated to form a combined feature set that incorporates both image and sales data.
- **Output**:
  - A final dense layer with 3 neurons and a softmax activation function is applied to produce the output, which classifies the products into low, medium, and high-selling categories.
- **Compilation**:
  - The model is compiled using the Adam optimizer, with sparse categorical cross-entropy as the loss function and accuracy as the primary evaluation metric.

This simplest architecture provides a quick starting point, allowing the model to learn basic interactions between the two inputs. However, due to its limited capacity, it may not be sufficient for learning more complex patterns.

Model 2: Deeper Model (More Layers and Neurons)

The second model introduces additional layers and neurons to potentially capture more complex relationships between the image and sales data. It also includes dropout layers to prevent overfitting.

- **Inputs**:
    - Like the baseline model, this architecture accepts two inputs: one for image features and one for sales data.
- **Layers**:
    - The image input is processed through a denser layer with 256 neurons, and the sales input is passed through a layer with 128 neurons, both using ReLU activation.
- **Dropout**:
    - To mitigate overfitting, dropout layers are introduced after the dense layers, with a dropout rate of 0.5. This randomly drops 50% of the neurons during training, forcing the model to generalize better.
- **Concatenation**:
    - After the dropout layers, the features from both inputs are concatenated.
- **Additional Dense Layer**:
    - An additional dense layer with 128 neurons is applied after the concatenation to further refine the combined features before the final output layer.
- **Output**:
    - A final dense layer with 3 neurons and a softmax activation function outputs the classification into low, medium, and high-selling categories.
- **Compilation**:
    - The model is compiled using the Adam optimizer, with sparse categorical cross-entropy as the loss function and accuracy as the metric.

By adding more neurons and layers, this model aims to capture more nuanced relationships between the image and sales data, which the baseline model might miss. The inclusion of dropout layers helps the model avoid overfitting, especially when dealing with limited data.

Model 3: Wide Model (Wider Layers with More Neurons)

The third model increases the capacity of the architecture by using wider layers with more neurons. Like the deeper model, it includes dropout layers, but the dense layers have significantly more neurons, making the model capable of learning more complex relationships between the data.

- **Inputs**:
    - The architecture is similar in structure to the previous models, with two inputs: one for the image features and one for the sales data.
- **Layers**:

    - The image input is processed through a dense layer with 512 neurons, and the sales input is passed through a dense layer with 256 neurons, both with ReLU activation.
- **Dropout**:
    - Dropout layers with a 0.5 rate are applied to both dense layers to prevent overfitting, as the model capacity is significantly larger.
- **Concatenation**:
    - The output from the dropout layers is concatenated to combine the information from the image and sales data.
- **Additional Dense Layer**:
    - A dense layer with 256 neurons is added after the concatenation to further process the combined features.
- **Output**:
    - The final dense layer has 3 neurons and a softmax activation function, classifying the products into low, medium, and high-selling categories.
- **Compilation**:
    - The model is compiled using the Adam optimizer, with sparse categorical cross-entropy as the loss function and accuracy as the metric.

5.3.3 Model Training and Evaluation

The three architectures (baseline, deeper, and wide models) were trained using the preprocessed sales and image data. Each model was evaluated using classification metrics to determine its effectiveness in predicting whether a product belongs to the low, medium, or high-selling categories.

Training Process

The following steps were taken during the training process:

- **Training and Validation Split**: The dataset was split into training and validation sets, with a typical split ratio of 80% training data and 20% validation data. This split ensured that the models were evaluated on unseen data during training to prevent overfitting and to assess how well the models generalized.
- **Batch Size and Epochs**: The models were trained using a batch size of 32, which was chosen to balance between memory efficiency and model convergence. The models were trained

for a set number of epochs (typically 50 or more), with early stopping applied to prevent overfitting if validation performance stopped improving after several epochs.

- **Loss Function and Optimizer**: All models were compiled with the Adam optimizer, which is well-suited for deep learning tasks due to its adaptive learning rate. The loss function used was sparse_categorical_crossentropy, which is appropriate for multi-class classification tasks where the target variable (sales category) is represented as integer labels.

