



Future of Sri Lankan Apparel Industry: Proposal for the B2B Sales Trend Analysis Using Machine Learning Approach

**A dissertation submitted for the Degree of Master of
Business Analytics**

W M V D Wasala


University of Colombo School of Computing

2019

DECLARATION


Name of the student: W M V D Wasala
Registration number: 2020/BA/040
Name of the Degree Program: Master of Business Analytics
Project/Thesis title: Future of Sri Lankan Apparel Industry: Proposal for the B2B Sales Trend Analysis Using Machine Learning Approach

1. The project/thesis is my original work and has not been submitted previously for a degree at this or any other University/Institute. To the best of my knowledge, it does not contain any material published or written by another person, except as acknowledged in the text.
2. I understand what plagiarism is, the various types of plagiarism, how to avoid it, what my resources are, who can help me if I am unsure about a research or plagiarism issue, as well as what the consequences are at University of Colombo School of Computing (UCSC) for plagiarism.
3. I understand that ignorance is not an excuse for plagiarism and that I am responsible for clarifying, asking questions and utilizing all available resources in order to educate myself and prevent myself from plagiarizing.
4. I am also aware of the dangers of using online plagiarism checkers and sites that offer essays for sale. I understand that if I use these resources, I am solely responsible for the consequences of my actions.
5. I assure that any work I submit with my name on it will reflect my own ideas and effort. I will properly cite all material that is not my own.
6. I understand that there is no acceptable excuse for committing plagiarism and that doing so is a violation of the Student Code of Conduct.

Signature of the Student	Date (DD/MM/YYYY)
	18/09/2024

Certified by Supervisor(s)

This is to certify that this project/thesis is based on the work of the above-mentioned student under my/our supervision. The thesis has been prepared according to the format stipulated and is of an acceptable standard.

	Supervisor
Name	Prof. M.G. Noel A.S Fernando
Signature	
Date	18/09/2024

DEDICATION

I would like to dedicate this thesis to my parent

ACKNOWLEDGEMENTS

I am blessed to surround with wonderful people who extended their blessings, encouragement, helpful advice and honest opinions.

Firstly, I am grateful to Mr M N M Nasif, Group Manager IT and Expo Industrial Engineering PVT LTD for facilitating me the required data for the research. Without his support, this research would not have been possible.

Thank you Professor Noel Fernando, my academic supervisor and my mentor for your guidance, immense support and encouragement bestowed on me right from inception until the successful completion of this research. I truly respect him for his insights, patience and support given throughout the research. Without his inputs, valuable comments and feedback, this work would have not been a reality.

I would also wish to extend my warm appreciation toward my co-worker, Mr Sasith for giving his valuable feedback and comment on the methodology that I have used in this research.

More importantly, I would not have had the courage to pursue this without my father and mother whom always being the pillar of my success in every aspect of my life.

Last but definitely not least, I would like to thank all my friends who have supported me throughout this long journey to end it with a great success.

ABSTRACT

In the backdrop of global economic challenges, Sri Lanka's apparel export industry, a significant contributor to the nation's economy, faces threats amidst the country's severe economic crisis. Despite its reputation for ethical sourcing and high-quality garments, Sri Lanka's market share in the global garment industry is relatively small compared to the dominating country, China. Recognizing the challenges associated with this reduced market share, Expo Group of Industries, a leading engineering plant in Sri Lanka, acknowledges the necessity of adopting a data-driven approach to navigate complexities and maintain competitiveness. The company is committed to leveraging data-driven strategies to overcome industry challenges, ensuring it can continue to provide tailored solutions for its clients in the fashion industry.

Sri Lanka is grappling with a severe economic crisis since March 2022, marked by a drastic drop in foreign reserves and a subsequent impact on industries, notably the apparel sector. The crisis, rooted in a dollar shortage and exacerbated by electricity tariff hikes and unfavorable tax policies, has led to increased production costs, shipping challenges, and delays in order fulfillment. The political and economic instability has eroded trust among foreign buyers, resulting in reduced orders and job losses in the apparel industry. Amid these challenges, a proposed research project aims to develop a tailored forecasting model using machine learning and time series analysis to improve B2B sales predictions in the Sri Lankan apparel industry, addressing a critical knowledge gap and offering practical insights for industry stakeholders.

The study investigates sales forecasting techniques in the B2B apparel industry, revealing that SARIMAX, Random Forest Regression, and XGBoost are effective models. While LSTM lags due to data limitations, Random Forest Regression and XGBoost consistently outperform ARIMA-based models, with XGBoost emerging as the superior performer based on lower MSE, higher R², and Explained Variance Score. These findings align with prior research highlighting the efficacy of machine learning models in sales prediction. The study fills gaps in B2B sales forecasting literature for the apparel industry, emphasizing the importance of data-driven decision-making and customer profiling for maximizing financial performance. Despite limitations, the research provides a robust foundation for evidence-based decision-making in navigating challenges and capitalizing on opportunities in the Sri Lankan apparel industry.

Table of Contents

DECLARATION	i
Certified by Supervisor(s)	i
DEDICATION	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
LIST OF FIGURES	vii
LIST OF TABLES	viii
CHAPTER 1: INTRODUCTION.....	1
1.1. Motivation.....	1
1.2. Statement of Problem.....	4
1.3. Research Aims and Objectives	7
1.4. Scope of Work and Limitations	8
1.5. Structure of the Thesis	9
CHAPTER 2: LITERATURE REVIEW	11
2.1. Business-to Business (B2B) Sales in Apparel Industry	11
2.2. Machine Learning Models	11
2.3. Time-Series Forecasting Models	16
2.4. Research Gap	19
CHAPTER 3: METHODOLOGY	21
3.1. Systematic Approach	21
(i) Data Understanding and Data Gathering	21
(ii) Data Preprocessing	22
(iii) Feature Engineering and Selecting	28
(iv) Model Selection	32
(a) Time-Series Model Selection.....	32
(b) Machine Learning Model Selection.....	34
(v) Model Validation	36
3.2. High-level Architecture Diagram	37
CHAPTER 4: RESULTS AND EVALUATION.....	39
4.1. Time Series Model Evaluation	39
4.1.1. ARIMA (AutoRegressive Integrated Moving Average) Results	39

i. Rolling Statistics and Augmented Dickey-Fuller (ADF) Test	40
ii. Differencing data for non-stationary	41
4.1.2. SARIMA (Seasonal ARIMA)	46
4.1.3. SARIMAX (Seasonal ARIMA with eXogenous factors)	49
4.2. Machine Learning Model Evaluation	52
4.2.1. Random Forest Regression.....	52
4.2.2. Extreme Gradient Boosting (XGBoost) Regression	54
4.2.3. Long Short-Term Memory Model (LSTM)	56
4.3. Model Comparison	59
CHAPTER 5: CONCLUSION AND FUTURE WORK.....	61
5.1. Overview of the Study	61
5.2. Conclusion	61
5.3. Limitations and Further improvements of the study.....	63
5.3.1. Limitations	63
5.3.2. Further improvements of the study	63
APPENDICES	I
Appendices A	I
Appendices B	II
Appendices C	III
References	XIX

LIST OF FIGURES

Figure 1: Data types in the dataset.....	22
Figure 2: Encoded dataset	28
Figure 3: Creating new column called InvoiceValue(LKR)	29
Figure 4: Correlation heat map for time series variables.....	31
Figure 5: High-level Architecture Diagram.....	38
Figure 6: Rolling Mean and Standard Deviation test result	41
Figure 7: Before and after Differencing	42
Figure 8: ACF and PACF plots after first differencing	43
Figure 9: Hyperparameters based on Akaike Information Criteria	45
Figure 10: Forecasted Model Using ARIMA	45
Figure 11: Seasonal and Trend decomposition using Loess graph	47
Figure 12: Hyperparameters based on Akaike Information Criteria	48
Figure 13: SARIMA forecasting model	48
Figure 14: Hyperparameters based on Akaike Information Criteria	50
Figure 15: SARIMAX forecasting model	51
Figure 16: Residual Plot	53
Figure 17: Predicted vs. Actual Plot.....	53
Figure 18: Actual Vs. Forecasted Sales.....	54
Figure 19: Residual plot	55
Figure 20: Predicted Vs Actual Plot.....	56
Figure 21: Time series forecasting with LSTM.....	58

LIST OF TABLES

Table 1: Data source, No of Records and Attributes	22
Table 2: Define Missing Values	23
Table 3: Missing complete rows.....	25
Table 4: ADF test results	41
Table 5: ADF test results after First Differencing.....	43
Table 6: Time Series Model comparison for Test Data	59
Table 7: Machine Learning Model comparison for Test Data	59

CHAPTER 1

INTRODUCTION

1.1. Motivation

Business-to-business (B2B) sales play a crucial role in the global economy, encompassing transactions between two or more businesses that involve the exchange of goods, services, or raw materials. These transactions can take various forms, such as a business purchasing raw materials to manufacture its own goods, purchase additional value-added services to enhance operational efficiency, or re-sells goods and services produced by other businesses (Giri, et al., 2019). In today's business setting, B2B companies encounter a multitude of challenges that demand their attention and strategic decision-making. These challenges include market disruptions, intensified competition, evolving customer demands, and the expectations of investors for profitable and sustainable growth. To navigate these complexities and stay ahead of the curve, B2B companies are increasingly turning to data-driven decision-making. The advancement of technology has revolutionized the way businesses operate, generating vast amounts of data in the process. B2B companies are now presented with an opportunity to leverage this data to gain valuable insights and make informed decisions. By adopting a data-driven approach, organizations can analyze large volumes of information, extract meaningful patterns and trends, and uncover actionable intelligence.

In a B2B environment, trend analysis and forecasting future sales are quite crucial as the entire production and supply depend on these forecasting. Analyzing historical sales happened with previous timeframes to identify sales patterns and project future trends and accurate forecasting affects not only sales, but also other areas within the business, such as strategic planning, finance, marketing, operations, and company performance assessment (Lu & Kao, 2016). With accurate estimations, decision-makers can adapt to changing market signals, make smarter business decisions and accordingly adjust their procurement and production plans. Further, these proactive approaches will lead to early adoptions for the increasing demand and quick response to any declines which will have a positive or negative impact on revenue generation. In addition, trend analysis and forecasting can increase the profitability of the company by cost prediction and decision-makers can anticipate sales volume and plan for resource allocation, such as adding more

workers and other resources, to ensure smooth operations during peak seasons (Haselbeck, et al., 2022). This strategic workforce planning helps avoid understaffing or overstaffing, optimizing productivity and reducing unnecessary costs. Additionally, sales trend analysis and forecasting provides valuable insights into the entire sales pipeline. Decision-makers can identify and address any weak links or bottlenecks in the sales process ahead of time. By proactively resolving issues, businesses can ensure stable performance throughout the term, minimize disruptions, and maintain customer satisfaction. Moreover, accurate sales trend analysis allows businesses to align their marketing strategies with projected sales volumes. Marketing campaigns can be planned and executed based on anticipated demand, ensuring efficient utilization of resources and maximizing return on investment.

Sri Lanka is indeed facing significant economic challenges, which have been described as one of the worst crises the country has experienced in decades. In March 2022, the Sri Lankan government declared itself bankrupt, unable to meet its financial obligations, including defaulting on more than \$55 billion of its foreign debts (Wickramasingha, 2023). The consequences of this economic crisis have been far-reaching. One of the immediate impacts has been a shortage of foreign reserves, severely limiting the government's ability to import essential goods and services. As a result, Sri Lanka has been struggling yearlong to provide its citizens with basic necessities, such as fuel, electricity, gas, essential drugs, and food. The scarcity of foreign currency reserves has led to challenges in importing vital commodities, causing shortages and price hikes in the market. The availability and affordability of essential items have become major concerns for the population, affecting their daily lives and overall well-being. The government has been working to mitigate the effects of these shortages, but the scale of the crisis has made it a daunting task.

In the midst of Sri Lanka's economic and political challenges, the apparel export industry, which has earned a strong reputation globally, is facing a serious threat. Over the years, Sri Lankan apparel manufacturers have established themselves as a trusted destination for ethical apparel sourcing, manufacturing high-quality garments adhering to international standards (Samanthi, 2022). These factors have positioned Sri Lanka as a trusted sourcing destination for renowned global fashion brands and the label "Made in Sri Lanka" has become synonymous with quality, reliability, and accountability in the industry (Samanthi, 2022).

The apparel export industry in Sri Lanka has played a crucial role in the country's economy, serving as the primary source of foreign exchange earnings and contributing significantly to employment

opportunities. According to a report by Online Clothing Study, the apparel industry accounted for approximately 52% of Sri Lanka's total export revenue in 2021 (Rammandala, 2022), highlighting its significance as the leading foreign exchange earner for the country. This statistic underlines the pivotal role played by the apparel sector in generating foreign exchange and strengthening the nation's economy. In addition to its economic significance, the apparel industry is also a major employer in Sri Lanka. It employs a substantial portion of the country's workforce, with approximately 15% of the total workforce being engaged in the industry (Mirza & Ensign, 2021). This demonstrates the industry's contribution to job creation and livelihoods, providing employment opportunities to a significant number of individuals. The substantial share of export revenue and the significant workforce employed by the apparel industry highlight its importance to the Sri Lankan economy.

Moreover, Sri Lanka's apparel industry is indeed home to some of the largest and high-end garment manufacturers in the world. Sri Lankan manufacturers have developed strong partnerships with these brands, supplying their products and contributing to their global supply chains. Several well-known global fashion brands have sourced their products from Sri Lanka. These brands include Banana Republic, Speedo, Ralph Lauren, H&M, Tommy Hilfiger, GAP, Marks & Spencer, Victoria's Secret, Patagonia, PVH (Calvin Klein), NIKE, Calzedonia, Levi's, and Puma (Linh, 2022). This reflects the industry's reputation for high-quality production, ethical practices, and reliability. The presence of these brands contributes to the growth and global recognition of Sri Lanka's apparel industry, creating opportunities for further development and collaborations.

Though, Sri Lanka annually exports garments approximately worth around \$2.2 billion to the United States of America, approximately \$900 million to the United Kingdom, almost \$350 million to Italy, and around \$200 million to Belgium (Anon., 2023), the global garment industry is currently dominated by China with almost a 40% apparel industry market share and Sri Lanka is holding the market share of almost 1.2%, which means Sri Lanka is not yet among the top 10 garment manufacturing countries in the world (Anon., 2023). However, it is important to note that the global garment industry is currently dominated by China, which holds a market share of almost 40% (Anon., 2023). In comparison, Sri Lanka's market share stands at around 1.2%, indicating that it is not yet among the top 10 garment manufacturing countries in the world (Anon., 2023). Nevertheless, Sri Lanka's consistent exports to major economies signify its competitiveness and ability to supply garments that meet international standards and customer expectations.

Expo Group of Industries (PVT) LTD. is a prominent engineering plant situated in the Katunayake Export Processing Zone. Originally owned by Miyaura Lanka Limited, a Japanese engineering plant, Expo Group of Industries was later acquired by the Expo Industrial Group of companies. Expo Group of Industries has established itself as a leader in the industry by offering a diverse range of services that cater to the specific needs of its clients. By excelling in sustainable branding, labeling, and print packaging, the company has earned a reputation as a trusted partner who provides environment friendly solutions for luxury global fashion retailers and apparel manufacturers in Sri Lanka. The motivation to conduct the research work at Expo Group of Industries stems primarily from client requirements. The company recognizes the importance of establishing a data-driven culture and making appropriate investments to fully unlock the potential of data and remain competitive in the marketplace. This approach will help the company to gain valuable insights and develop tailored solutions that address the unique challenges faced by its clients and it will ensure that the company remains at the forefront of innovation and can continuously adapt its offerings to exceed client expectations.

1.2. Statement of Problem

Sri Lanka has been an eye of global inquiry since 2022 March, as the country has been in the fray of a financial crisis with crippling inflation and an energy dearth. In February 2022, the country's foreign reserves plummeted by a staggering 70%, reaching to US \$2.30 billion (Varshney, 2022) and the scarcity of foreign exchange has severely limited the country's ability to import goods and materials necessary for various industries.

The repercussions of this crisis have profound impact on the apparel industry in Sri Lanka, causing numerous apparel companies to encounter significant challenges in meeting their end goals. The root causes of this predicament can be traced back to the dollar crisis that commenced in 2019 and has steadily worsened over time. The devaluation of the Sri Lankan rupee against the US dollar has led to soaring costs for imported goods, including essential raw materials for the apparel sector. This has significantly strained the financial resources of apparel companies, impeding their ability to purchase the necessary raw materials for production. In addition to the dollar crisis, the newly introduced electricity tariff has worsened the challenges faced by the industry. This tariff has resulted in a substantial hike in electricity prices, further burdening apparel manufacturers who rely

heavily on energy-intensive processes. The increased cost of electricity has put immense pressure on companies' budgets, forcing them to make difficult decisions to balance expenses and maintain profitability. Furthermore, the prevailing tax regime has imposed a significant financial burden on businesses, making it harder for them to allocate funds towards production and growth initiatives. The surge in costs associated with production have placed a heavy strain on apparel companies. Notably, the industry has experienced a massive increase in raw material prices, further squeezing profit margins. As a result, apparel manufacturers in Sri Lanka have found it increasingly challenging to sustain their operations.

Adding to the complications, the dollar crisis has resulting in many shipping companies opting out calling vessels into Colombo port leading to higher transportation costs for importing raw materials and exporting finished goods. The challenges faced by garment factories in Sri Lanka have become growth impediments on multiple fronts, impacting their operations and the overall competitiveness of the industry. The two significant challenges that hinder its growth are the rising cost of shipping and logistics, and the timely delivery of finished goods. The increasing expenses associated with shipping and logistics pose a considerable obstacle for apparel companies, making it increasingly difficult for them to maintain profitability. Additionally, ensuring on-time delivery of finished goods has become a critical concern for garment manufacturers. The scarcity of shipping services and the financial crisis have disrupted the supply chain, resulting in delays and uncertainties in order fulfillment. This not only jeopardizes relationships with buyers but also raises concerns about the ability to meet delivery deadlines. The unreliability in meeting customer expectations due to these challenges can have adverse effects, including dissatisfied customers, damaged buyer relationships, and potential loss of future orders.

The fragile political and economic conditions in Sri Lanka have significantly impacted the garment industry, leading to lower-than-anticipated orders from foreign buyers. As highlighted by Dias (Dias, 2022), concerns about the country's stability and potential disruptions to order fulfillment have caused foreign buyers to worry, resulting in reduced orders for export-oriented manufacturers. Such economic crises rooted in political issues often lead to a lack of trust and confidence among the international buying fraternity. As highlighted by Business Times, this mistrust has prompted some brands and retailers to shift their sourcing orders from Sri Lanka to neighboring countries, aiming to mitigate the perceived risks (Dias, 2022). Consequently, the scarcity of orders has forced many larger apparel companies in Sri Lanka to scale back their operations. The ramifications of

this situation have been particularly devastating for the livelihoods of workers in the apparel industry. With limited orders, these companies are operating at reduced capacity, typically three to four days per week. Furthermore, to align with the decreased production levels and cost constraints, some companies have been compelled to downsize their existing workforce (Dias, 2022). The adverse effects of the industry's challenges are visible in the significant job losses that have already occurred in the apparel sector since 2022. Business Times reports that more than 10,000 individuals have lost their jobs in the industry (Dias, 2022).

The apparel industry is characterized by plethora of factors that make predicting future product demand or sales. These factors include the wide variety of product styles, patterns, short life cycles, and fluctuating consumer demands which can lead to flawed or less accurate predictions (Giri, et al., 2019). On the other hand, the complexity of business dynamics further complicates decision-making, often leading to subjective judgments based on personal experiences and mental models. To address these challenges and improve business performance in the fashion industry, cutting-edge data analytics tools and robust trend analysis and forecasting models can play a crucial role. By effectively handling the vast amounts of data available, businesses can gain valuable insights and make more accurate predictions. However, based on the past research, Machine learning (ML) techniques and time-series analysis have shown promise in enhancing revenue forecasting in the apparel industry (Bohanec, et al., 2017). These advanced techniques can be applied to various prediction problems, enabling businesses to make data-driven decisions and optimize their operations. Moreover, from a Sri Lankan perspective, it is particularly important to closely monitor sales trends, especially in the Business-to-Business (B2B) segment of the apparel industry. By keeping a vigilant eye on B2B sales, companies can gain a deeper understanding of market dynamics, identify emerging opportunities, and navigate the current financial and geopolitical challenges more effectively. This awareness can help them adapt their strategies and make informed decisions to stay competitive in a rapidly changing business environment.

While Machine Learning (ML) techniques and time series analysis have been extensively studied in data-driven decision-making, there is a noticeable gap in the literature when it comes to B2B sales forecasting. To gain a comprehensive understanding of which model performs well in B2B sales forecasting and trend analysis within the apparel industry, it is necessary to conduct empirical research using real-world data from this specific sector. The existing studies and reports in this area have explored various algorithms and models in different industries, but their findings may not

directly translate to the apparel industry and mostly consist of exploratory analyses, and there is a lack of scientific focus on B2B sales specifically in the context of the apparel industry in Sri Lanka. By collecting and analyzing data that captures the seasonal changes and other relevant factors unique to the apparel industry, researchers can evaluate different trend analysis and forecasting techniques and identify the most effective model for accurate sales predictions.

To address this knowledge gap, the proposed project aims to bridge the theoretical and practical aspects by incorporating supervised ML and time series analysis specifically tailored to the apparel industry, considering the challenges and characteristics specific to B2B sales in this sector. Such research will provide valuable insights and practical guidance for businesses in the apparel industry seeking to improve their sales accuracy and by leveraging these techniques, the project seeks to develop a robust trend analysis and forecasting model for B2B sales in the apparel industry, specifically tailored to the Sri Lankan context. The findings can help industry stakeholders make data-driven decisions, optimize their sales strategies, and enhance their competitiveness in the market.

1.3. Research Aims and Objectives

In today's competitive world, trend analysis and accurate sales forecasting plays a vital role in inventory management preventing overproduction and overstocking while maximizing revenue generation (Ensafia, et al., 2022). Predominantly in a business-to-business (B2B) environment, forecasting future demand holds great significance, as it supports corporate analysis and decision-making, providing a potential competitive advantage across various domains. Thus, this research project is motivated by the need of the client in terms of analyzing and forecasting sales trends in the apparel industry. Even though the analysis is based on the local producers, the outcome of the project will provide broader insight into global fashion brands and their market behaviour in the forthcoming years. Moreover, data visualization is often considered the most powerful means to communicate crucial information to the stakeholders enabling them to understand and act upon it. Hence the objectives of the project are as follows.

1. To explore the trend in the apparel export industry based on historical data by performing regression forecasting techniques such as SARIMA, SARIMAX.and others.

2. To determine the accuracy of classification machine learning models in predicting and forecasting the behaviour of the apparel export industry.
3. To determine the most effective method to forecast apparel industry data that exhibits seasonality or non-seasonality, by comparing regression and classification machine learning models.
4. To visualize patterns and trends, providing top management with actionable insights to facilitate decision-making and enhance export performance.

By accomplishing these objectives, the project aims to enhance the understanding of sales trends in the Sri Lankan apparel industry, equip decision-makers with reliable trend analysis and forecasting tools, and provide valuable visual representations that support effective decision-making processes, ultimately leading to improved export performance.

1.4. Scope of Work and Limitations

This research project aims to contribute new knowledge by bridging the existing knowledge gap in the literature related to trend analysis and sales forecasting in the context of B2B sales in the apparel industry. By leveraging data-driven decision-making, the study will assist policymakers in the apparel industry to make informed choices based on data analysis rather than relying solely on intuition. Additionally, the project aims to identify the most profitable customers, enabling the company to focus its efforts on generating additional revenue and maximizing its financial performance.

The scope of the research is primarily defined by the data set provided by the stakeholder, Expo Group of Industries (PVT) LTD, covering the period from 2019 to 2023 and the analysis will focus on forecasting trends in terms of client-wise, brand-wise, and item-wise sales. By exploring these different dimensions, the project seeks to gain a comprehensive understanding of the underlying sales patterns within the Sri Lankan apparel industry.

However, it is important to acknowledge the limitations of the study. Firstly, the research will not address the global market demand pertaining to specific products or countries, as the necessary

data is not available with the client. Hence, the focus will be only on analyzing and forecasting sales trends within the specific context of the provided dataset. Even though the findings may provide insights into broader market behavior, it is important to interpret the results within the scope of the Sri Lankan apparel industry.

Additionally, the research project will not consider unstructured data such as emails and invoices. Although these sources could potentially carry valuable insights, they are excluded from the analysis due to the project's defined scope and available resources. The research will primarily rely on the structured data provided by the client to conduct the forecasting and analysis.

Despite these limitations, the project aims to provide valuable insights into the sales trends within the Sri Lankan apparel industry, enabling decision-makers equip with necessary information to navigate the challenges posed by the current financial and geopolitical climate effectively. The findings obtained through this research endeavor will serve as a solid foundation for evidence-based decision-making, empowering stakeholders to make informed choices that enhance and deepen their understanding of the performance of the Sri Lankan apparel industry. By shedding light on key insights and trends, this study aims to contribute to the improvement and comprehensive understanding of the industry's dynamics, navigate challenges and seize the opportunities in the future.

1.5. Structure of the Thesis

The thesis organized itself to five chapters and the first chapter begins with the motivation, followed by elaborating the research problem. Then the research objectives had been discussed and the significance of the study had been defined. After that, the assumptions and limitations of the study had been illustrated and finally, the chapter concluded after outlining the structure of the thesis.

The review of relevant literature is presented in chapter two and it begins with the definitions of B2B Business. Then the behaviors of machine learning models and time series models had been exposed. The chapter concluded after explaining the research gap.

The third chapter describes the research methodology and it begins with an overview of the research philosophy. Then it explains the five steps involved in the selected research approach.

The fourth chapter presented the results and evaluations of this study. First, it explains three time series models used in this study and the outcome is explained. Then, three machine learning models used in this study is explained in details along with its results. Chapter concludes with the comparison of outputs received from time series models and machine learning models.

Then the final chapter presented the overview of the study. Then it provides a conclusion and then finally limitations of the study and future improvements had been discussed.

CHAPTER 2

LITERATURE REVIEW

2.1. Business-to Business (B2B) Sales in Apparel Industry

In the apparel industry, B2B sales refer to the business-to-business model, where one business sells its products and services to another business rather than directly to consumers. This B2B sales model involves longer sales cycles, extended contracts, and the establishment of enduring relationships with other companies. However, the apparel industry faces unique challenges in forecasting future demand for its products. These challenges arise from factors such as the short life cycle of apparel products, limited historical data availability, high market demand uncertainty, and seasonal trends. To address these challenges, researchers have attempted to develop forecasting models. However, these models have only achieved moderate accuracy in predicting future demand for apparel products. Despite the abundance of data and advancements in statistical and machine learning techniques that have significantly improved data-driven decision-making in various domains, there is limited literature available on machine learning techniques specifically applied to B2B sales forecasting in the apparel industry. This scarcity of research makes it difficult to comprehend the recent developments in this particular domain of the apparel industry.

Many organizations in the apparel industry have yet to embrace machine learning techniques at the B2B sales forecasting stage of their supply chain. As a result, the adoption of advanced forecasting methods and the exploration of machine learning techniques in this context are still relatively limited. Further research and exploration in this area could contribute to improved forecasting accuracy, better decision-making, and enhanced operational efficiency in B2B sales within the apparel industry.

2.2. Machine Learning Models

Over the years, researchers and analysts have implemented various machine learning algorithms and time series techniques to aid in trend analysis and forecasting problems that impact business

decision-making. Some studies have indicated that machine learning models offer improved predictive capabilities compared to traditional time series models, although further confirmation is still needed. Machine learning models are known for their computational efficiency, making them suitable for handling large datasets. They excel in dealing with multi-dimensional data and are capable of handling outliers effectively. Additionally, machine learning models have a high learning rate compared to other algorithms and demonstrate robustness in the presence of noise. One advantage of machine learning models is their versatility, as they can be applied to both classification and regression problems. They have been utilized in recent research and real-world scenarios across various fields. These models offer potential for enhancing the accuracy and effectiveness of trend analysis and forecasting tasks, thereby supporting informed decision-making in business contexts.

In 2015, Bohanec et al. (Bohanec, et al., 2017) developed a supervised machine learning model that employed the double-loop learning technique to address trend analysis and B2B sales forecasting. Due to the limited availability of instances, they employed a feature engineering approach using the R package to create additional instances (attributes) for training the model. The study concluded that the random forest model emerged as the top-performing classifier, achieving an accuracy of 96% and showing great promise for B2B sales forecasting. However, it is important to note that the research had limitations in terms of the number of training instances. The majority of these instances were artificially created, which might explain the high levels of classification accuracy obtained using this approach. While the results of the study are promising, further research is needed to validate the findings with larger and more diverse datasets. The use of artificially created instances raises questions about the generalizability of the model to real-world scenarios. Therefore, it is crucial to conduct future research using more extensive and representative datasets to assess the true effectiveness and reliability of the random forest model for B2B sales forecasting in the apparel industry.

In 2019, Mortensen et al., (Mortensen, et al., 2019) developed a model to assess B2B sales using different classification algorithms. After applying feature engineering techniques, the research utilized a total of 15 attributes. However, only four attributes were found to be highly significant, while the others exhibited less importance and were excluded from some models (Mortensen, et al., 2019). The main attributes used in the study were: Type (This attribute indicated whether an opportunity was new, incremental, or a renewal), Amount (It represented the size of an opportunity

or the expected amount that a customer would pay), Task Count (This attribute was calculated by tallying the total number of tasks created and linked to an opportunity's parent account) and Complexity (It referred to the complexity level of a product being offered to a customer). The researchers employed various classification models, including Multiple Logistic Regression, Decision Tree, Random Forest (RF), and XGBoost. Among these models, the Random Forest algorithm achieved the highest accuracy rate of 80%. It had a precision of 86% and recall of 77%. However, the researchers acknowledged that the accuracy was relatively low due to the poor quality of the data used. They recommended further improvements to enhance the model's performance. The findings of this study highlight the importance of selecting relevant attributes and employing suitable classification algorithms for B2B sales forecasting. The relatively low accuracy emphasizes the need for better data quality and additional enhancements to the model. Future research can focus on refining the model and exploring other techniques to improve its predictive capabilities in B2B sales analysis.

In the same year, Arif et al. (Arif, et al., 2019) conducted an analysis using three popular machine learning (ML) algorithms for B2B sales forecasting. The K-Nearest Neighbor (KNN), Gaussian Naïve Bayes, and Decision Tree Classifier algorithms were tested to determine the best technique for forecasting sales demand. The dataset was split into an 80:20 ratio, with 80% of the data used for training the model and 20% for testing. They evaluated the algorithms based on metrics such as accuracy, precision, sensitivity, and F1 score to determine the best technique for forecasting sales demand. The results indicated that Gaussian Naïve Bayes achieved the highest accuracy of 58.92% and was considered the best algorithm for demand forecasting among the three tested models. However, it is important to note that this accuracy level may be relatively low compared to other studies. One notable finding from this research was that the geographical area had an impact on the prediction. However, further research is needed to explore additional factors and improve the accuracy of demand forecasting models in the B2B context.

The relationship between past data and sales forecasting can be complex, and different studies may present varying perspectives on the most appropriate approach. Pavlyshenko (Pavlyshenko, 2019) argues that sales prediction is a regression problem rather than strictly a time series problem, suggesting that regression-based models may be more suitable for sales forecasting. The research draws upon a dataset from the "Rossmann Store Sales" Kaggle competition, comprising a substantial volume of sales data from Rossmann stores. The dataset, featuring numerous data

points, serves as the foundation for examining the application of machine learning models in sales forecasting. The investigation begins with descriptive analytics and data visualization, uncovering valuable insights into sales distributions, correlations, and influential factors. Notably, the research introduces the concept of reframing sales prediction as a regression problem, highlighting the potential advantages of regression-based machine learning models over traditional time series approaches. A key contribution lies in the exploration of machine-learning generalization, demonstrating its capacity to improve prediction accuracy, even when historical data is limited, such as during the launch of new products or stores. Furthermore, the study explores stacking techniques, a method of combining predictions from multiple models to enhance forecasting precision. However, certain research gaps emerge within these findings. The reliance on a single Kaggle dataset raises questions about the generalizability of the approaches to other industries and datasets. Future research should seek to validate these methods across diverse contexts. Additionally, the study's focus on accuracy as the primary evaluation metric may not fully capture the complexities of real-world sales forecasting scenarios, warranting the exploration of industry-specific evaluation metrics to ensure practical relevance.

In 2020, Rezazadeh (Rezazadeh, 2020) utilized two promising supervised classification algorithms, XGBoost and LightGBM, for predicting B2B sales. The voting ensemble method was employed to make predictions on an unseen dataset, specifically to predict the likelihood of winning sales opportunities. The study utilized a dataset consisting of 25,578 closed sales opportunity records from January 2015 to August 2019. A total of 20 features were used, including sales project attributes (Opportunity Type, General Nature of Work, Detailed Nature of Work, Project Location, Project Duration, Total Contract Value, Status), customer information attributes (Account, Account Location, Key Account Energy, Key Account Finance, Key Account Healthcare), and resource allocation attributes (Business Unit, Engagement Manager, Sales Lead, Probability, Sub-practice, Practice, Group Practice, Segment, User-entered Probability). The ML model was developed on the Azure ML platform, and the data was extracted from the CRM cloud database. In this study, the ML workflow accurately classified 87% of the unseen sales data, outperforming user-entered predictions which had an accuracy of only 67%. However, a notable drawback of the research was the presence of imbalanced data, with more lost sales instances than won sales instances. To address this issue, the study suggested using under-sampling techniques in conjunction with an ensemble model. The findings of above studies highlight the effectiveness of

machine learning algorithms in B2B sales forecasting. Gaussian Naïve Bayes demonstrated the highest accuracy for demand forecasting, while the ensemble model incorporating XGBoost and LightGBM achieved significant accuracy in predicting sales opportunities. Future research could focus on refining these models further and addressing challenges related to imbalanced data to improve the accuracy and reliability of B2B sales forecasts.

In a study conducted by Wisesa et al. (Wisesa, et al., 2020), the focus was on B2B telecommunication sales, and four machine learning algorithms were analyzed: Decision Tree (DT), Generalized Linear Model (GLM), Random Forest (RF), and Gradient Boost Tree (GBT). The research utilized sales data from the period of 2016-2018, and regression metrics were employed to evaluate the models' performance. The features considered in this study included Category, City, Type of items and its opportunity-ID, Quarter, Product Name, Sub Service Product, Service Product (MIDI or Non-MIDI), and Sales Revenue. To assess the model's performance, metrics such as Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Performance Metrics were utilized. Based on the results of MSE and MAPE, it was concluded that the Generalized Linear Model (GLM) exhibited the smallest error values and outperformed the other algorithms. Additionally, the study suggested the use of ADAboost to enhance the performance of the Gradient Boost Tree (GBT) model. Overall, the research by Wisesa et al. (Wisesa, et al., 2020) demonstrated the effectiveness of machine learning algorithms, particularly GLM, in predicting B2B telecommunication sales. The suggestion to utilize ADAboost to improve GBT performance provides insights into potential approaches for enhancing the accuracy of sales forecasting in this domain.