Models' evaluation and selection

The models were evaluated on their ability to classify products into low, medium, and high-selling categories. The performance metrics analyzed included Test Loss and Test Accuracy. The figure 5.10 below shows a breakdown of the results:

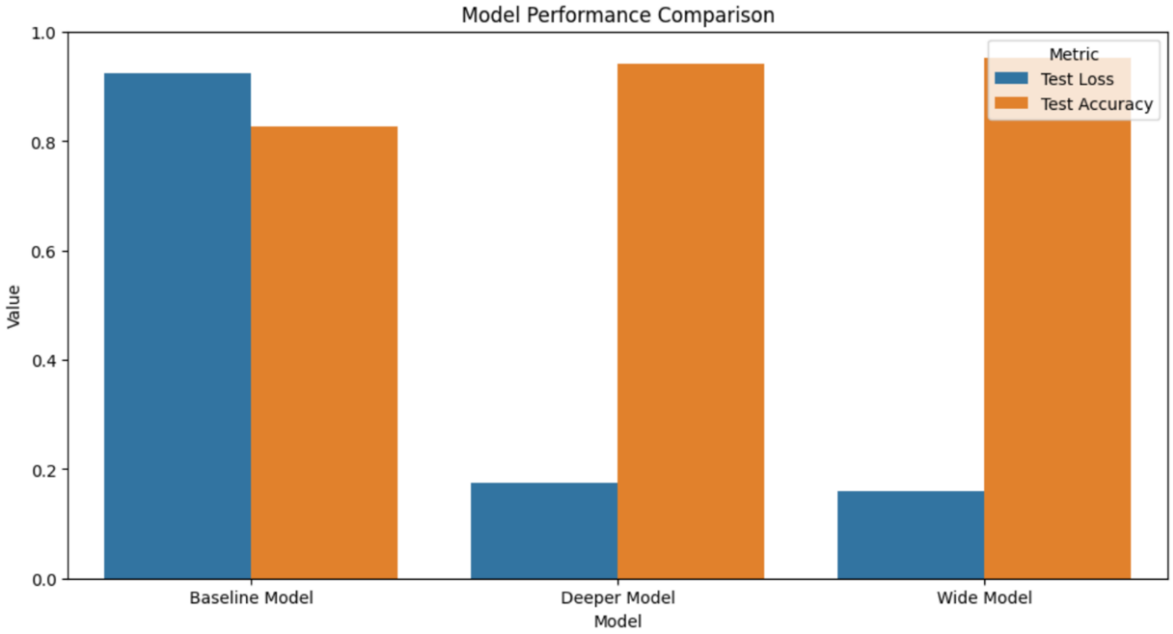


Figure 5.10 - Multimodal image classification, models' comparison

- **Baseline Model**: The baseline model performed moderately well with a test accuracy of 82.6% but had a high test loss of 0.92. This indicates that while it could make correct predictions to some extent, it had trouble minimizing errors, leading to suboptimal learning. Given its simplicity, this result is expected and demonstrates the model's limited capacity to capture complex relationships in the multimodal data.
- **Deeper Model**: The deeper model significantly improved on the test loss (0.175), reducing errors substantially compared to the baseline model. It also achieved a much higher test accuracy of 94.1%. The addition of layers and dropout helped the model generalize better,

and the reduced test loss shows that this model learned more robust representations of the combined image and sales data. This model is particularly effective when minimizing prediction errors is the primary goal.

- **Wide Model**: The wide model outperformed both the baseline and deeper models in terms of test accuracy, achieving 95.3%. However, the test loss of the wide model (0.160) was only marginally better than the deeper model. This indicates that the wide model made slightly more correct predictions but at a similar level of error minimization compared to the deeper model. The wider layers allowed the model to capture more complex patterns between the sales and image data, making it the top performer in terms of overall accuracy.