Moreover, in the year 2021, Raizada & Jatinderkumar (Raizada & Jatinderkumar, 2021) explores the application of data mining and machine learning techniques in predicting sales for retail businesses, with a focus on Walmart stores. Sales forecasting is a critical task for the success of retail organizations, and this study aims to compare various supervised machine learning algorithms to build accurate prediction models. The paper discusses the use of several machine learning algorithms, including Multiple Linear Regression, Random Forest Regression, K-NN Algorithm, Support Vector Machine (SVM), and Extra Tree Regression. These algorithms are applied to predict the sales of 45 retail outlets of Walmart located in different geographical regions. The prediction models take into account various features such as date, weekly sales, holiday flags, temperature, fuel prices, Consumer Price Index (CPI), and unemployment rates in the respective

states. The key contributions of this paper include helping business owners decide which approach to use when predicting sales for their supermarkets. By considering different scenarios and factors, this research aids in making informed decisions regarding promotional and marketing strategies for products. The paper provides a detailed methodology for data preprocessing, model training, and evaluation. It presents the results of applying these machine learning techniques to Walmart sales data from three different years (2010, 2011, and 2012) and compares their performance in terms of Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The findings indicate that the Extra Tree Regression technique consistently outperforms other models, achieving high accuracy in sales prediction across all three years. This suggests that ensemble learning methods, such as Extra Tree Regression and Random Forest Regression, are effective for sales forecasting in retail. On the other hand, simple linear regression models are less accurate for short-term sales predictions. Finally, it emphasizes the importance of considering external factors and provides practical guidance for business owners looking to enhance their sales prediction capabilities.

2.3. Time-Series Forecasting Models

In the field of sales forecasting, time-series analysis plays a crucial role, and several statistical models have been employed in addition to the machine learning algorithms mentioned earlier. These models are designed specifically for time-series analysis and sales forecasting, taking into account the temporal nature of the data. Some commonly used models in time-series analysis and sales forecasting include Auto Regression (AR), Moving Average (MA), Autoregressive Moving Average (ARMA), Auto Regressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), Seasonal Support Vector Regression (SSVR), Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors (SARIMAX), Vector Auto Regression (VAR), Vector Auto Regression Moving-Average (VARMA), Vector Auto Regression Moving-Average with eXogenous factors (VARMAX) and Holt Winter (HW) (Ensafia, et al., 2022) (Pavlyshenko, 2019) (Anil, et al., 2023).

ARIMA is a widely used model that captures the autocorrelation and moving average components in the time series data. SARIMA and SARIMAX are extensions of ARIMA that specifically address the seasonal component of the series. These models are utilized when the data exhibits

seasonality or when the seasonal component needs to be considered in the forecasting process. SSVR is a regression-based model that incorporates the seasonal patterns into the forecasting process. Holt Winter (HW) is another popular model that captures both the trend and seasonal components in the data. These time-series forecasting models are trained on historical sales data to identify trends and seasonality, enabling them to make predictions about future sales. Accurate sales forecasting is crucial for strategic decision-making and planning, and as a result, considerable efforts have been made in the literature to improve the accuracy of these forecasting models.

Contrary to the prevalence of ML algorithms in sales forecasting literature, several studies (Haselbeck, et al., 2022) (Stephan, 2022) (Makridakis, et al., 2018) (Makridakis, et al., 2020) have indicated that it is not necessarily clear whether these algorithms are inherently superior to classification models. The effectiveness of forecasting models depends on the specific application and the nature of the data being analyzed. In fact, some studies have found that classical models can outperform more complex ML approaches, such as Artificial Neural Networks (ANN) (Haselbeck, et al., 2022) (Stephan, 2022) (Makridakis, et al., 2018) (Makridakis, et al., 2020). The suitability and performance of forecasting models are highly context-dependent. While ML algorithms have gained popularity for their ability to handle complex and high-dimensional data, they may not always be the best choice for every forecasting task. Classical models, such as ARIMA or Holt-Winters, have been well-established and have a strong theoretical foundation. In some cases, these classical models have been found to provide more accurate forecasts than sophisticated ML algorithms like ANN. Therefore, it is essential to carefully consider the specific requirements of the forecasting problem, the characteristics of the data, and the limitations of different modeling approaches when selecting the appropriate forecasting model. Ultimately, the choice between classical models and ML algorithms should be based on empirical evidence and an understanding of the specific forecasting task at hand.

The study conducted by Haselbeck et al. (Haselbeck, et al., 2022) aimed to determine the best prediction model by combining both classical and ML models for horticultural plant sales. Although the research focused on horticultural sales, the presence of seasonality in the data makes it relevant for understanding sales forecasting in the apparel industry, which also experiences seasonal trends. The study considered three classical models: Exponential Smoothing (ES), Seasonal ARIMA (SARIMA), and Seasonal ARIMA with exogenous factors (SARIMAX). In addition, several ML models were utilized, including Lasso Regression, Ridge Regression, Elastic

Net Regression, Artificial Neural Network (ANN), Long Short-Term Memory Network (LSTM), Extreme Gradient Boosting (XGBoost), Bayesian Ridge Regression, Automatic Relevance Determination, and Gaussian Process Regression (GPR). The findings of the research indicated that SARIMAX achieved results close to the best ones among the classical models. GPR and LSTM also delivered competitive results, but XGBoost outperformed the other models. These results suggest that XGBoost, which is a popular ML algorithm known for its ensemble learning and gradient boosting capabilities, was the most effective model for predicting horticultural plant sales in the study. While the research focused on horticultural plant sales, the findings imply that XGBoost could potentially be a suitable model for sales forecasting in the apparel industry as well, given the similar presence of seasonality in the data. However, it is important to note that the performance of the models may vary depending on the specific characteristics of the apparel industry data. Therefore, further research and experimentation would be necessary to validate the effectiveness of XGBoost or other models for sales forecasting in the apparel industry.

Amrutkar and Mahadik (Amrutkar & Mahadik, 2022) also express dissatisfaction with the predictions obtained from traditional time series methods. On the other hand, based on previous research outcomes, machine learning algorithms have been shown to provide better results compared to traditional time series methods (Pavlyshenko, 2019) (Amrutkar & Mahadik, 2022). These algorithms offer the advantage of being able to identify complex patterns in sales dynamics, allowing for more accurate predictions (Pavlyshenko, 2019) (Amrutkar & Mahadik, 2022). By leveraging supervised machine learning methods, it becomes possible to capture and analyze intricate relationships between various factors that influence sales, leading to improved forecasting accuracy. Further, Pavlyshenko (Pavlyshenko, 2019) identified several limitations in using time series approaches for sales forecasting. Firstly, these methods typically require long historical data to capture seasonality, which may be lacking when launching new products. In such cases, leveraging sales data from similar products with similar sales patterns becomes essential. Secondly, sales data often contain outliers and missing values, necessitating data cleaning and interpolation before applying time series techniques. Lastly, the complexity of sales forecasting is compounded by the need to consider numerous exogenous factors that influence sales. These challenges underscore the importance of alternative approaches, such as machine learning-based regression models, which can address these limitations more effectively.

The studies mentioned earlier have provided insights into different approaches and models used in sales forecasting, but they may not directly apply to the specific context of the apparel industry. Given the nature of the apparel industry, with its short product life cycles, uncertain market demand, and seasonal trends, it is crucial to conduct research specifically focused on this domain. The existing studies have explored various algorithms and models in different industries, but their findings may not directly translate to the apparel industry.

2.4. Research Gap

One of the primary gaps in the literature is the scarcity of research specifically focused on B2B trend analysis and sales forecasting in the apparel industry. While there have been studies exploring machine learning and time-series forecasting models in other domains, such as horticultural sales and telecommunication, these findings may not directly apply to the unique challenges faced by the apparel industry, such as short product life cycles, limited historical data availability, high market demand uncertainty, and seasonal trends. Although some studies have indicated the potential of machine learning models in B2B sales forecasting, there is still a lack of comprehensive research that thoroughly explores the effectiveness and reliability of various machine learning techniques in the context of the apparel industry. Moreover, the existing literature does not extensively cover the combination of classical time-series forecasting models with machine learning algorithms to leverage the strengths of both approaches. Several studies have highlighted the importance of data quality in achieving accurate sales forecasts. However, some of the research mentioned in the literature review acknowledged issues related to poor data quality, imbalanced data, and artificially created instances, which may affect the reliability and generalizability of the forecasting models.

- To address the gaps identified in the literature and achieve the objectives of improving B2B sales forecasting accuracy in the apparel industry, the following methodologies can be proposed:
- Conduct an extensive data collection effort to gather historical B2B sales data from the apparel industry. The data should include information on product attributes, customer profiles, sales cycles, seasonal patterns, and market demand fluctuations. Preprocess the

data to handle missing values, outliers, and imbalances. Utilize advanced techniques for feature engineering to capture relevant insights from the data.

- Develop a hybrid approach that combines the strengths of classical time-series forecasting models (e.g., ARIMA, SARIMA, Holt-Winters) with machine learning algorithms (e.g., XGBoost, Random Forest). This hybrid approach can be designed to handle the temporal nature of the data while leveraging the ability of machine learning models to capture complex patterns and interactions among variables. Implement ensemble modeling techniques to further improve the forecasting accuracy.
- Ensemble methods, such as stacking or blending, can combine the predictions from multiple forecasting models to create a more robust and accurate final forecast.

Utilize appropriate evaluation metrics, such as Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), to assess the performance of the forecasting models. Validate the models using out-of-sample data and conduct sensitivity analysis to ensure their reliability.

CHAPTER 3

METHODOLOGY

3.1. Systematic Approach

In this study, a systematic approach is employed, which consists of six distinct steps. (i) Data understanding and data gathering, (ii) Data pre-processing , (iii) Feature engineering and selection , (iv) Model selection , (v) Model validation and (vi) Visualization.

The subsequent sections will provide a comprehensive explanation of each phase of these approaches and how they have been implemented within the study.

(i) Data Understanding and Data Gathering

With the explicit consent from all relevant stakeholders, the sales data has been obtained from Expo Group of Industries (PVT) LTD. This acquisition has been carried out in strict adherence to ethical and legal requirements. The dataset encompasses a total of 76,486 records, spanning from January 2018 to March 2023. The dataset utilized in this research draws information from three distinct tables: Invoice Master, Invoice Items, and Evaluation. These tables are interconnected through primary keys, allowing for a comprehensive analysis.

- **InvoiceMaster** : provides a high-level overview of each invoice.
- **InvoiceItems** : offers detailed information about item descriptions associated with individual invoice numbers.
- **Evaluation** : table contains intricate product details.

Through careful consideration of all three tables, 7 attributes (as listed in Table 1) have been selected due to their anticipated substantial impact on predictive modeling efforts.

Table 1: Data source, No of Records and Attributes

Data	No of Data Records	Attributes
InvoiceMaster	76,486	InvoiceID InvoiceDate BuyersName InvoiceAmount ExchangeRate
InvoiceItems	76,486	InvoiceID Qty UnitPrice ProductID
Evaluation	8790	ProductID Description

(ii) Data Preprocessing

Data preprocessing is a crucial step in empirical research that ensures the quality and suitability of the data for analysis and it takes up to 90% of the project time. Below are the key considerations and steps involved in data preprocessing.

- **Data type investigation:** The purpose of this step is to understand the data types of the variables in the dataset (e.g., numerical, categorical, text) and handle them appropriately. Different data types may require different preprocessing techniques, such as scaling numerical variables, encoding categorical variables, and handling text data through text preprocessing methods like tokenization or stemming. As per the figure 1, the dataset contains 3 data types: int64, datetime64[ns], and float64.

```
InvoiceDate      datetime64[ns]
BuyerName        object
UnitPrice        float64
Qty              float64
InvoiceAmount (USD) float64
ExchangeRate     float64
ProductID        object
Description       object
dtype: object
```

Figure 1: Data types in the dataset

- **Data cleaning:** It includes cleaning dataset by addressing any errors, inconsistencies, or missing values. This involves removing duplicate entries, correcting errors, imputing missing data, and excluding outliers. Identifying and removing duplicate records from the dataset helps ensure that each data point is unique and reduces the potential for biases caused by duplicate information. Data cleaning enhances the overall quality of the dataset and improves its reliability for subsequent analysis. Further, many real-world datasets contain missing values for various reasons. They are often encoded as NaNs, blanks or any other placeholders.

The below mentioned "Missing Data Summary" report (Table 2) presents information about missing values within the dataset. As evident from the summary, the "Description" column contains a notable number of missing values, specifically 52, while the "ExchangeRate" and "ProductID" columns exhibit 14 and 13 missing values respectively. In terms of missing values as a percentage of the total data, the "Description" column accounts for approximately 0.067% of missing values, whereas the "ExchangeRate" and "ProductID" columns have 0.018% and 0.016% missing values respectively.

Table 2: Define Missing Values

	Missing Values	Missing Percentage
Description	52	0.067986
ExchangeRate	14	0.018304
ProductID	13	0.016997
InvoiceDate	0	0
BuyerName	0	0
UnitPrice	0	0
Qty	0	0
InvoiceAmount (USD)	0	0

Training a model with a dataset that has missing values is crucial as it can significantly affect the quality of the machine learning model. One way to handle this problem is to get rid of the observations that have missing data. However, it will risk losing data points with valuable information. A better strategy would be to impute the missing values and there are

different imputation methods that can be applied to address this issue. However, handling missing data is vital because there are different types of missing data patterns:

- **Missing Completely at Random (MCAR):** Data is missing entirely randomly, and there is no relationship between the missingness of data and any observed or unobserved variables.
- **Missing at Random (MAR):** Data is missing based on some observed variables but not others. There is a relationship between the missingness and other measured data.
- **Not Missing at Random (NMAR):** Data is missing in a way that is related to unobserved or unmeasured variables. This can be the most challenging type of missing data to handle.

While addressing missing values in the dataset, the following particular concerns have been taken into consideration:

- **Missing Product Descriptions and ProductID:** It was observed that the Evaluation table had missing product descriptions and Product ID (Table 2). In response, the study employed Mode Imputation as the chosen method for addressing this issue. The rationale behind opting for mode imputation lies in the assumption that the absence of product descriptions is somewhat correlated with other observable variables in the dataset, aligning with the "MAR" (Missing at Random) category. This imputation technique enables to preserve valuable data points and uphold the integrity of the dataset while mitigating the impact of missing values.

Below are some key points regarding this approach:

1. In a sales dataset, the "Description" and "ProductID" columns typically contain text data describing the products and Product Code. These data can be diverse and challenging to impute accurately based on numerical or other columns.
2. The mode (most frequent) value in the "Description" and "ProductID" column can be a reasonable imputation strategy because it assumes that the most common product description is likely to be a good representation of the missing values.

3. This method is simple, computationally efficient, and can work well when certain products are sold more frequently than others.

- **Missing Exchange Rates:** As observed 0.018% data (14 records) related to exchange rate is missing in the dataset (Table 2) and given the specific nature of this data, forward fill imputation (also known as previous value imputation) has been employed to address these missing entries. Forward fill imputation is a suitable method for scenarios involving time series data, where data points are recorded at regular time intervals (e.g., daily, monthly). The rationale behind this choice is grounded in the reasonable assumption that exchange rates tend to change gradually over time. Therefore, the most recent known exchange rate is a plausible approximation for the missing value in the immediate future. This imputation strategy leverages the temporal order of data points and ensures that the imputed values maintain a logical and coherent progression, aligning with the dynamics of exchange rate fluctuations.
- **Missing Complete Records:** It is observed that there are 100 missing dates, which accounts for approximately 5.22% of the total data are missing in the dataset. A "missing date" signifies that the entire row of data is absent, meaning that the entire entry is incomplete. Hence, new date rows have been generated and UnitPrice and Qty, have been generated using mean imputation method. That means imputing missing values by taking the average of the nearest non-missing values. ExchangeRate is generated using Forward Fill imputation. InvoiceAmount (USD) is calculated using $\text{UnitPrice} * \text{Qty}$ formula. To impute ProductId and Description, mode imputation is used. With that, the data set is containing total of 76,586 records.

Table 3: Missing complete rows

	Missing Values	Missing Percentage
Description	100	0.130572
ExchangeRate	100	0.130572
ProductID	100	0.130572

InvoiceDate	100	0.130572
BuyerName	100	0.130572
UnitPrice	100	0.130572
Qty	100	0.130572
InvoiceAmount (USD)	100	0.130572

- **Remove Outliers:** Outliers in a dataset can arise from various sources, such as data entry errors, measurement inaccuracies, or rare and extreme events. These atypical data points don't represent the typical behavior of the dataset and have the potential to introduce bias and inaccuracies in statistical analyses and machine learning models. If left unaddressed, outliers can lead to skewed results, incorrect conclusions, and adversely affect predictive model performance. Many machine learning algorithms are sensitive to outliers, as they can disproportionately influence model training and predictions. Removing outliers is crucial to create robust and accurate models that can generalize well to new data. Additionally, outliers can distort data visualizations and plots, making it challenging to discern meaningful patterns and trends.

The Interquartile Range (IQR) is a widely accepted statistical technique employed for detecting and eliminating outliers within a dataset. It signifies the spread of the middle 50% of the data and is computed as the difference between the 75th percentile (Q3) and the 25th percentile (Q1) of the dataset. In this study, the IQR method has been utilized to detect and address outliers by establishing a range within which data points are considered normal.

This approach has been applied in this study to address outlier detection. It has been observed that the dataset contains upper-bound outliers amounting to 0.05%. To ensure the accuracy of the study's outcomes and machine learning models, a decision was made to remove these outliers. Following the removal of outliers, the dataset has been reduced to 76,464 records.

- **Encoding Categorical Columns using Label Encoding:** The dataset contains three categorical columns with non-numeric values. To ensure compatibility with machine learning models, a label encoding process was applied. This process transforms categorical values into numerical representations and it was employed to overcome the challenge that machine learning models often encounter when directly working with text data. The following categorical columns have undergone encoding:

- **BuyerName:** This column contains buyer names represented using alphabetic characters. Since machine learning models typically cannot process text fields directly, a label encoding process has been applied. In this process, each unique buyer name has been assigned a unique numerical value, allowing the model to work with the data effectively
- **Description:** In this column, product descriptions contain a mixture of letters, numbers, and special characters. Given that machine learning models usually cannot handle text data directly, a label encoding process has been utilized. During this process, each description has been mapped to a distinct numerical value, enabling the model to process and analyze the data efficiently.
- **ProductID:** This column presented a unique challenge as it contained a mix of both numeric and character values. To maintain uniformity in data types and facilitate meaningful machine learning analysis, label encoding has been employed. Through this process, each distinct value has been assigned a unique numeric representation.