Choosing the Best Model

Based on the evaluation results, the decision on the best model depends on the primary goal of the classification task:

- **Deeper Model**: If minimizing errors (lowering the test loss) is more important, the deeper model is the best choice. It achieves a very low test loss of 0.175, indicating that it is better at reducing the model's overall error in predicting the correct categories. This model is useful in scenarios where making fewer incorrect predictions is crucial.
- **Wide Model**: If maximizing correct predictions (higher accuracy) is more important, the wide model is the optimal choice. With a test accuracy of 95.3%, the wide model provides the highest rate of correct classifications. This model is preferred in situations where the number of correct predictions is more critical than minimizing small errors.

In this study, we prioritize accuracy over test loss, meaning that the model with the highest accuracy should be selected. Therefore, the Wide Model is chosen as the best-performing model for the multimodal classification task. Despite its marginally higher test loss compared to the deeper model, the wide model excels in making more correct predictions, which aligns with our goal of maximizing classification accuracy. Both the deeper and wide models were hyperparameter-tuned and trained to optimize their performance, but for further analysis, the wide model will be the focus of our future work.

Evaluation Metrics

The final models were evaluated using several classification metrics to assess their performance. Figure 5.11 compares the performance of the two best models (Wide Model and Deeper Model) across four key metrics: Precision, Recall, F1 Score, and Accuracy. Both models were tuned for optimal hyperparameters and evaluated on the test dataset.

- **Precision**: The precision values for both the Wide Model and Deeper Model are almost identical, indicating that both models were effective at minimizing false positives. This suggests that both models are reliable when predicting which products are in the low, medium, or high-selling categories.
- **Recall**: Similar to precision, the recall values for both models are very close, indicating that both models performed well in identifying the correct selling categories without missing too many positive instances (low, medium, high-selling products).
- **F1 Score**: The F1 Score, which balances both precision and recall, shows a similar trend for both models, indicating that they are both effective at achieving a balance between precision and recall.
- **Accuracy**: The accuracy metric shows only a slight improvement for the Wide Model over the Deeper Model, confirming that the Wide Model made slightly more correct predictions overall. However, the difference in accuracy is minimal, and both models demonstrate strong classification performance.

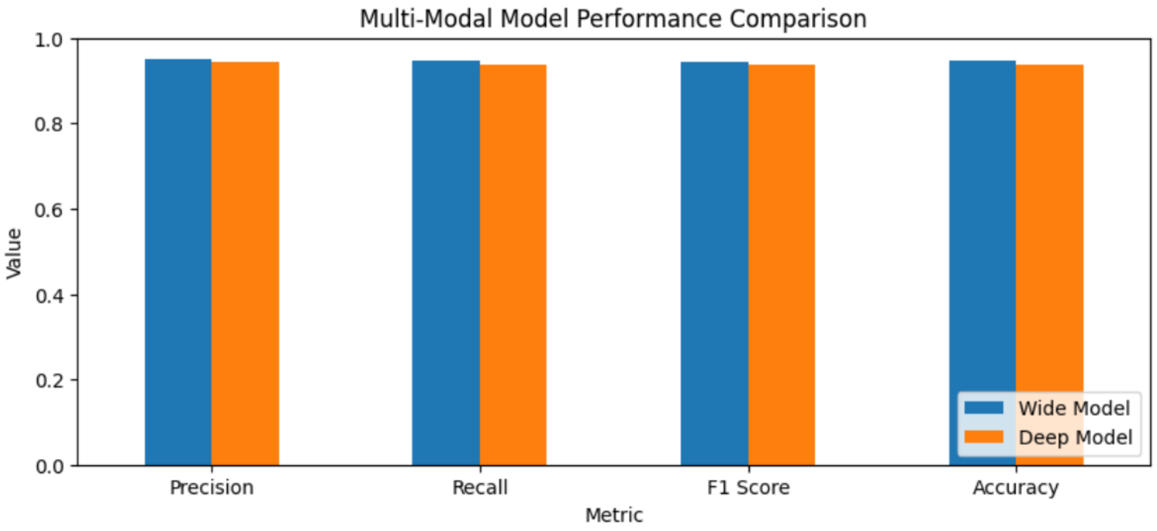


Figure 5.11 - Multimodal image classification, models' performance metrics

While both the Wide Model and Deeper Model performed similarly across all metrics, the Wide Model had a marginal edge in terms of accuracy. This makes the Wide Model the preferred choice when the goal is to maximize correct predictions. A confusion matrix has been plotted for each model (Figure 5.12) to confirm the results.

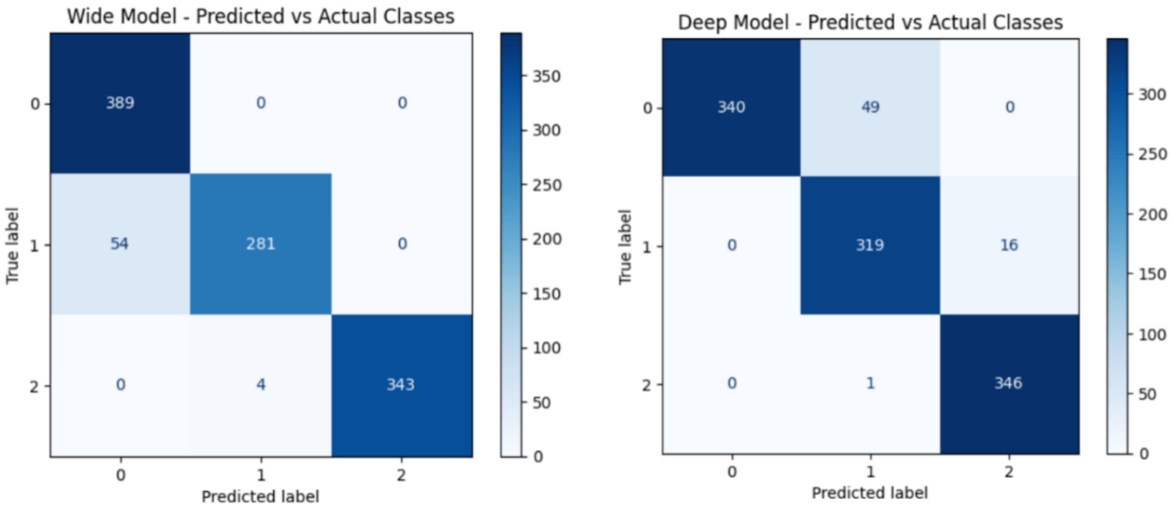


Figure 5.12 - Classification comparison of Wide and Deep Model

5.4 Challenges and Justifications

Throughout the research, several challenges were encountered that influenced the outcomes and decisions made:

- **Data Aggregation**: Aggregating the sales and restocks data by product ID reduced the granularity of the dataset, which impacted the model's ability to learn from fine-grained variations. This challenge was mitigated by focusing on more general trends and patterns, which are still valuable for inventory management in a fast fashion context.
- **Data Quality and Availability**: The datasets provided by the company contained not enough data points for a time series as well as dates inconsistencies, which required extensive cleaning and preprocessing. While this step was time-consuming, it was necessary to ensure the reliability of the models and the validity of the research findings.
- **Model Complexity**: The choice of models was influenced by the trade-off between complexity and interpretability. While advanced models like neural networks offered the potential for higher accuracy, they also required more computational resources and were less interpretable.
- **Integration of Multimodal Data**: Combining sales data with image data posed technical challenges, including aligning different data types. Despite these challenges, the multimodal model demonstrated the value of integrating diverse data sources to enhance predictive accuracy.

The results of this research highlight the potential of data analytics in improving inventory management and reducing waste in the fast fashion industry. By leveraging machine learning models, time series analysis, and multimodal approaches, the study provides a comprehensive framework for predicting restocks, understanding sales trends, and classifying products based on sales performance and visual attributes. The findings underscore the importance of data-driven decision-making in achieving sustainable practices in fast fashion and provide a foundation for future research in this area.

6. Discussion

This chapter discusses the findings of the study in relation to the existing literature and the research objectives outlined earlier. It evaluates the implications of using data analytics to improve inventory management and reduce waste in the fast fashion industry, identifies the strengths and weaknesses of the approaches used, and suggests potential directions for future research.

6.1 Interpretation of Findings

The study aimed to explore the application of data analytics in minimizing waste and optimizing inventory management in the fast fashion industry. The results demonstrated that leveraging machine learning models, time series analysis, and multimodal approaches can provide significant insights and improvements in managing inventory more effectively.