The categorical columns, BuyerName, Description, and ProductID, have been encoded using the label encoding method. Below (figure 2) is the representation of the data after this process, marking the completion of the data preprocessing phase.

```

First five records of columns before encoding:
      BuyerName      Description \
0  ATG HAND CARE (PVT) LTD      SQUARE HEAT STICKER ( 03 COLOUR )
1  ATG HAND CARE (PVT) LTD      SQUARE HEAT STICKER ( 03 COLOUR )
2  ATG HAND CARE (PVT) LTD      SQUARE HEAT STICKER ( 03 COLOUR )
3  ATG HAND CARE (PVT) LTD      SQUARE HEAT STICKER ( 03 COLOUR )
4      BRANDIX APPAREL LTD  MADE IN SRI LANKA HEAT SEAL ( 46065/A )

      ProductID
0  ATGSQ500000-HTS
1  ATGSQ500000-HTS
2  ATGSQ500000-HTS
3  ATGSQ500000-HTS
4      BRA46065-HTS

First five records of columns after encoding:
      Encoded_BuyerName  Encoded_Description  Encoded_ProductID
0                9                3242                157
1                9                3242                157
2                9                3242                157
3                9                3242                157
4               20                2388                284

Encoding completed and saved to encoded_sales_data.xlsx

```

Figure 2: Encoded dataset

(iii) Feature Engineering and Selecting

Feature engineering is a crucial step, where new features are created or selected from existing ones to improve the performance and accuracy of a model. It involves transforming the raw data into a format that is more suitable for machine learning algorithms.

- **Feature Creation:** Feature creation is a vital aspect which can significantly impact the performance of predictive models. In the context of data analysis and machine learning, feature creation refers to the process of generating new attributes or variables from the existing ones to extract more meaningful and relevant information. This step aims to enhance the dataset's capability to capture important patterns, relationships, or insights that may not be readily apparent in the original features.

The primary objective of introducing "InvoiceValue (LKR)" to the dataset is to quantify and represent the value associated with each invoice in Sri Lankan Rupees (LKR). In B2B sales, the total value of an invoice can be a critical factor in understanding sales trends and predicting

future sales performance. The "InvoiceValue (LKR)" feature is computed by $\text{ExchangeRate} * \text{InvoiceAmount (USD)}$ (Figure 3). The total invoice value provides context about the scale and magnitude of each transaction, which can be valuable when making predictions or identifying patterns related to high-value or low-value sales.

```
import pandas as pd

file_path = 'encoded_dataset.xlsx'
df = pd.read_excel(file_path)

df['InvoiceValue (LKR)'] = df['InvoiceAmount (USD)'] * df['ExchangeRate']

new_file_path = 'new_encoded_dataset.xlsx'
df.to_excel(new_file_path, index=False)

print(f"Dataset with 'InvoiceValue (LKR)' column saved to {new_file_path}")
```

Dataset with 'InvoiceValue (LKR)' column saved to new_encoded_dataset.xlsx

Figure 3: Creating new column called InvoiceValue(LKR)

- **Feature Transformation:** Data transformation is the process of altering the original dataset to meet specific analytical or modeling requirements. This process enhances the dataset's usability, facilitates better insights, and ensures that it conforms to the assumptions of the analysis or modeling techniques to be applied. Common transformations include logarithmic or exponential transformations, scaling, or normalization.

In this study, feature transformation was performed on the dataset to facilitate feature selection. An issue was encountered during feature selection due to the presence of non-numeric data in the dataset. This resulted in a "TypeError: float() argument must be a string or a number, not 'Timestamp'" error. To address this, feature transformation was carried out on the "InvoiceDate" column, which originally contained both date and time information prior to the feature selection process. This transformation involved converting "InvoiceDate" into date, day, month, year without time, simplifying the data for analysis and resolving the inconsistency in data types.

After encoding the dataset, it was determined that there are 158 unique BuyerNames, 3765 unique Description, and 3785 unique ProductIDs present. Additionally, there were multiple transactions recorded for the same day. To address this, the decision was made to calculate the mean values for the 'Qty,' 'ExchangeRate,' and 'UnitPrice' for each day. Based on these

calculated values, the 'InvoiceAmount' was regenerated. Furthermore, in order to obtain meaningful results, it was determined that the 'BuyerNames,' 'Descriptions,' and 'ProductID' columns should be removed from the dataset. As a result of these data preprocessing steps, the dataset now comprises 1916 records.

- **Feature Selection:** Performing feature selection for a dataset involves identifying and choosing the most relevant and informative attributes (features) while discarding those that may not contribute significantly to the analysis or modeling process. Choosing the right features are crucial as it can help improve model performance, reduce overfitting, and enhance interpretability while reducing computational complexity. This can be done through various techniques, such as statistical tests, feature importance ranking, or with the help of domain expertise.

First, the target variable is set to 'InvoiceAmount (USD),' which is the column needs to be predicted. A set of features is then selected to serve as independent variables for the machine learning model. Following the selection of the target variable, feature selection is carried out using the SelectKBest method. This method identifies the top 'k' features with the strongest correlation to the target variable. To achieve this, the 'f_regression' scoring function is applied, which is appropriate for regression tasks.

The feature selection process identified the top five features based on their correlation with the target variable are as follows. They were chosen to be included in the machine learning model.

```
Selected Features: Index(['UnitPrice', 'Qty', 'ExchangeRate', 'InvoiceValue (LKR)', 'Year'], dtype='object')
```

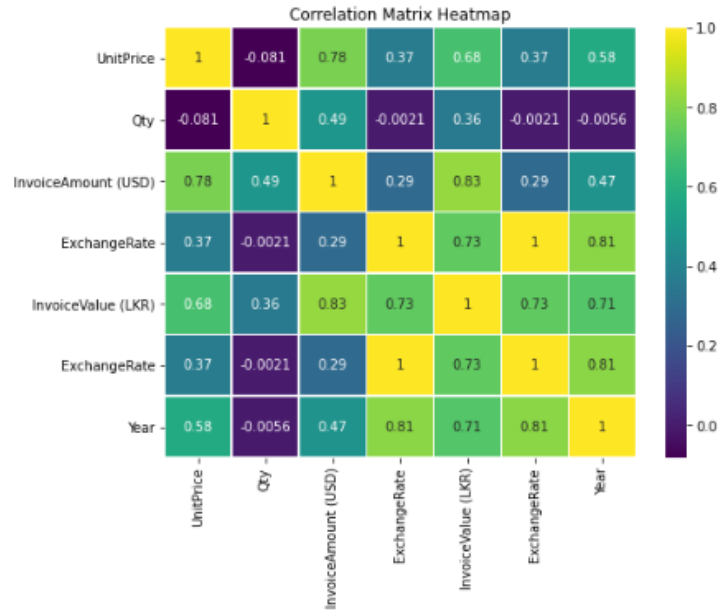


Figure 4: Correlation heat map for time series variables

- An exploratory analysis was conducted on the dataset, focusing on continuous variables such as 'UnitPrice,' 'Qty,' 'ExchangeRate,' 'InvoiceValue (LKR),' and 'Year.' A heat map was generated to visually assess the pairwise correlations between these variables. As per the Figure 4, heat map, below significant correlations indicate that changes in these pairs of variables tend to move together, either positively or negatively, in a meaningful and consistent way.
 1. UnitPrice and InvoiceAmount (USD) have a strong positive correlation of 0.782285. This suggests that as the unit price increases, the invoice amount in USD also tends to increase significantly. This correlation is significant.
 2. Qty and InvoiceAmount (USD) have a moderate positive correlation of 0.486724. This indicates that as the quantity increases, the invoice amount in USD also tends to increase. While not extremely strong, this correlation is still significant.
 3. InvoiceAmount (USD) and InvoiceValue (LKR) have a strong positive correlation of 0.826814. This suggests that as the invoice amount in USD increases, the

corresponding invoice value in Sri Lankan Rupees (LKR) tends to increase significantly. This correlation is significant.

4. UnitPrice and InvoiceValue (LKR) have a strong positive correlation of 0.681052. This means that as the unit price increases, the invoice value in LKR tends to increase significantly. This correlation is significant.
5. ExchangeRate and InvoiceValue (LKR) have a strong positive correlation of 0.734362. As the exchange rate increases, the invoice value in LKR tends to increase significantly. This correlation is significant.

(iv) Model Selection

(a) Time-Series Model Selection

Time-Series Model selection is a pivotal stage within time series analysis, where the selection of the most suitable forecasting model holds the potential to greatly influence prediction accuracy. This process entails a comprehensive evaluation of the dataset, and informed by the insights from the literature review, we explore five distinct models, each offering its specific advantages and considerations. Through this exploration, the study aim to identify the optimal approach for achieving precise and reliable sales trend forecasts.

➤ **ARIMA (AutoRegressive Integrated Moving Average):**

ARIMA, which stands for AutoRegressive Integrated Moving Average, functions as a resilient and robust time series forecasting model that seamlessly blends autoregressive (AR) and moving average (MA) components with differencing techniques to render time series data stationary. This methodology finds particular applicability in sales data analysis due to its adeptness at addressing both short-term and long-term dependencies within the data. It excels in capturing intricate patterns, including seasonality and trends, thereby emerging as a valuable asset for sales forecasting. ARIMA stands out as a formidable tool in the realm of time series forecasting due to

its ability to comprehensively capture diverse facets of the data. It skillfully discerns trends, whether they exhibit linear or nonlinear characteristics, making it especially well-suited for modeling extended-term dependencies. Furthermore, ARIMA accommodates data exhibiting seasonal patterns by integrating seasonal differencing, often represented as 'S' in the context of SARIMA, an extension of ARIMA tailored for seasonal data. The MA component within ARIMA proves instrumental in modeling short-term dependencies and mitigating the influence of noise inherent in the data.

- **AutoRegressive (AR) Component:** This component captures the relationship between the current value in the time series and its past values. It assumes that the current value can be predicted based on a linear combination of its previous values. The "p" parameter in ARIMA, denoted as AR(p), specifies the number of lagged observations included in the model.
- **Integrated (I) Component:** This component focuses on differencing the time series data to make it stationary. Stationarity is essential because many time series models, including ARIMA, assume that the data are stationary, meaning that statistical properties like mean and variance do not change over time. The "d" parameter in ARIMA, denoted as I(d), represents the number of differencing operations required to achieve stationarity. If the data are already stationary, $d=0$.
- **Moving Average (MA) Component:** The MA component models the relationship between the current value and past white noise (random) errors in the time series. It helps account for any short-term fluctuations or noise in the data. The "q" parameter in ARIMA, denoted as MA(q), specifies the number of lagged forecast errors included in the model.

➤ **SARIMA (Seasonal ARIMA):**

Seasonal ARIMA, abbreviated as SARIMA, represents extended version of the ARIMA framework, specifically designed to address the presence of seasonality in time series data. SARIMA exhibits exceptional efficacy in the context of sales data analysis, especially when

dealing with conspicuous seasonal patterns. The key distinguishing feature of SARIMA lies in its capacity to incorporate seasonal components. These components empower SARIMA to comprehend and effectively model recurring patterns within the data. By doing so, SARIMA can discern and forecast short-term and long-term fluctuations in sales data, thus enhancing its resilience and accuracy in capturing the intricacies of retail sales dynamics.

- **Seasonal Component (S):** Unlike standard ARIMA, SARIMA includes a seasonal component, denoted as "S," which represents the seasonal part of the time series. This component captures the repeating patterns that occur at regular intervals, such as daily, weekly, monthly, or yearly seasonality. It helps account for the impact of these seasonal patterns on the data

➤ **SARIMAX (Seasonal ARIMA with Exogenous Variables):**

SARIMAX, which stands for Seasonal Autoregressive Integrated Moving Average with Exogenous Variables, is a powerful time series forecasting model that builds upon the SARIMA framework by incorporating exogenous variables.

- **Exogenous Variables:** The key feature of SARIMAX is the inclusion of exogenous variables. Exogenous variables are external factors or predictors that are not part of the time series itself but can have a significant impact on the observed data. In this research exogenous variable is economic indicator exchange rate. This external factor is believed to have a causal relationship with the time series of interest.

(b) Machine Learning Model Selection

With the capacity to analyze large volumes of sales data, machine learning models can identify intricate patterns, predict future trends, and recommend data-driven strategies. Whether it's predicting customer demand, optimizing pricing strategies, or improving inventory management, machine learning models play a pivotal role in maximizing sales and revenue.

Though there are two types of machine learning models, in this study, we have chosen regression machine learning models over classification models. The reason to choose regression machine

learning model is the target variable in this study is 'InvoiceAmount (USD),' which is a continuous variable. Classical models, often referred to as classical statistical methods, are typically designed for categorical or discrete variables. They may not be well-suited for modeling and predicting continuous numeric values like 'InvoiceAmount (USD).' Regression models, on the other hand, are specifically designed for this purpose, making them a more appropriate choice.

Based on the literature review, below are the models chosen to conduct the analysis.

➤ **Random Forest Regression:**

Random Forest Regression is a versatile and powerful ensemble learning technique that addresses a wide range of regression problems. By combining predictions from multiple decision trees, it effectively mitigates the risk of overfitting, resulting in a more robust and accurate regression model. Its adaptability to diverse data types and complexities makes it suitable for both simple and complex regression tasks. Random Forest Regression excels in capturing intricate nonlinear relationships between predictors and the target variable, a vital capability for real-world data with complex patterns. It exhibits resilience to outliers and missing data, ensuring dependable predictions even with noisy or incomplete datasets. The method also provides a feature importance ranking, aiding feature selection and uncovering influential factors for the target variable. Known for its ability to generalize effectively, Random Forests maintain consistent performance on unseen data, reducing the likelihood of overfitting. Furthermore, fine-tuning hyperparameters can further enhance its performance, solidifying its position as a versatile and influential tool in regression analysis.

➤ **Gradient Boosting (e.g., XGBoost, LightGBM) Regression:**

Gradient Boosting, including popular implementations like XGBoost and LightGBM, is a powerful regression technique known for its exceptional predictive accuracy. Unlike traditional decision trees, Gradient Boosting builds an ensemble of decision trees sequentially, each one correcting the errors of the previous tree. This sequential learning process allows Gradient Boosting models to capture intricate relationships within the data, making them particularly effective for regression tasks involving complex, high-dimensional datasets. One of the key advantages of Gradient

Boosting regression is its ability to handle various data types, including numerical and categorical variables, without the need for extensive data preprocessing. This versatility simplifies the modeling process and can save valuable time. Additionally, Gradient Boosting models are robust to outliers and can effectively handle missing data, making them suitable for real-world datasets that may contain noise or incomplete information. They also provide insights into feature importance, helping analysts identify which variables have the most significant impact on the regression predictions.

➤ **Long Short-Term Memory Model:**

Long Short-Term Memory (LSTM) is a remarkable recurrent neural network (RNN) architecture designed to overcome the limitations faced by traditional RNNs in handling sequential data. Unlike standard neural networks, LSTM integrates feedback connections, making it adept at processing entire sequences of data rather than individual data points. This property enables LSTM to excel in capturing long-term dependencies and patterns in sequential data, such as time series, text, and speech. One of LSTM's fundamental innovations is the introduction of a memory cell, a specialized container capable of retaining information over extended periods. This memory cell is governed by three gates—the input gate, forget gate, and output gate—each serving a distinct purpose in determining what information to add, remove, or output from the cell. This selective information flow allows LSTM networks to effectively learn and model long-term dependencies, making them invaluable tools for tasks like language translation, speech recognition, and time series forecasting. LSTM's capacity to work with sequential data has driven breakthroughs in artificial intelligence and deep learning across various domains.

(v) Model Validation

The primary goal of this research is to define a dependable model for forecasting future sales, ensuring precision and reliability in predictions. In the context of machine learning and time series data analysis, particularly with regression techniques, the main objective is to achieve precise predictions of continuous values. To assess the performance of chosen regression models, following evaluation methods have been employed. Further, apart from the below-mentioned

statistical measurements, the real-world evaluation will be done by the industry experts and the top management of the stakeholders at Expo Group of Industries PVT (LTD.).

- **Mean Absolute Error (MAE)** - MAE is a metric used to assess the accuracy of predictions or forecasts by measuring the average absolute difference between predicted values and actual observed values. It provides a straightforward measure of how reliable the forecasts are, with lower MAE values indicating better predictive accuracy.
- **Mean Absolute Percentage Error (MAPE)** – MAPE is an indicator of forecasting accuracy expressed as a percentage. It quantifies the average absolute percentage difference between predicted and actual values. MAPE is particularly useful when you want to evaluate the forecast's performance in terms of the magnitude of errors relative to the actual values.
- **Mean Squared Error (MSE)** - MSE is a metric that calculates the average of the squared differences between predicted values and actual values. It provides a way to penalize larger errors more heavily than smaller ones. Like MAE, lower MSE values suggest better predictive accuracy, but it gives greater weight to outliers.
- **Root Mean Square Error (RMSE)** - RMSE is a variant of MSE that is often used because it provides a measure of error in the same units as the original data. It's the square root of the MSE and gives an idea of the typical size of errors made by the forecasting model. RMSE is valuable for comparing the accuracy of different models or assessing the overall fit of a model to the data.

3.2. High-level Architecture Diagram

The following high-level architecture diagram visually represent the overall framework of the study, highlighting the processes involved, outlining the different stages involved, from data collection and analysis to interpretation and conclusion.