6.1.1 Alignment with Existing Literature

This study adds to an extensive collection of research on the application of machine learning and data analytics to inventory management in the fast fashion sector. Prior research, such that done by (Caro and Gallien, 2010), highlighted the importance of machine learning in accelerating the supply chains of fast fashion by using predictive models to optimize stock distribution across retail networks. Similarly, (Giri, Thomassey, and Zeng, 2019) emphasized the relevance of data-driven insights when managing customer demand, demonstrating analytics' potential to minimize waste and improve decision-making.

This research provides a novel multimodal strategy to categorize items based on sales performance by merging product photos with sales data, however many studies have mainly relied on sales and restock data for predictive modeling. This strategy differs from the more popular emphasis on consumer purchasing patterns, as demonstrated by research such as (Nguyen, Le, and Ho, 2020), who used consumer data to segment markets and forecast sales patterns. While consumer data offers insightful information, it was purposefully left out of this study in order to emphasize the effectiveness of product-level analytics on its own. This choice was also reinforced by the ethical implications and data privacy issues pertaining to consumer data, which (Oh, 2020) brought up in the study of big data in retail.

Other studies, such as (Silva, Hassani, and Madsen, 2020), have shown the effectiveness of big data techniques in fast fashion, but predominantly through numerical data. This research expands on those findings by incorporating Convolutional Neural Networks (CNNs) to analyze product images, a method commonly used in other sectors but less so in fast fashion. The multimodal model created in this study is in accordance with recent developments in machine learning and artificial intelligence, such those investigated by (Thomassey and Fiordaliso, 2006), who used decision tree and clustering techniques to estimate sales. However, the introduction of visual data in this project adds a new layer of complexity and accuracy to product classification, filling a gap in the literature that has primarily focused on consumer data or historical sales trends.

In summary, this research complements existing work by demonstrating that visual attributes—such as product color, texture, and style—can be valuable predictors in sales performance, offering a fresh perspective on data analytics in fast fashion that goes beyond the traditional reliance on consumer data or purely numerical metrics.

6.1.2 Contribution to the Field

This study contributes to the field by not only confirming the utility of data analytics in fast fashion but also introducing a multimodal approach that combines sales data with product images. The use of a multimodal model represents a novel application in the context of inventory management, where traditionally, data-driven decision-making has relied heavily on numerical data. By integrating visual data into the predictive models, this study has shown that combining diverse data types can enhance the accuracy of sales predictions and provide a more holistic understanding of product performance.

The effective use of a multimodal model to classify things into low, medium, and high-selling categories suggests that visual characteristics like color and style are important indicators of product popularity.

Businesses can reduce waste and enhance sustainability by using the visual qualities of items to inform their decisions about product development and restocking. This research, therefore, contributes not only to inventory management practices but also to the broader discourse on sustainability in retail by demonstrating how data analytics can align business operations with eco-friendly practices.

6.2 Implications for the Fast Fashion Industry

The findings from this study have several practical implications for the fast fashion industry.

6.2.1 Improving Inventory Management

The integration of data analytics into inventory management practices can greatly enhance the ability of fast fashion companies to predict demand and align production with actual market needs. By employing advanced machine learning models, companies can move away from reactive restocking strategies toward more proactive and predictive approaches. This shift can lead to more efficient inventory management, reduced overproduction, and lower inventory costs, ultimately contributing to a more sustainable business model.

The study's findings suggest that the adoption of data-driven models could enable fast fashion companies to better anticipate demand fluctuations and adjust their inventory accordingly. This is particularly important in a highly dynamic market where trends can change rapidly, and consumer preferences are increasingly volatile. By leveraging data analytics, companies can reduce the risk of stockouts and overstocking, improving both customer satisfaction and financial performance.

6.2.2 Enhancing Sustainability

Sustainability has become a critical issue for the fast fashion industry, which is often criticized for its environmental impact. This study demonstrates that data analytics can play a pivotal role in supporting sustainable practices by reducing waste and optimizing resource use. By accurately predicting restocks and classifying products based on sales potential, companies can minimize excess inventory and prevent unsold items from ending up in landfills.

Furthermore, the multimodal approach used in this study highlights the potential for integrating sustainability considerations into product design and merchandising decisions. By understanding which visual attributes are associated with high sales, companies can make more informed decisions about product development, focusing on designs that are more likely to sell and thus reduce waste.

6.3 Future Research Directions

Based on the findings and limitations of this study, several avenues for future research are suggested:

1. **Exploring Other Machine Learning Techniques**: Future studies could explore additional machine learning models and techniques, such as reinforcement learning, to further enhance the predictive accuracy of inventory management models. These models could be particularly useful for optimizing dynamic pricing strategies or managing promotional campaigns in the fast fashion industry.
2. **Incorporating Real-Time Data**: The integration of real-time data, such as social media trends or online search behavior, could provide valuable insights into emerging fashion trends and consumer preferences. Future research could investigate the use of real-time analytics and streaming data to improve demand forecasting and inventory management in a rapidly changing market.
3. **Expanding the Scope of Multimodal Analysis**: While this study focused on combining sales data with product images, future research could explore other combinations of data modalities, such as text analysis of customer reviews or sentiment analysis of social media posts. These additional data sources could provide a more comprehensive understanding of consumer behavior and preferences, further enhancing the accuracy of predictive models.
4. **Testing in Different Retail Contexts**: The findings of this study are specific to the fast fashion industry, but the methodologies and models developed could be applied to other retail contexts, such as electronics, groceries, or home goods. Future research could test the generalizability of the models and explore their applicability across different retail sectors to identify common patterns and challenges.
5. **Developing Explainable AI Models**: Given the importance of interpretability in practical applications, future research could focus on developing explainable AI models that provide transparent and understandable predictions. This could involve the use of techniques such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) to enhance model transparency and build trust with decision-makers (www.markovml.com, n.d.).

7. Conclusion

The fast fashion industry, characterized by rapid production cycles and high inventory turnover, faces significant challenges in optimizing inventory management and reducing waste. This study explored the application of data analytics, particularly machine learning models and multimodal analysis, to address these issues. By leveraging the Visuelle 2.0 dataset, the research demonstrated that advanced machine learning techniques significantly enhance predictive accuracy for inventory restocking, outperforming traditional methods. Moreover, integrating sales data with product images improved classification accuracy of sales performance, highlighting the value of multimodal data in understanding consumer preferences.

The findings align with existing literature and contribute to the growing body of knowledge on sustainable practices in fashion retail, bringing a novel approach of being able to classify items combined with traditional stock forecasting techniques.

The findings suggest that data-driven approaches can support more efficient inventory management and contribute to sustainability efforts by reducing overproduction and unsold inventory. These results have important implications for fast fashion businesses aiming to align with consumer demands for sustainability while maintaining operational efficiency.

Despite the study's contributions, limitations such as data quality and computational complexity underscore the need for further research, particularly in developing scalable and interpretable models. As the industry evolves, the integration of advanced analytics will be crucial for enhancing decision-making and promoting sustainable practices in fast fashion.

References

Abdullah, S.M. (2020). Introduction to ThunderSVM: A Fast SVM Library on GPUs and CPUs. [online] Available at: https://medium.com/analytics-vidhya/how-to-install-and-run-thundersvm-in-google-colab-de1fe49eef85.

Amat, J. (2014). Weighted time series forecasting - Skforecast Docs. [online] Skforecast.org. Available at: https://skforecast.org/0.10.0/user_guides/weighted-time-series-forecasting [Accessed 29 Sep. 2024].

amsten (2020). *ML | Evaluation Metrics*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/metrics-for-machine-learning-model/.

Aviv, Y. (2003). A Time-Series Framework for Supply-Chain Inventory Management. *Operations Research*, 51(2), pp.210–227. doi:https://doi.org/10.1287/opre.51.2.210.12780.

Bajaj, A. (2021). *Performance Metrics in Machine Learning [Complete Guide]*. [online] neptune.ai. Available at: https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide.