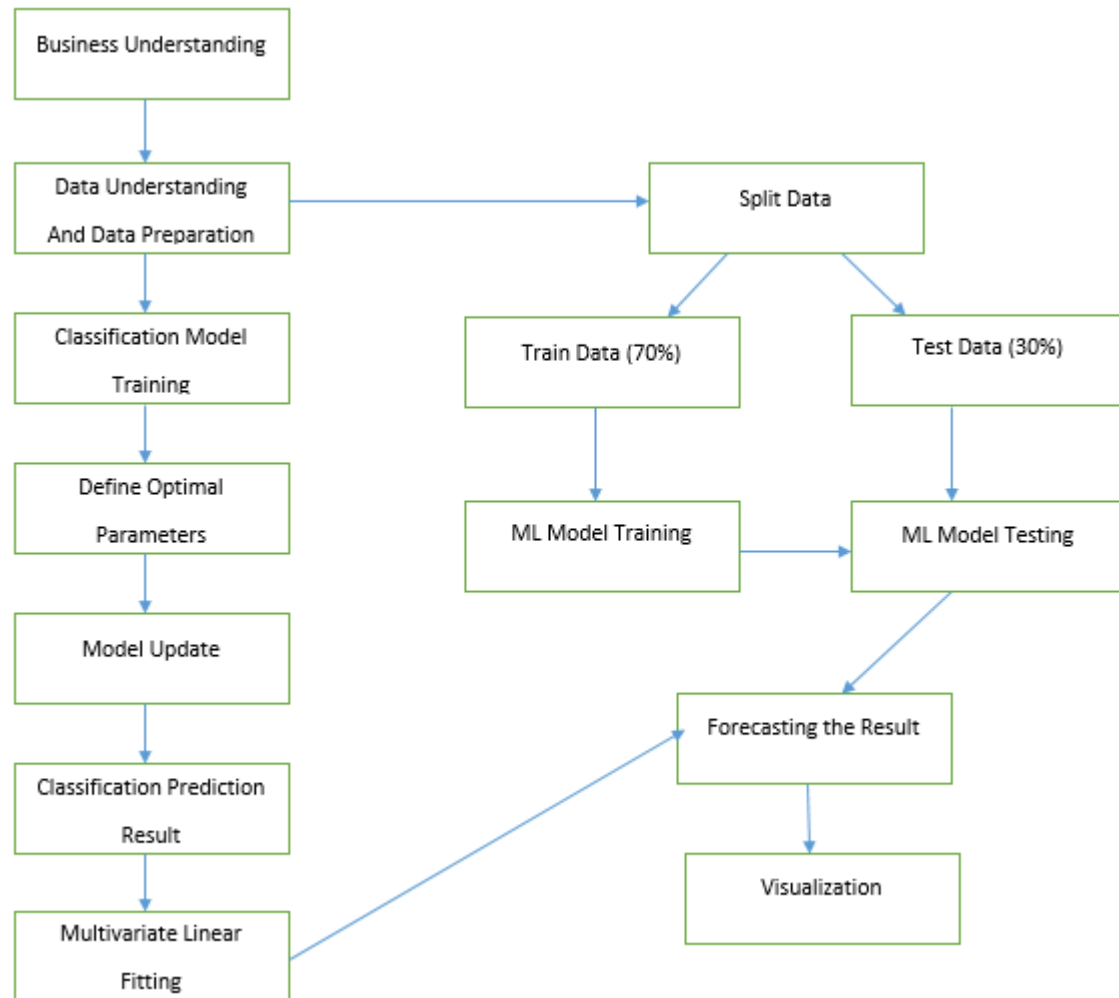


Figure 5: High-level Architecture Diagram

CHAPTER 4

RESULTS AND EVALUATION

The evaluation section adopts a comprehensive approach by considering three distinct time series models and applying three diverse machine learning models to effectively forecast sales within the apparel industry. This intricate evaluation process encompasses several essential components. Firstly, it involves the meticulous preprocessing of time series data, addressing issues such as missing values, outliers, and scaling to ensure that the datasets are in optimal condition for modeling. Subsequently, the evaluation proceeds with the individual training and testing of each of the five machine learning models using the dataset. This process incorporates the utilization of rigorous cross-validation techniques aimed at minimizing bias and ensuring the robustness of the findings. To assess the performance of each model, a range of evaluation metrics is employed to identify the model that consistently outperforms the others when applied to the sales dataset. While performing data preprocessing step, it was determined that the dataset contained upper-bound outliers amounting to 0.97%. To ensure the accuracy of the outcomes and machine learning models, it was decided to remove these outliers. Following their removal, the dataset now consists of 76,464 records. Further, multiple transactions were recorded for the same day and to streamline the dataset for meaningful analysis, averages were then used to regenerate 'InvoiceAmount' resulting in 1916 records.

4.1. Time Series Model Evaluation

4.1.1. ARIMA (AutoRegressive Integrated Moving Average) Results

The crucial consideration in the ARIMA model is the assumption of data stationarity. When dealing with non-stationary data, the model's ability to accurately capture underlying patterns becomes compromised, resulting in inaccurate outcomes and insights. Hence, it is essential to confirm the model's appropriateness before commencing the analysis. This confirmation can be accomplished through two approaches: utilizing Rolling Statistics and performing the Augmented Dickey-Fuller Test.

i. Rolling Statistics and Augmented Dickey-Fuller (ADF) Test

In time series analysis, the combination of Rolling Statistics and the Augmented Dickey-Fuller (ADF) Test serves as a powerful tool to assess and characterize the stationarity of a time series dataset.

Rolling statistics involve the calculation of key statistical measures, such as the mean and standard deviation, over a moving window of data points within the time series. This technique helps visualize how these statistics change over time and reveals trends or patterns that might not be apparent when looking at the entire data set. Notably, alterations in the rolling mean or standard deviation - whether ascending or descending - can serve as indicators of evolving data characteristics. Stationarity often aligns with a relatively consistent and unchanging rolling mean and standard deviation.

On the other hand, Augmented Dickey-Fuller (ADF) Test, is a formal statistical method used to determine whether the time series is stationary or not. Moreover, more negative ADF Statistic indicates stronger evidence against a unit root and supports stationarity and small p-value (typically below a significance level 0.05) suggests that the data is stationary. The reason being is that we need differencing only if the series is non-stationary.

When analyzing the chart illustrating the rolling mean and standard deviation, it helps detect periods of stability or stationarity within the data. As illustrated in the visualization below (Figure 6), it is clear that both the rolling mean and rolling standard deviation display a consistent pattern over time. Consequently, the analysis indicates that the time series data is non-stationarity.

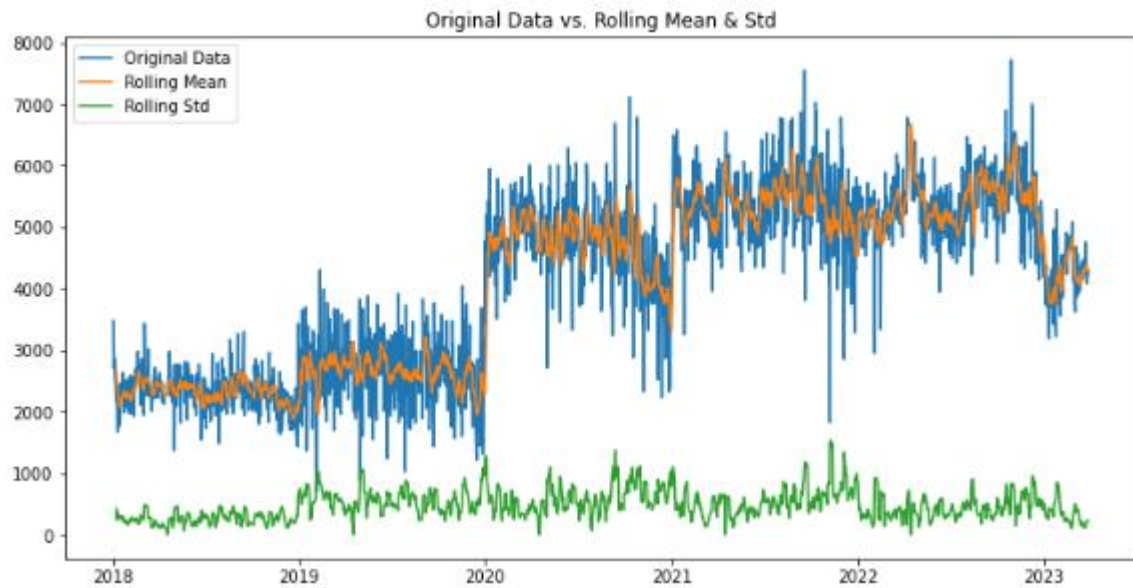


Figure 6: Rolling Mean and Standard Deviation test result

Table 4: ADF test results

ADF Statistic	p-value	1% Critical Value	5% Critical Value	10% Critical Value
-2.0039	0.2848	-3.4304	-2.8615	-2.5667

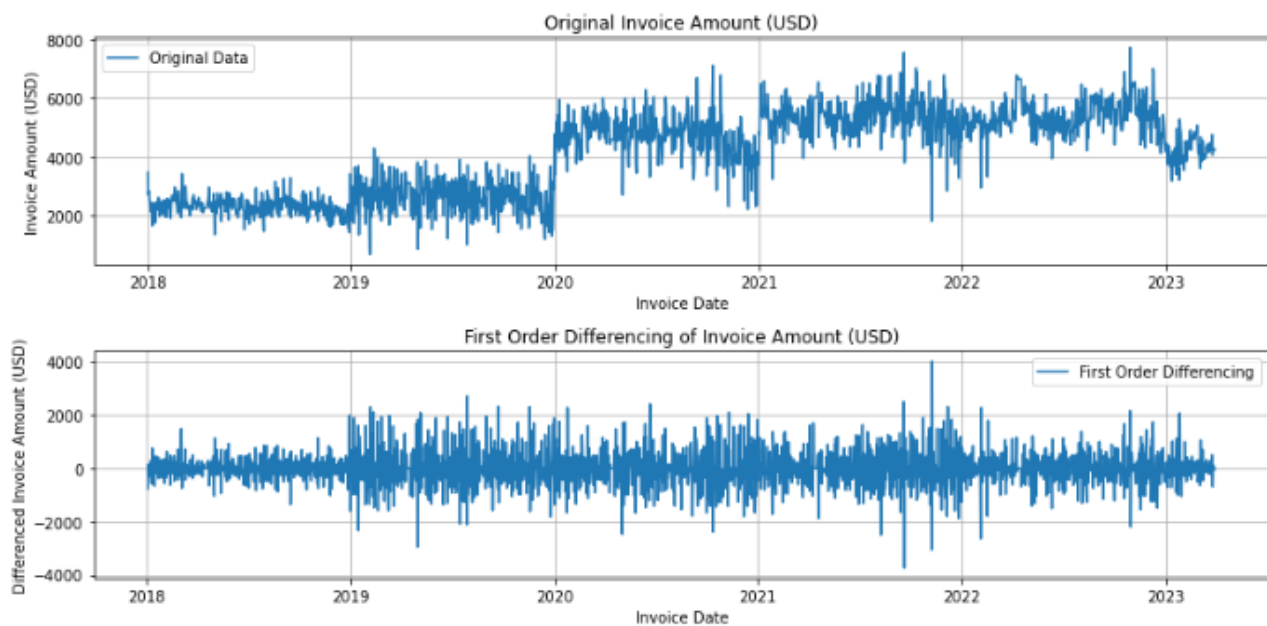
As per the Table 4, ADF statistic registers at -2.0039, which exceeds the critical values at both the 5% significance level (-2.8616) and the 1% significance level (-3.4305) respectively. This provides strong evidence of non-stationarity. Furthermore, the calculated p-value of 0.2848 is greater than the common significance level of 0.05, further supporting the conclusion of a non-stationary data series. It is highly suggestive that the null hypothesis cannot be rejected. Therefore, the above statistical evidence firmly supports that time series is deemed non-stationary.

ii. Differencing data for non-stationary

Based on the observations mentioned earlier (Table 4), it is evident that the provided data exhibits non-stationary characteristics. Therefore, differencing is required, to make it stationary and to

effectively utilize it for ARIMA and SARIMA analysis. Achieving stationarity is imperative for obtaining accurate estimates from the models. A widely employed technique to facilitate this transformation is differencing.

Differencing involves the subtraction of the current value in a time series from either its previous value or a lagged value. For instance, in the context of a monthly sales time series, differencing is achieved by subtracting the sales of the prior month from the current month's sales. This operation yields a new time series representing the month-to-month change in sales. The process can be iterated to obtain higher-order differences, such as the change in the change in sales, and so forth. The primary objective is to eliminate any trend or seasonality that renders the original time series



non-stationary.

Figure 7: Before and after Differencing

As shown in the Figure 7, Differencing stabilized the mean of the dataset by removing changes in the level of a time series, and therefore eliminating (or reducing) trend and seasonality.

After differencing, the next step involves testing the stationarity or non-stationarity of the data. This evaluation aims to determine whether further differencing is required or if the data has already achieved the desired stationary state. To assess stationarity, ADF testing is repeated to reevaluate the stationarity or non-stationarity of the differenced time series data.

ADF Statistic	p-value	1% Critical Value	5% Critical Value	10% Critical Value
-11.1293	3.3127e-20	-3.4304	-2.8615	-2.5667

Table 5: ADF test results after First Differencing

After the first differencing, as indicated in the Table 5, ADF statistic registers at -11.1293, which is notably smaller than the critical values at both the 5% significance level (-2.8616) and the 1% significance level (-3.4305) respectively. This provides robust evidence of stationarity. Moreover, the calculated p-value of 3.3127e-20 is significantly smaller than the typical significance level of 0.05, further reinforcing the assertion of a stationary data series. The evidence is compelling enough to confidently reject the null hypothesis. Therefore, it can be concluded that, following the first differencing, the time series data has achieved stationarity. The value of differencing (d) can be confirmed as d=1.

The ARIMA parameters (p, d, q), which represents Auto Regressive (AR), differencing and Moving Average (MA) components respectively, have been derived from the patterns observed in the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF). These functions aid in the estimation of the parameters essential for making forecasts using the ARIMA model.

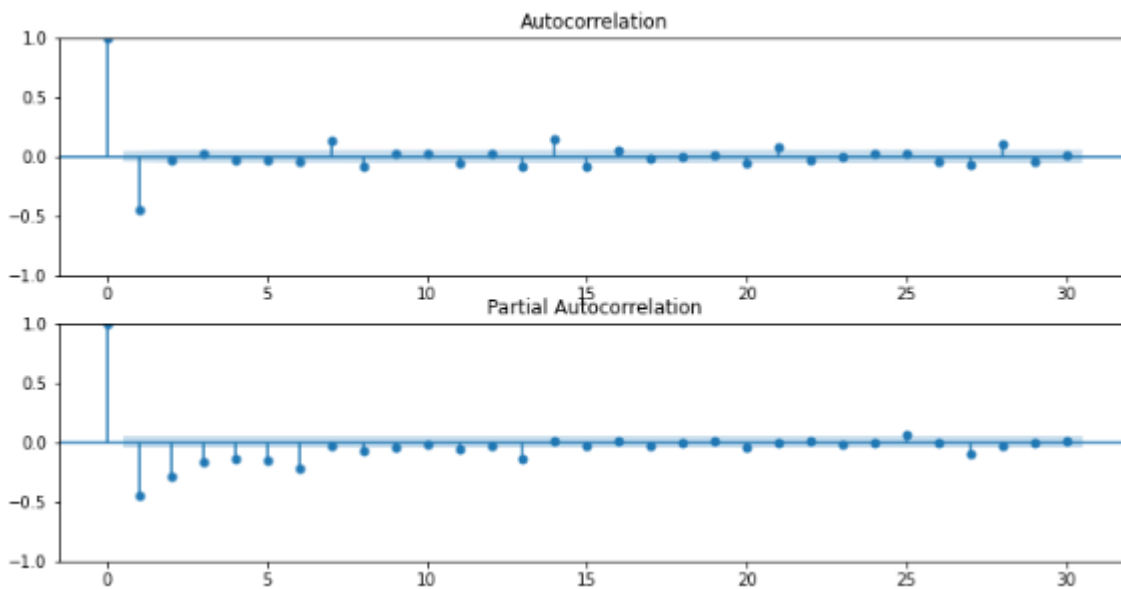


Figure 8: ACF and PACF plots after first differencing

Based on the ACF and PACF plots (Figure 8), after first differencing, p, q, d values are as follows.

p = 2, d = 1 and q= 1

Since the data is non-stationary, it is consider adjusting the auto-ARIMA hyperparameters to address this issue. It is an automatic module that helps to define the max p, max q, and d values and find the optimal value using the error matrix called “AIC (Akaike Information Criteria)” which quantifies the goodness of fit of the model (Figure 9). In this study the dataset used is non-stationary, but differencing is included in the modeling process (by specifying d=1 and max_d=1), and making stationary=True allowing the model to handle the differencing as part of the modeling process.

Therefore, a modified configuration for appropriate differencing parameters are as follows:

```
model = auto_arima(train_data,
                    start_p=1, start_q=1,
                    test='ocsb',
                    max_p=5, max_q=5, max_d=1, # Include differencing order (d)
                    m=1,
                    d=1, # Include differencing order (d)
                    seasonal=True,
                    stationary=True,
                    start_P=0,
                    D=None,
                    trace=True,
                    error_action='ignore',
                    suppress_warnings=True,
                    stepwise=True)
```

Performing stepwise search to minimize aic

```
ARIMA(1,0,1) (0,0,0) [0] intercept : AIC=23882.824, Time=2.61 sec
ARIMA(0,0,0) (0,0,0) [0] intercept : AIC=26611.713, Time=0.05 sec
ARIMA(1,0,0) (0,0,0) [0] intercept : AIC=24480.510, Time=0.15 sec
ARIMA(0,0,1) (0,0,0) [0] intercept : AIC=25658.262, Time=0.51 sec
ARIMA(0,0,0) (0,0,0) [0]          : AIC=29828.816, Time=0.03 sec
ARIMA(2,0,1) (0,0,0) [0] intercept : AIC=23874.530, Time=1.77 sec
ARIMA(2,0,0) (0,0,0) [0] intercept : AIC=24175.060, Time=0.18 sec
```

```

ARIMA(3,0,1)(0,0,0)[0] intercept : AIC=23886.038, Time=2.42 sec
ARIMA(2,0,2)(0,0,0)[0] intercept : AIC=inf, Time=2.26 sec
ARIMA(1,0,2)(0,0,0)[0] intercept : AIC=23877.137, Time=2.02 sec
ARIMA(3,0,0)(0,0,0)[0] intercept : AIC=24051.559, Time=0.34 sec
ARIMA(3,0,2)(0,0,0)[0] intercept : AIC=inf, Time=2.62 sec
ARIMA(2,0,1)(0,0,0)[0] intercept : AIC=23876.231, Time=0.37 sec

```

Best model: ARIMA(2,0,1)(0,0,0)[0] intercept
Total fit time: 15.337 seconds

Figure 9: Hyperparameters based on Akaike Information Criteria

The dataset is divided in to 80:20 ratio to find the best model to predict sale data. Since ARIMA models are known to perform well with stationary data, the dataset used here has undergone differencing to remove trends and seasonal patterns. As depicted in Figure 10, a forecast was conducted on the training data to assess the model's accuracy.