Bradlow, E.T., Gangwar, M., Kopalle, P. and Voleti, S. (2017). The Role of Big Data and Predictive Analytics in Retailing. *Journal of Retailing*, 93(1), pp.79–95. doi:https://doi.org/10.1016/j.jretai.2016.12.004.

Brahmadeep and Thomassey, S. (2016). Intelligent demand forecasting systems for fast fashion. *Information Systems for the Fashion and Apparel Industry*, pp.145–161. doi:https://doi.org/10.1016/b978-0-08-100571-2.00008-7.

Brownlee, J. (2017). Why One-Hot Encode Data in Machine Learning? [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/.

Caro, F. and Gallien, J. (2010). Inventory Management of a Fast-Fashion Retail Network. *Operations Research*, 58(2), pp.257–273. doi:https://doi.org/10.1287/opre.1090.0698.

Chen, I-Fei. and Lu, C.-J. (2021). Demand Forecasting for Multichannel Fashion Retailers by Integrating Clustering and Machine Learning Algorithms. *Processes*, 9(9), p.1578. doi:https://doi.org/10.3390/pr9091578.

Choi, T.-M., Hui, C.-L., Ng, S.-F. and Yu, Y. (2012). Color Trend Forecasting of Fashionable Products with Very Few Historical Data. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6), pp.1003–1010. doi:https://doi.org/10.1109/tsmcc.2011.2176725.

Gardner, E.S. and Diaz-Saiz, J. (2002). Seasonal adjustment of inventory demand series: a case study. *International Journal of Forecasting*, 18(1), pp.117–123. doi:https://doi.org/10.1016/s0169-2070(01)00108-x.

Garg, A. (2022). *Image Classification Using Resnet-50 Deep Learning Model*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2022/09/image-classification-in-stl-10-dataset-using-resnet-50-deep-learning-model/.

Giri, C., Thomassey, S. and Zeng, X. (2019). Customer Analytics in Fashion Retail Industry. *Functional Textiles and Clothing*, pp.349–361. doi:https://doi.org/10.1007/978-981-13-7721-1_27.

Github.io. (2017). *HumaticsLab | Visuelle 2.0*. [online] Available at: https://humaticslab.github.io/forecasting/visuelle.

Hani, A., Al-Obeidat, F., Benkhelifa, E. and Adedugbe, O. (2020). A Framework for Online Social Network Volatile Data Analysis: A Case for the Fast Fashion Industry. *JUCS - Journal of Universal Computer Science*, 26(1), pp.127–155. doi:https://doi.org/10.3897/jucs.2020.008.

Hotz, N. (2024). *What is CRISP DM?* [online] Data Science Project Management. Available at: https://www.datascience-pm.com/crisp-dm-2/.

HumaticsLAB (2022). *visuelle2.0-code/LICENSE.txt at main · HumaticsLAB/visuelle2.0-code*. [online] GitHub. Available at: https://github.com/HumaticsLAB/visuelle2.0-code/blob/main/LICENSE.txt [Accessed 28 Sep. 2024].

keras.io. (n.d.). Keras documentation: KerasTuner. [online] Available at: https://keras.io/keras_tuner/.

Long, X. and Nasiry, J. (2019). Sustainability in the Fast Fashion Industry. *SSRN Electronic Journal*, 1(1).

McNeill, L. and Moore, R. (2015). Sustainable fashion consumption and the fast fashion conundrum: Fashionable consumers and attitudes to sustainability in clothing choice. *International Journal of Consumer Studies*, 39(3), pp.212–222. doi:https://doi.org/10.1111/ijcs.12169.

Mustafa Ayobami Raji, Hameedat Bukola Olodo, Timothy Tolulope Oke, Wilhelmina Afua Addy, Onyeka Chrisanctus Ofodile and Adedoyin Tolulope Oyewole (2024). Real-time data analytics in retail: A review of USA and global practices. *GSC Advanced Research and Reviews*, [online] 18(3), pp.059–065. doi:https://doi.org/10.30574/gscarr.2024.18.3.0089.

Nguyen, H.T., Le, D.M.D., Ho, T.T.M. and Nguyen, P.M. (2020). Enhancing sustainability in the contemporary model of CSR: a case of fast fashion industry in developing countries. *Social Responsibility Journal*, 17(4), pp.578–591. doi:https://doi.org/10.1108/SRJ-03-2019-0108.

Niinimäki, K., Peters, G., Dahlbo, H., Perry, P., Rissanen, T. and Gwilt, A. (2020). The Environmental Price of Fast Fashion. *Nature Reviews Earth & Environment*, 1(4), pp.189–200. doi:https://doi.org/10.1038/s43017-020-0039-9.

NVIDIA Developer. (2012). CUDA GPUs. [online] Available at: https://developer.nvidia.com/cuda-gpus.

Oh, K. (2020). The roles of data analytics in the fashion industry. *Journal of Textile Engineering & Fashion Technology*, 6(3). doi:https://doi.org/10.15406/jteft.2020.06.00237.

Point 8 Insights (2022). *Training a Convolutional Neural Network with Multiple Input Features*. [online] Available at: https://medium.com/@Point8/training-a-convolutional-neural-network-with-multiple-input-features-c34ce56f32b5.

Scikit-Learn (2019). *sklearn.preprocessing.StandardScaler — scikit-learn 0.21.2 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

Sethi, M.H. (2021). Sindhi Patchwork, Artisans and Fashion Industry. *Journal of Textile Science & Fashion Technology*, 7(5). doi:https://doi.org/10.33552/jtsft.2021.07.000673.

Silva, E.S., Hassani, H. and Madsen, D.Ø. (2020). Big Data in fashion: Transforming the Retail Sector. *Journal of Business Strategy*, 41(4).

Skenderi, G., Joppi, C., Denitto, M., Scarpa, B. and Cristani, M. (2022). The multi-modal universe of fast-fashion: the Visuelle 2.0 benchmark. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. doi:https://doi.org/10.1109/cvprw56347.2022.00245.

Stringer, T., Mortimer, G. and Payne, A.R. (2020). Do ethical concerns and personal values influence the purchase intention of fast-fashion clothing? *Journal of Fashion Marketing and Management: An International Journal*, [online] 24(1), pp.99–120. doi:https://doi.org/10.1108/JFMM-01-2019-0011.

Surabani, S. and Rodriguez, B. (2023). Data Analytics Application in Fashion Retail SMEs (A Case Study in Caracas Fashion Store). *International Journal of Information Technology and Computer Science Applications*, 1(1), pp.35–43. doi:https://doi.org/10.58776/ijitcsa.v1i1.17.

Sztandera, L. (2020). Thoughts on the Future of Data Analytics Education in College Fashion Curricula. *Journal of Textile Science & Fashion Technology*, 6(1). doi:https://doi.org/10.33552/jtsft.2020.06.000627.

TensorFlow. (n.d.). *Module: tf.keras.applications.resnet50 | TensorFlow Core v2.9.0*. [online] Available at: https://www.tensorflow.org/api_docs/python/tf/keras/applications/resnet50.

Thomassey, S. and Fiordaliso, A. (2006). A hybrid sales forecasting system based on clustering and decision trees. *Decision Support Systems*, 42(1), pp.408–421. doi:https://doi.org/10.1016/j.dss.2005.01.008.

Thomassey, S. and Happiette, M. (2007). A neural clustering and classification system for sales forecasting of new apparel items. *Applied Soft Computing*, 7(4), pp.1177–1187. doi:https://doi.org/10.1016/j.asoc.2006.01.005.

Verhoef, P.C., Kooge, E. and Walk, N. (2016). *Creating Value with Big Data Analytics*. Routledge. doi:https://doi.org/10.4324/9781315734750.

WWF (2013). *The Impact of a Cotton T-Shirt | Stories | WWF*. [online] World Wildlife Fund. Available at: https://www.worldwildlife.org/stories/the-impact-of-a-cotton-t-shirt.

www.markovml.com. (n.d.). *LIME vs SHAP: A Comparative Analysis of Interpretability Tools*. [online] Available at: https://www.markovml.com/blog/lime-vs-shap.