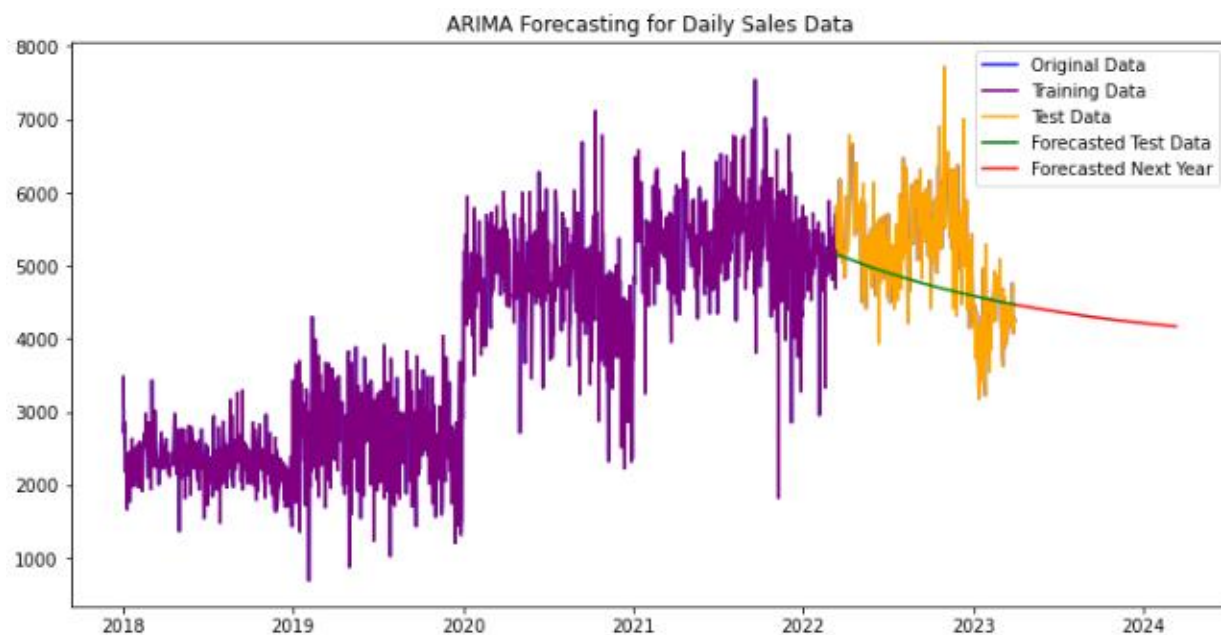


Figure 10: Forecasted Model Using ARIMA

The model is evaluated using MAE, MSE, RMSE and MAPE values where MAE is a fundamental metric for measuring predictive accuracy, MSE computes the average of the squared differences between predicted and actual values, RMSE is useful for comparing the accuracy of different models or assessing a model's overall fit to the data and MAPE expresses forecasting accuracy as a percentage.

Model parameters are as follows.

```
Mean Absolute Error (MAE): 642.77  
Mean Squared Error (MSE): 657349.30  
Root Mean Square Error (RMSE): 810.77  
Mean Absolute Percentage Error (MAPE): 11.92%
```

The above mentioned MAE, MSE, RMSE, and MAPE values indicate that the ARIMA model's predictions aren't highly accurate. The presence of a descending curve in both test data and future forecasts generated by an ARIMA model (Figure 10) suggests the model's shortcomings in capturing intrinsic data patterns. Such a scenario could arise when the dataset lacks evident autocorrelation or seasonality, elements that ARIMA models aim to identify. To enhance forecast accuracy in such cases, it might be prudent to explore alternative modeling approaches.

4.1.2. SARIMA (Seasonal ARIMA)

SARIMA also uses past values but takes into account any seasonality patterns in the data. Since SARIMA brings in seasonality as a parameter, it's significantly more powerful than ARIMA in forecasting complex data spaces which containing cycles. The Autoregressive (AR), Integrated (I), and Moving Average (MA) parts of the model remain as that of ARIMA. The addition of Seasonality adds robustness to the SARIMA model.

As SARIMA and SARIMAX are well-suited for handling seasonal data, the initial step involved an assessment of the data's seasonality and trend. As depicted in the output below (Figure 11), it is evident that the data exhibits seasonality. This observation sheds light on why ARIMA performed inadequately in forecasting.

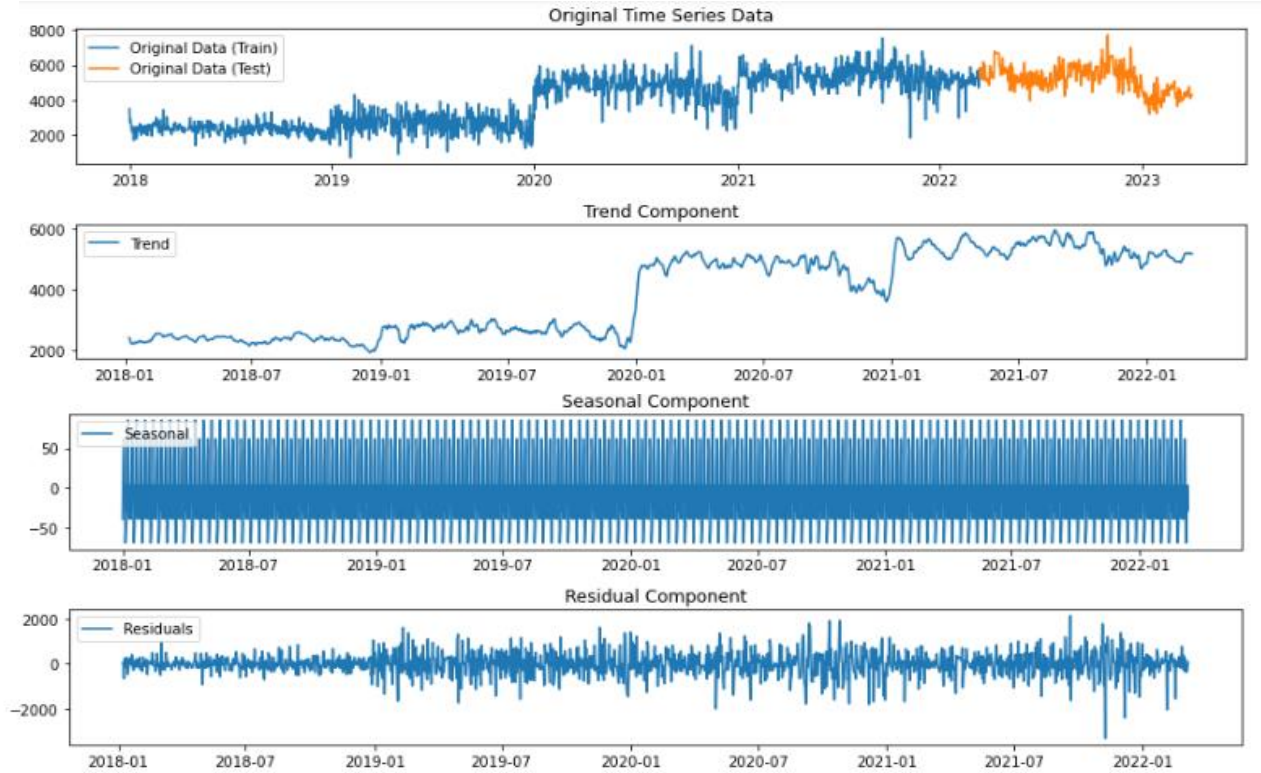


Figure 11: Seasonal and Trend decomposition using Loess graph

To maintain the accuracy, hyperparameters have been generated using auto-ARIMA method. It is identified $(5,1,1)(0,0,0)[12]$ is the best model to use in SARIMA forecasting (Figure 12).

Performing stepwise search to minimize aic

```

ARIMA(2,1,2) (1,0,1) [12] intercept : AIC=23857.178, Time=7.76 sec
ARIMA(0,1,0) (0,0,0) [12] intercept : AIC=24566.945, Time=0.06 sec
ARIMA(1,1,0) (1,0,0) [12] intercept : AIC=24192.065, Time=0.72 sec
ARIMA(0,1,1) (0,0,1) [12] intercept : AIC=23861.674, Time=3.15 sec
ARIMA(0,1,0) (0,0,0) [12] intercept : AIC=24564.949, Time=0.05 sec
ARIMA(2,1,2) (0,0,1) [12] intercept : AIC=23854.119, Time=5.38 sec
ARIMA(2,1,2) (0,0,0) [12] intercept : AIC=23857.087, Time=2.49 sec
ARIMA(2,1,2) (0,0,2) [12] intercept : AIC=23856.009, Time=16.92 sec
ARIMA(2,1,2) (1,0,0) [12] intercept : AIC=23854.094, Time=6.03 sec
ARIMA(2,1,2) (2,0,0) [12] intercept : AIC=23856.019, Time=16.08 sec
ARIMA(2,1,2) (2,0,1) [12] intercept : AIC=23856.807, Time=23.03 sec
ARIMA(1,1,2) (1,0,0) [12] intercept : AIC=23860.572, Time=3.47 sec
ARIMA(2,1,1) (1,0,0) [12] intercept : AIC=23852.460, Time=5.00 sec
ARIMA(2,1,1) (0,0,0) [12] intercept : AIC=23851.945, Time=1.52 sec
ARIMA(2,1,1) (0,0,1) [12] intercept : AIC=23852.480, Time=5.29 sec
ARIMA(2,1,1) (1,0,1) [12] intercept : AIC=inf, Time=9.03 sec
ARIMA(1,1,1) (0,0,0) [12] intercept : AIC=23853.469, Time=2.72 sec

```



```

ARIMA (2,1,0) (0,0,0) [12] intercept : AIC=24053.653, Time=0.71 sec
ARIMA (3,1,1) (0,0,0) [12] intercept : AIC=23851.441, Time=2.54 sec
ARIMA (3,1,1) (1,0,0) [12] intercept : AIC=23852.296, Time=7.24 sec
ARIMA (3,1,1) (0,0,1) [12] intercept : AIC=23853.006, Time=7.12 sec
ARIMA (3,1,1) (1,0,1) [12] intercept : AIC=23860.297, Time=7.40 sec
ARIMA (3,1,0) (0,0,0) [12] intercept : AIC=24015.455, Time=0.25 sec
ARIMA (4,1,1) (0,0,0) [12] intercept : AIC=23850.674, Time=3.62 sec
ARIMA (4,1,1) (1,0,0) [12] intercept : AIC=23851.188, Time=8.02 sec
ARIMA (4,1,1) (0,0,1) [12] intercept : AIC=23851.446, Time=9.19 sec
ARIMA (4,1,1) (1,0,1) [12] intercept : AIC=23858.215, Time=9.23 sec
ARIMA (4,1,0) (0,0,0) [12] intercept : AIC=23987.528, Time=0.91 sec
ARIMA (5,1,1) (0,0,0) [12] intercept : AIC=23847.931, Time=5.20 sec
ARIMA (5,1,1) (1,0,0) [12] intercept : AIC=23848.554, Time=10.24 sec
ARIMA (5,1,1) (0,0,1) [12] intercept : AIC=23848.782, Time=9.55 sec
ARIMA (5,1,1) (1,0,1) [12] intercept : AIC=23853.256, Time=15.61 sec
ARIMA (5,1,0) (0,0,0) [12] intercept : AIC=23947.402, Time=2.08 sec
ARIMA (5,1,2) (0,0,0) [12] intercept : AIC=23849.894, Time=17.50 sec
ARIMA (4,1,2) (0,0,0) [12] intercept : AIC=23852.724, Time=6.62 sec
ARIMA (5,1,1) (0,0,0) [12] intercept : AIC=23850.100, Time=2.06 sec

```

Best model: ARIMA(5,1,1)(0,0,0)[12] intercept
Total fit time: 234.659 seconds

Figure 12: Hyperparameters based on Akaike Information Criteria

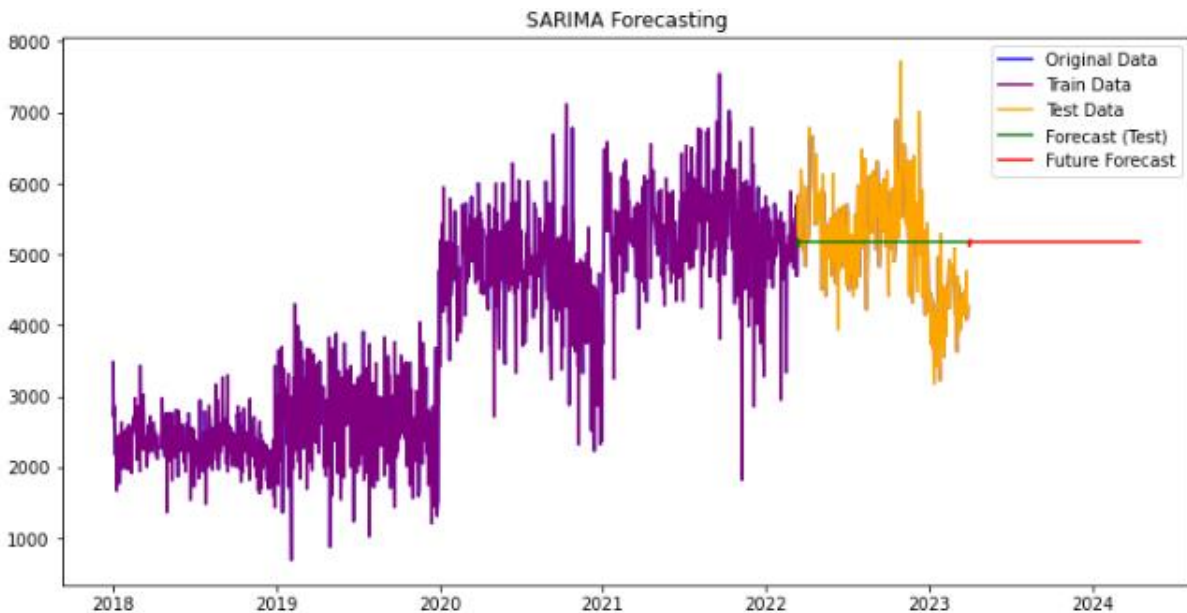


Figure 13: SARIMA forecasting model

The model evaluation parameters are as follows.

```
Mean Absolute Error (MAE) for Test Data: 616.78
Mean Squared Error (MSE) for Test Data: 585796.21
Root Mean Squared Error (RMSE) for Test Data: 765.37
Mean Absolute Percentage Error (MAPE) for Test Data: 12.52%
```

The above MAE, MSE, RMSE, and MAPE values for the test data suggest that the SARIMA model's predictions are not very accurate. This aligns with the observation of straight lines in both test data forecasting and future forecasting in Figure 13. The straight-line forecasts indicate that the model is failing to capture the essential patterns or variations in the data, likely due to the absence of significant seasonality or autocorrelation. Since SARIMA models are specifically designed to handle seasonality and autocorrelation, their effectiveness is limited when these characteristics are not prominent in the data. Hence, In conclusion, exploring machine learning approaches like Random Forest or XGBoost, which can model complex patterns without relying on specific assumptions about seasonality or autocorrelation will provide an accurate model.

4.1.3. SARIMAX (Seasonal ARIMA with eXogenous factors)

SARIMAX is a generalization of the ARIMA model that considers both seasonality and exogenous variables. SARIMAX models are among the most widely used statistical models for forecasting, with excellent forecasting performance. In the SARIMAX model notation, the parameters p, d, and q represent the autoregressive, differencing, and moving-average components, respectively. P, D, and Q denote the same components for the seasonal part of the model, with m representing the number of periods in each season.

In this study m is considered as 12. p,d,q and P,Q,D values have been generated through auto-arima model to maintain the accuracy and as per AIC output shown in Figure 14, P,Q,D values are (5,1,1)(0,0,0)[12]. The exogenous variables considered for the evaluation are 'UnitPrice', 'Qty' and 'ExchangeRate'.

```
Performing stepwise search to minimize aic
ARIMA(2,1,2)(1,0,1)[12] intercept : AIC=23857.178, Time=6.69 sec
ARIMA(0,1,0)(0,0,0)[12] intercept : AIC=24566.945, Time=0.06 sec
ARIMA(1,1,0)(1,0,0)[12] intercept : AIC=24192.065, Time=0.62 sec
ARIMA(0,1,1)(0,0,1)[12] intercept : AIC=23861.674, Time=2.45 sec
ARIMA(0,1,0)(0,0,0)[12]          : AIC=24564.949, Time=0.03 sec
ARIMA(2,1,2)(0,0,1)[12] intercept : AIC=23854.119, Time=4.68 sec
```

```

ARIMA (2,1,2) (0,0,0) [12] intercept : AIC=23857.087, Time=2.17 sec
ARIMA (2,1,2) (0,0,2) [12] intercept : AIC=23856.009, Time=13.50 sec
ARIMA (2,1,2) (1,0,0) [12] intercept : AIC=23854.094, Time=4.60 sec
ARIMA (2,1,2) (2,0,0) [12] intercept : AIC=23856.019, Time=14.58 sec
ARIMA (2,1,2) (2,0,1) [12] intercept : AIC=23856.807, Time=22.02 sec
ARIMA (1,1,2) (1,0,0) [12] intercept : AIC=23860.572, Time=4.00 sec
ARIMA (2,1,1) (1,0,0) [12] intercept : AIC=23852.460, Time=5.10 sec
ARIMA (2,1,1) (0,0,0) [12] intercept : AIC=23851.945, Time=1.56 sec
ARIMA (2,1,1) (0,0,1) [12] intercept : AIC=23852.480, Time=4.62 sec
ARIMA (2,1,1) (1,0,1) [12] intercept : AIC=inf, Time=8.79 sec
ARIMA (1,1,1) (0,0,0) [12] intercept : AIC=23853.469, Time=1.14 sec
ARIMA (2,1,0) (0,0,0) [12] intercept : AIC=24053.653, Time=0.25 sec
ARIMA (3,1,1) (0,0,0) [12] intercept : AIC=23851.441, Time=2.54 sec
ARIMA (3,1,1) (1,0,0) [12] intercept : AIC=23852.296, Time=7.25 sec
ARIMA (3,1,1) (0,0,1) [12] intercept : AIC=23853.006, Time=6.68 sec
ARIMA (3,1,1) (1,0,1) [12] intercept : AIC=23860.297, Time=7.66 sec
ARIMA (3,1,0) (0,0,0) [12] intercept : AIC=24015.455, Time=0.32 sec
ARIMA (4,1,1) (0,0,0) [12] intercept : AIC=23850.674, Time=3.49 sec
ARIMA (4,1,1) (1,0,0) [12] intercept : AIC=23851.188, Time=8.64 sec
ARIMA (4,1,1) (0,0,1) [12] intercept : AIC=23851.446, Time=8.09 sec
ARIMA (4,1,1) (1,0,1) [12] intercept : AIC=23858.215, Time=9.87 sec
ARIMA (4,1,0) (0,0,0) [12] intercept : AIC=23987.528, Time=0.49 sec
ARIMA (5,1,1) (0,0,0) [12] intercept : AIC=23847.931, Time=3.87 sec
ARIMA (5,1,1) (1,0,0) [12] intercept : AIC=23848.554, Time=10.50 sec
ARIMA (5,1,1) (0,0,1) [12] intercept : AIC=23848.782, Time=8.40 sec
ARIMA (5,1,1) (1,0,1) [12] intercept : AIC=23853.256, Time=13.67 sec
ARIMA (5,1,0) (0,0,0) [12] intercept : AIC=23947.402, Time=4.73 sec
ARIMA (5,1,2) (0,0,0) [12] intercept : AIC=23849.894, Time=7.04 sec
ARIMA (4,1,2) (0,0,0) [12] intercept : AIC=23852.724, Time=6.91 sec
ARIMA (5,1,1) (0,0,0) [12] intercept : AIC=23850.100, Time=2.49 sec

```

Best model: ARIMA (5,1,1) (0,0,0) [12] intercept
Total fit time: 221.819 seconds

Figure 14: Hyperparameters based on Akaike Information Criteria

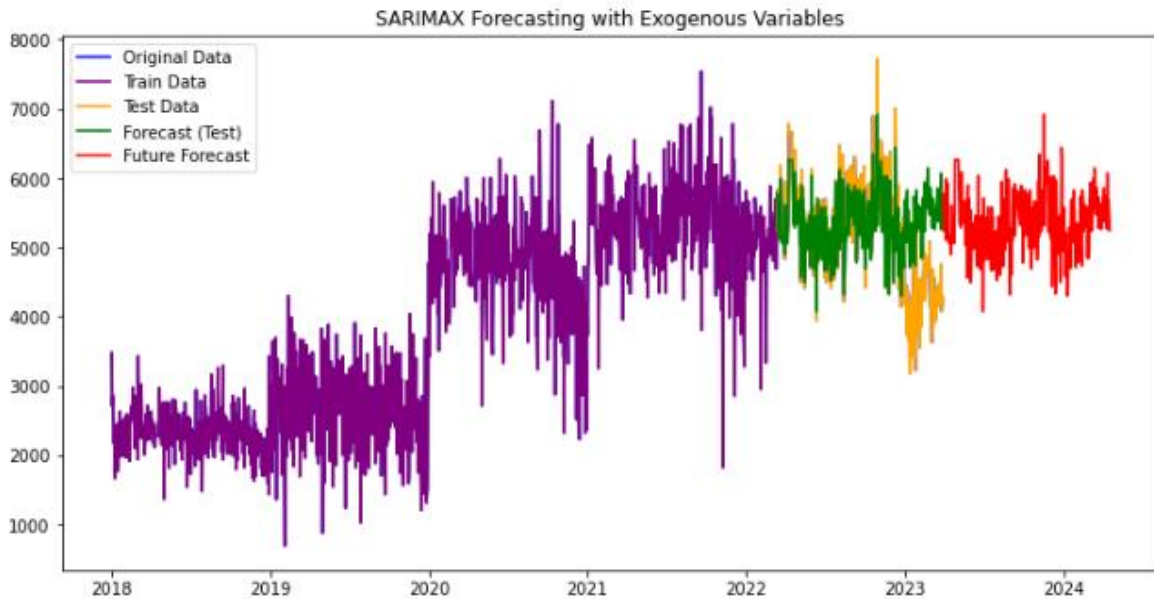


Figure 15: SARIMAX forecasting model

:

The model evaluation parameters are as follows.

Mean Absolute Error (MAE) for Test Data: 408.47
Mean Squared Error (MSE) for Test Data: 411742.93
Root Mean Squared Error (RMSE) for Test Data: 641.67
Mean Absolute Percentage Error (MAPE) for Test Data: 9.11%

SARIMAX leverages the power of exogenous variables, external factors that can impact time series data, to capture seasonality and improve predictive accuracy. By considering these additional features, the model becomes highly adaptable and gains the ability to detect subtle patterns and seasonal fluctuations that might elude a purely data-driven approach. This adaptability is a key strength of SARIMAX, enabling it to generate precise forecasts for both test data and future time points. When relevant exogenous variables are accessible, SARIMAX stands as a valuable asset for robust time series analysis, providing insights that might otherwise remain hidden.

As per the SARIMAX model showed in Figure 15, the model appears to perform well in capturing the overall trend and seasonality of the data, especially in the earlier parts of the test data. However, the underestimation of the recent upward trend in the test data suggests some limitations in the model's ability to predict rapid changes. Further, confidence intervals or prediction ranges are not present in the graph, making it difficult to assess the uncertainty associated with the future

forecast. When it comes to trends and long-term behavior, the original data and the model's forecast both suggest a clear upward trend over time. The model seems to capture this trend reasonably well, although it might underestimate the rate of increase in the future.

In conclusion, the SARIMAX model appears to be a good fit for the data, capturing the trend, seasonality, and general patterns effectively. However, it might benefit from further adjustments to better capture rapid changes and provide more accurate forecasts for periods with steeper trends. Examining the residuals and considering alternative model specifications could potentially improve the model's performance.

4.2. Machine Learning Model Evaluation

4.2.1. Random Forest Regression

Random Forest Regression is a highly regarded machine learning algorithm in the domain of supervised learning, particularly well-suited for predictive modeling with continuous target variables. It exhibits exceptional predictive accuracy, demonstrating its prowess in handling extensive datasets and unraveling intricate patterns within the data. A standout feature of Random Forest Regression is its remarkable capacity for rapid model training, a crucial asset in scenarios where time efficiency is paramount. Random Forest Regression can handle both stationary and non-stationary data, making it a flexible choice for regression problems involving time series or other types of data. Its notable strengths extend beyond classification tasks, making it a formidable tool for regression analysis.

In below residual plot (Figure 16), the residuals (the differences between actual and predicted values) against the predicted values are plotted. It helps to check for patterns or non-linearity in the residuals. As per the figure 4.11, random scatter of residuals are around zero, with no visible patterns.

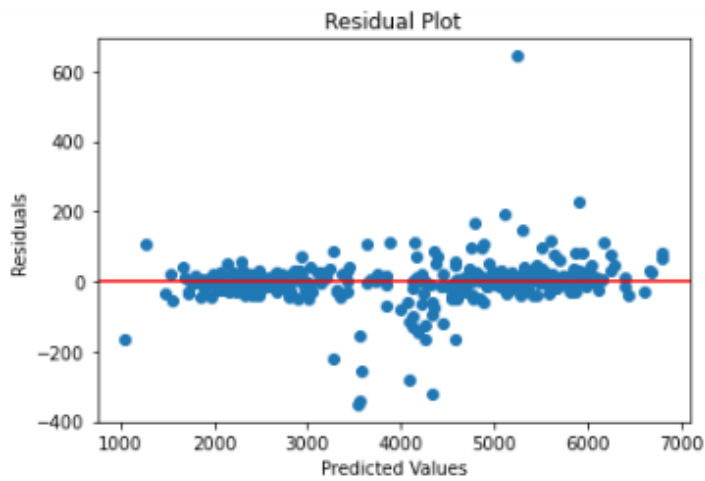


Figure 16: Residual Plot

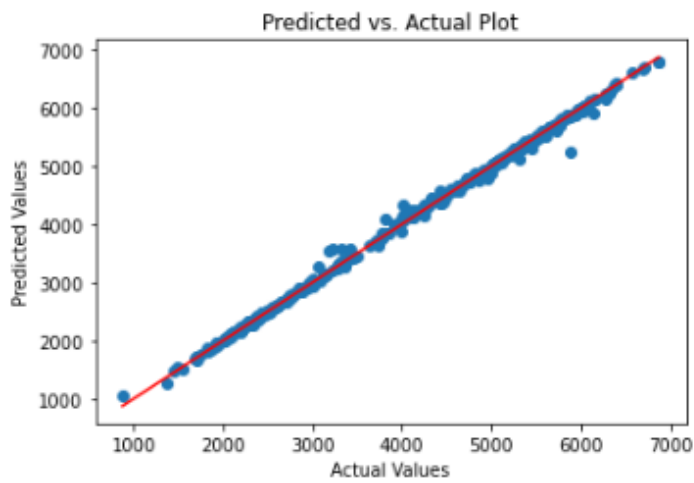


Figure 17: Predicted vs. Actual Plot

In the Figure 17, Predicted vs. Actual Plot, a comparison is made between the actual target values and the predicted values. When the predictions closely match the actual values, as demonstrated in Figure 17, the data points typically cluster around a diagonal line that runs from the bottom-left to the top-right of the plot. This alignment along the diagonal signifies the accuracy of the predictions. Conversely, a substantial deviation from this diagonal line suggests a lack of model adequacy, indicating that the random forest regression model which is build, effectively forecasts sales trends.

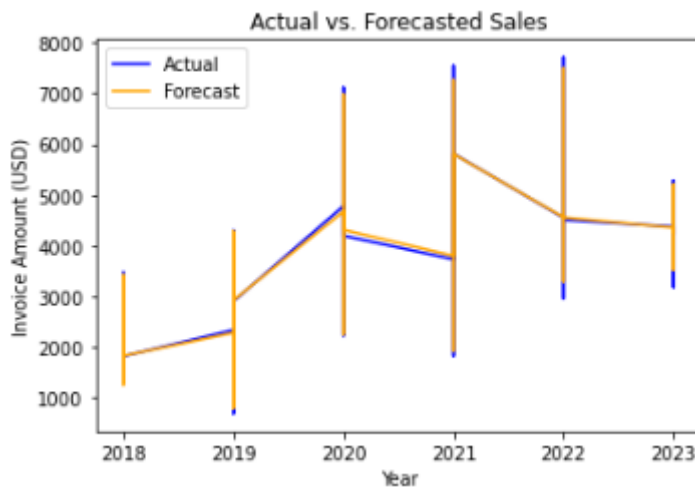


Figure 18: Actual Vs. Forecasted Sales

Model Parameters are as follows.

Mean Squared Error (MSE): 4332.451798638762
Mean Absolute Error (MAE): 32.99519401041704
R-squared (R2): 0.9979348621244586
Explained Variance Score: 0.9979357997218169

The above metrics suggest that the Random Forest model is performing exceptionally well in capturing the patterns in the data and making accurate predictions (Figure 18). The extremely high R-squared and Explained Variance Score values indicate that the model is able to account for almost all of the variability in the target variable. The low MSE and MAE values further confirm the model's precision in predicting the target variable.

4.2.2. Extreme Gradient Boosting (XGBoost) Regression

XGBoost, or Extreme Gradient Boosting, is an efficient algorithm primarily used for predicting continuous numeric values, making it particularly suitable for tasks like forecasting sales trends. This algorithm is designed with speed and efficiency in mind, employing techniques such as parallelization, column block optimization, and sparsity-aware splitting to expedite the training process. A notable feature of XGBoost is its built-in support for handling missing values, negating the necessity for imputation, as it can intelligently determine the best path when a value is missing

during tree construction. Moreover, XGBoost incorporates pruning, a technique that eliminates splits that do not contribute significantly to reducing loss, thereby ensuring that trees are not overly complex, preventing overfitting. It is worth noting that XGBoost is versatile and does not strictly require stationary data, as its applicability depends more on the specific problem and the quality of data preprocessing.

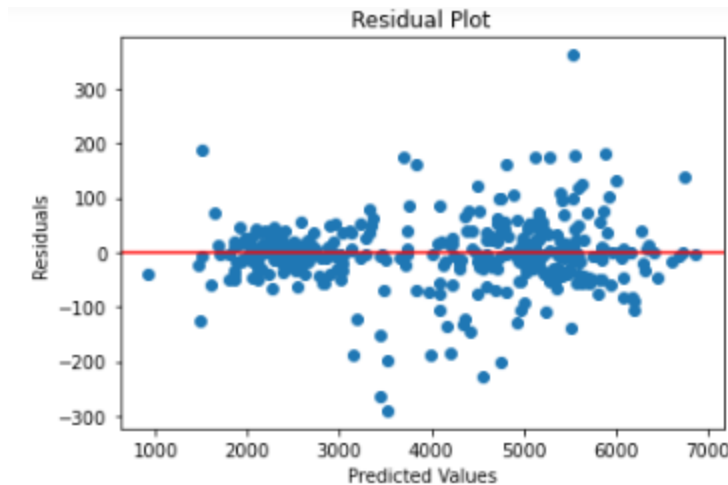


Figure 19: Residual plot

The residual plot (Figure 19) displays a random scatter of data points distributed around the horizontal line positioned at $y = 0$. This pattern reveals the absence of discernible trends or systematic errors in the model's predictions. However, the presence of patterns, such as a funnel shape in the residuals, can be indicative of potential issues like heteroscedasticity or non-linearity within the data. However, in the model under consideration, these patterns are notably absent, suggesting the model's adequacy and accuracy in capturing the underlying data patterns.

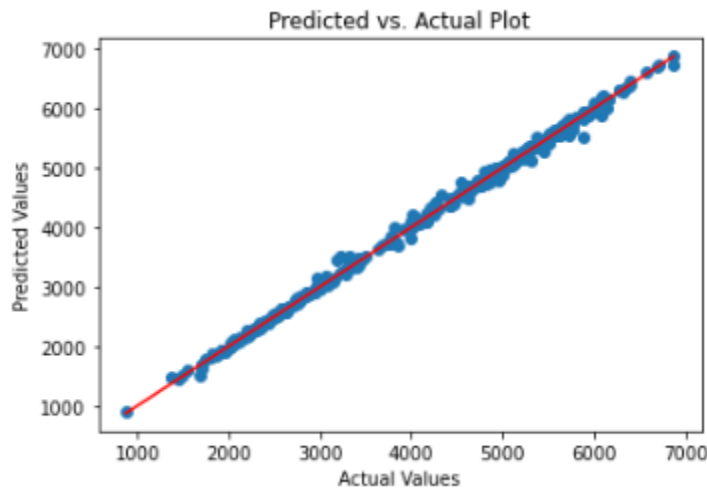


Figure 20: Predicted Vs Actual Plot

Model Parameters are as follows.

```
Mean Squared Error (MSE): 3996.047417565643
Mean Absolute Error (MAE): 40.84002135594686
R-squared (R2): 0.9980952150749681
Explained Variance Score: 0.9980989604214201
```

The model (Figure 20) appears to be performing exceptionally well, with exceptionally high R-squared and Explained Variance Score values (0.99809). This indicates that the model is able to explain nearly all of the variance in the target variable, suggesting a very strong fit between the model's predictions and the actual data. The MSE and MAE values (3996.04 and 40.84, respectively) are relatively low, further supporting the model's accuracy in predicting the target variable.

4.2.3. Long Short-Term Memory Model (LSTM)

The primary objective of Long Short-Term Memory (LSTM) is to effectively capture and learn long-term dependencies in sequential data. However, LSTM is not restricted to stationary data and LSTM's ability to handle non-stationary data is particularly valuable because many real-world time series datasets are non-stationary, including the sale's dataset using in this study. Non-stationary data may exhibit trends, seasonality, and other variations that change over time, and LSTM can still be used to model and forecast such data effectively. In practice, LSTM can be efficiently

implemented in Python using popular deep learning libraries like Keras or TensorFlow, making it a valuable tool for various applications, from time series forecasting to natural language processing.

Below are the part of logs from the training process of a neural network using the TensorFlow library. These logs show the progress of the training process over multiple epochs.

```
Epoch 1/100
24/24 [=====] - 4s 9ms/step - loss: 0.0522
Epoch 2/100
24/24 [=====] - 0s 9ms/step - loss: 0.0099
Epoch 3/100
24/24 [=====] - 0s 11ms/step - loss: 0.0082
Epoch 4/100
24/24 [=====] - 0s 9ms/step - loss: 0.0076
Epoch 5/100
24/24 [=====] - 0s 10ms/step - loss: 0.0074
Epoch 6/100
24/24 [=====] - 0s 9ms/step - loss: 0.0073
```

"Epoch 1/100": This indicates that the training is in its first epoch out of 100 total epochs. An epoch is one complete pass through the entire training dataset. "24/24": The first number (24 in this case) represents the number of batches processed in the current epoch. The second number (also 24) represents the total number of batches in the entire training dataset. During training, data is often divided into batches for more efficient processing. "- loss: 0.0522": This shows the loss value (measure of how well the model is performing) at the end of the current epoch. The loss is a metric that the model aims to minimize during training. In this case, the loss at the end of the first epoch is 0.0522.

Below is the summary of the model. "lstm_2 (LSTM) (None, 50) 10400" describes the first layer of the model. It is an LSTM (Long Short-Term Memory) layer with 50 units or neurons. The output shape is specified as (None, 50), where "None" indicates that the batch size can vary, and 50 represents the number of neurons in the layer. The number of parameters in this layer is 10,400. "dense_2 (Dense) (None, 1) 51" describes the second layer, which is a Dense layer. It has 1 unit (commonly used for regression tasks). The output shape is (None, 1), indicating a single output value. The number of parameters in this layer is 51.

Model: "sequential_2"

Layer (type)	Output Shape	Param #
lstm_2 (LSTM)	(None, 50)	10400
dense_2 (Dense)	(None, 1)	51

=====
Total params: 10451 (40.82 KB)
Trainable params: 10451 (40.82 KB)
Non-trainable params: 0 (0.00 Byte)

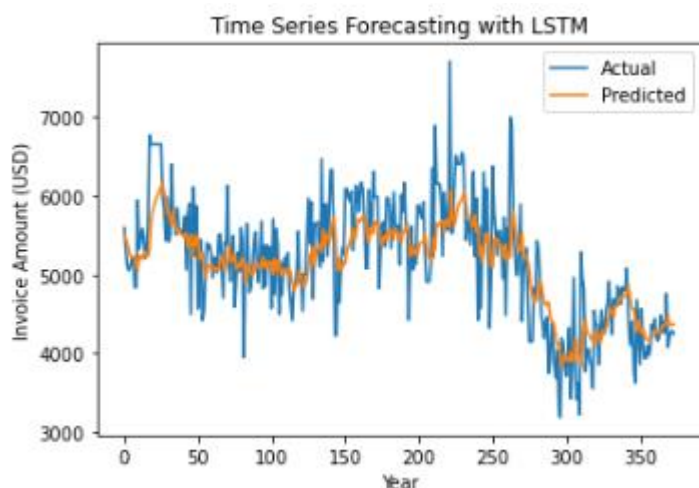


Figure 21: Time series forecasting with LSTM

Model parameters are as follows.

Mean Squared Error (MSE): 245550.7380189521
Mean Absolute Error (MAE): 380.02452607996315
R-squared (R2): 0.5864604581296673
Explained Variance Score: 0.5934946308177084

The R-squared and Explained Variance Score values (0.58646 and 0.59349, respectively) suggest that the LMST model (Figure 21) is able to explain around 58-59% of the variance in the target variable. This indicates a moderate fit between the model's predictions and the actual data, but there's room for improvement. The MSE and MAE values (245550.74 and 380.02, respectively) provide information about the average magnitude of the errors in the predictions. In conclusion,

the LSTM model's performance is not as strong as the previously discussed Random Forest and XGBoost models.

4.3. Model Comparison

Table 6 and table 7 shows the summary of the tested Time Series Models MAE, MSE, RMSE, and MAPE values for test data and Machine Learning Model MAE, MSE, RMSE, and MAPE values for test data respectively

Table 6: Time Series Model comparison for Test Data

Model	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Square Error (RMSE)	Mean Absolute Percentage Error (MAPE)
ARIMA	642.77	657349.30	810.77	11.92%
SARIMA	616.78	585796.21	765.37	12.52%
SARIMAX	408.47	411742.93	641.67	9.11%

Table 7: Machine Learning Model comparison for Test Data

Model	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	R-squared (R2)	Explained Variance Score
Random Forest Regression	32.99	4332.45	0.99	99.81%
Extreme Gradient Boost	40.84	3996.04	0.99	99.81%
LSTM	380.02	245550.73	0.58	59.35%

The evaluation results for sales forecasting indicate below insights:

- ARIMA provides forecasts with a relatively high Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Square Error (RMSE). These metrics suggest that ARIMA may not provide the most accurate forecasts. Furthermore, the Mean Absolute Percentage Error (MAPE) is 11.92%, indicating a substantial degree of forecast inaccuracy.

- SARIMA performs better than ARIMA in terms of MAE, MSE, and RMSE, but it still has relatively high errors. The MAPE, although lower, is at 12.52%, indicating some degree of forecast inaccuracy.
- SARIMAX exhibits improved performance compared to both ARIMA and SARIMA with the lowest MAE, MSE, RMSE, and MAPE. It seems to capture the seasonal and exogenous factors more effectively.
- Random Forest Regression delivers very low errors across the board. The low MAE, MSE, and RMSE values, along with high R-squared and Explained Variance Score of 0.9979, suggest that it provides highly accurate forecasts. It appears to be a strong candidate for sales forecasting.
- XGBoost also demonstrates very low errors, similar to Random Forest. It exhibits excellent accuracy with low MAE, MSE, and high R-squared and Explained Variance Score of 0.9981, making it a strong contender for sales forecasting.
- LSTM, while providing forecasts, has relatively high MAE, MSE, and RMSE values. The R-squared and Explained Variance Score (0.5935) is considerably lower than Random Forest and XGBoost, indicating that it may struggle to capture complex patterns and dependencies in sales data, making it less accurate for this specific application.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1. Overview of the Study

The central aim of this comprehensive study is to delve into the intricacies of time series sales data analysis, thereby uncovering a robust model that can effectively capture and interpret the dynamics of sales data. Building on the insights provided in chapter 2, it has become increasingly evident that the landscape of research pertaining to sales data analysis has been relatively sparse in recent years. Furthermore, the existing body of research primarily focuses on the broader perspective of sales data analysis, with limited attention paid to the nuances specific to the apparel industry.

Given the significance of the apparel industry in Sri Lanka, a prominent player in the global export market, the absence of substantial evidence elucidating the behavioral patterns within this sector is a critical gap that demands urgent attention. With the overarching goal of filling this void, the study not only contributes to the existing literature but also endeavors to paint a comprehensive picture of the trend analysis specifically tailored to the intricate dynamics of apparel sales. Through a meticulous examination of the data, this research seeks to offer valuable insights that can pave the way for informed decision-making and strategic planning within the realm of Sri Lanka's apparel industry.

5.2. Conclusion

Numerous studies and research efforts have explored machine learning algorithms and time series techniques for trend analysis and sales forecasting in various contexts and this study delves into the critical importance of precise trend analysis and sales forecasting within the B2B apparel industry, elucidating various forecasting approaches and their respective performances. Significantly, in the realm of trend analysis and sales forecasting, SARIMAX, Random Forest Regression, and XGBoost emerge as formidable contenders, consistently providing remarkably accurate forecasts, as demonstrated in Table 4.3 and Table 4.4. As per the generated results, out of

time series models tested in the study, SARIMAX, exhibits a moderate level of accuracy, while ARIMA and LSTM fall behind when compared to their counterparts in terms of forecast precision. While LSTM networks are generally robust models, excelling in handling sequential and time-series data owing to their capacity to retain information across extensive sequences, the results of this study indicate subpar performance. There could be several reasons for this outcome. Primarily, the inadequate quantity of data might impede the LSTM model's ability to effectively learn intricate patterns. Insufficient data is known to compromise the model's performance, often leading to less-than-optimal results.

Notably, when it comes to machine learning models, Random Forest Regression and XGBoost distinctly outshine ARIMA-based models, consistently delivering highly precise and reliable forecasts, a fact underscored by their consistently low error metrics. This highlights the robustness of Random Forest Regression and XGBoost as invaluable tools for generating dependable forecasts in the context of the B2B apparel industry. However, the best model is typically the one with the lower values for the Mean Squared Error (MSE) and Mean Absolute Error (MAE) and higher values for the R-squared (R²) and Explained Variance Score. Based on the test results, XGBoost has a lower MSE and higher R² and Explained Variance Score, indicating that it provides more accurate predictions and better explains the variance in the data. Therefore, based on these metrics, the second model XGBoost is the better-performing model in this study.

In some prior studies also researchers found that machine learning models are best in predicting sales data, which supports the findings of this study. For instance, Mortensen et al. [15] assessed B2B sales using various classification algorithms, finding that Random Forest achieved the highest accuracy. Similarly, Rezazadeh [18] achieved an accuracy rate of 87% with the XGBoost. Moreover, Haselbeck et al. [3] integrated classical and machine learning models for horticultural plant sales, finding that XGBoost was the most effective model, which suggests its potential relevance in the apparel industry.

Furthermore, the study contributes to the body of knowledge by filling the gaps in the literature related to B2B sales forecasting within the apparel industry, stressing the significance of data-driven decision-making over intuition. The customer profiling aspect also can assist companies in identifying and focusing on their most profitable customers, thus maximizing financial performance. While recognizing certain limitations, such as the exclusion of global market demand analysis and unstructured data, this research provides a robust foundation for evidence-based

decision-making, aiming to deepen our understanding of the Sri Lankan apparel industry's dynamics and equipping stakeholders to navigate challenges and exploit future opportunities effectively.

5.3. Limitations and Further improvements of the study

5.3.1. Limitations

- The study does not cover global market demand for specific products or countries due to unavailable data, focusing solely on analyzing and forecasting sales trends within the provided dataset. Hence, the findings should be interpreted within the context of the Sri Lankan apparel industry.
- Unstructured data sources like emails and invoices are excluded from analysis due to project scope and resource constraints.
- The study exclusively focuses on apparel sales activities and does not take into consideration other sales operations, including machinery sales costs and maintenance costs, which are integral components of the company's monthly sales.
- Findings may not be generalized to all apparel companies since the study uses data from a single company.

5.3.2. Further improvements of the study

- Despite the insights provided by previous studies, their applicability to the apparel industry remains uncertain due to the unique characteristics of this sector, such as short product life cycles and seasonal trends. Therefore, further research specifically tailored to the apparel industry is necessary to determine the most effective sales forecasting approaches in this context.
- Moreover, it is essential to acknowledge that the quality of the results may be influenced by the specific dataset and the features used in the modeling process. Therefore, a prudent approach would be to experiment with different models and employ various feature

engineering techniques to fine-tune and optimize sales forecasting further. This iterative process can help tailor the model selection to the unique characteristics of your sales data and business objectives and computational complexity.

APPENDICES

Appendices A

Letter from Client: *Note that Research topic had been changed later on*



Course Co-Ordinator - Master of Business Analytics
University of Colombo School of Computing,
No 35, Reid Avenue,
Colombo 07.
23.03.2023

Dear Sir,

Granting permission to access the Data

On Behalf of Expo Industrial Engineering (PVT) LTD., this is to confirm that we will grant permission and full access to our database to Wasala Mudiyansele Vindya Dinushani Wasala who is conducting a study on **"Future of B2B sales: A Trend Analysis on Sri Lankan Apparel Industry "** as a partial fulfillment of her Postgraduate studies.

Thank you.

A handwritten signature in blue ink, appearing to be "M N M Nasif". The signature is written over a horizontal dotted line.

M N M Nasif

Group Manager - IT

Expo Industrial Group

Appendices B

The part of the dataset used in this research study are shown in the below screenshots,

InvoiceID	InvoiceDate	BuyerName	UnitPrice	Qty	InvoiceAmount (USD)	ExchangeRate	InvoiceValue (LKR)	ProductID	Description
180100001	2018-01-01 10:09:57	ATG HAND CARE (PVT) LTD	7.2	1240	8928	152.85	1,364,644.80	ATGSQ500000-HTS	SQUARE HEAT STICKER (03 COLOUR)
180100002	2018-01-01 10:17:35	ATG HAND CARE (PVT) LTD	7.2	2545	18324	152.85	2,800,823.40	ATGSQ500000-HTS	SQUARE HEAT STICKER (03 COLOUR)
180100003	2018-01-01 10:28:35	ATG HAND CARE (PVT) LTD	7.2	2771	19951.2	152.85	3,049,540.92	ATGSQ500000-HTS	SQUARE HEAT STICKER (03 COLOUR)
180100004	2018-01-01 10:34:23	ATG HAND CARE (PVT) LTD	7.2	2865	20628	152.85	3,152,989.80	ATGSQ500000-HTS	SQUARE HEAT STICKER (03 COLOUR)
180100005	2018-01-01 14:07:31	BRANDIX APPAREL LTD	7.86	2841	22330.26	152.85	3,413,180.24	BRA46065-HTS	MADE IN SRI LANKA HEAT SEAL (46065/A)
180100006	2018-01-02 14:08:46	TIMEX GARMENTS (PVT) LTD	7.41	1878	13915.98	152.85	2,127,057.54	TIMEBLOC46300-HTS	BLOCH SIZE, CARE & COO LABEL
180100007	2018-01-02 15:12:40	TIMEX GARMENTS (PVT) LTD	7.41	2930	21711.3	152.85	3,318,572.21	TIMEBLOC46300-HTS	BLOCH SIZE, CARE & COO LABEL
180100008	2018-01-02 15:14:38	TIMEX GARMENTS (PVT) LTD	7.41	1491	11048.31	152.85	1,688,734.18	TIMEBLOC46300-HTS	BLOCH SIZE, CARE & COO LABEL
180100009	2018-01-02 15:18:08	BRANDIX APPAREL LTD	7.18	1231	8838.58	152.85	1,350,976.95	BRA-PYB-166-54077-HTS	DOG PYB-166/C11 HEAT SEAL (54077)
180100010	2018-01-02 18:45:13	BRANDIX APPAREL LTD	6.69	1767	11821.23	152.85	1,806,875.01	BRA-AM-004-54077-HTS	DOG YU18-AM-004 HEAT SEAL (54077)
180100011	2018-01-02 18:48:04	BRANDIX APPAREL LTD	7.58	2493	18896.94	152.85	2,888,397.28	BRA-AE-002-54077-HTS	DOG YU18-AE-002 HEAT SEAL (54077)
180100012	2018-01-02 18:59:00	BRANDIX APPAREL LTD	7.18	2723	19551.14	152.85	2,988,391.75	BRA-PYB-166-54077-HTS	DOG PYB-166/C11 HEAT SEAL (54077)
180100013	2018-01-02 19:03:28	BRANDIX APPAREL LTD	6.75	2705	18258.75	152.85	2,790,849.94	BRA-YU-AJ-002-54077-HTS	DOG YU18-AJ-002 HEAT SEAL (54077)
180100014	2018-01-02 19:09:19	BRANDIX APPAREL LTD	6.69	1296	8670.24	152.85	1,325,246.18	BRA-AM-004-54077-HTS	DOG YU18-AM-004 HEAT SEAL (54077)
180100015	2018-01-02 19:12:52	BRANDIX APPAREL LTD	7.18	1880	13498.4	152.85	2,063,230.44	BRA-PYB-166-54077-HTS	DOG PYB-166/C11 HEAT SEAL (54077)
180100016	2018-01-02 19:28:29	OMEGA LINE LTD	8.57	2693	23079.01	152.85	3,527,626.68	OMEGA52651-HTS	CASHMERE HEAT TRANSFER LABEL (52651)
180100017	2018-01-03 10:16:14	INDUSTRIAL CLOTHINGS LTD	8.85	1700	15045	153.47	2,308,956.15	IND42815-HTS	10000003557/ WURTH LOGO WITH STRAP

190103040	2019-07-24 14:29:39	JK GARMENTS (PVT) LTD	8.12	2038	16548.56	180.07	2,979,899.20	COUSRI20483-HTS	PUMA LOGO HEAT SEAL (20820/A) WHITE
190103041	2019-07-24 14:43:50	ANSELL TEXTILES LANKA (PVT)	10.5	2929	30754.5	180.07	5,537,962.82	ANSESPO0199-HTS	ACTIVE ARMOUR HEAT SEAL (59-408/IND02564) (ESP)
190103042	2019-07-24 15:52:48	HI FASHION HOLDING (PVT)	8.32	1959	16298.88	180.07	2,934,939.32	HIFESPO0590-HTS	LEVIS HEAT SEAL (NON CANADA) - ESPO0590
190103043	2019-07-24 15:54:47	CRYSTAL MARTIN CENTRAL (PVT) LTD	9.95	1168	11621.6	180.07	2,092,701.51	CRYEPU00091-LCR	NABAIJI HEAT SEAL - 8328710 (EPU00091) - 106073
190103044	2019-07-24 16:51:31	BRANDIX APPAREL LTD	10.08	1121	11299.68	180.07	2,034,733.38	BRASRI20517-LCR	PINK HEAT SEAL (YF19-UL-033) - 20517
190103045	2019-07-24 16:58:38	BRANDIX APPAREL LTD	9.68	2901	28081.68	180.07	5,056,668.12	BRA-PYB168-54217-HTS	PINK WORD HEAT SEAL - PYB-168/C12 (54217)
190103046	2019-07-24 20:37:58	NORLANKA MANUFACTURING LTD	10.93	1482	16198.26	180.07	2,916,820.68	NORSRI20591-LCR	ATG EU 2018 HEAT STICKER (120000)
190103047	2019-07-24 21:04:18	BRANDIX APPAREL LTD	9.73	2795	27195.35	180.07	4,897,066.67	BRA55682-HTS	WEDZE & LOGO HEAT SEAL-DKT-012A PINK (55682/B)
190103048	2019-07-24 21:12:09	BRANDIX APPAREL LTD	10.14	2896	29365.44	180.07	5,287,834.78	BRADCA51175-HTS	KALEN ISO TYPE HEAT SEAL - BLACK (51175/A)
190103049	2019-07-24 21:15:49	BRANDIX APPAREL LTD	8.91	2028	18069.48	180.07	3,253,771.26	BRADCA53367-HTS	KEY HEAT TRANSFER LABEL - DKT N05B (GREY) 53367
190103050	2019-07-25 08:52:33	JK GARMENTS (PVT) LTD	8.12	2784	22606.08	179.68	4,061,860.45	JKGSRI20820-LCR	PUMA LOGO HEAT SEAL (20820/A) WHITE
190103051	2019-07-25 08:54:24	BRANDIX APPAREL LTD	11.16	2616	29194.56	179.68	5,245,678.54	BRASRI20360-HTS	FLAG TOMMY LOGO SIZE HEAT SEAL - 051219 (20360)
190103052	2019-07-25 10:43:02	BRANDIX APPAREL LTD	11.16	1204	13436.64	179.68	2,414,295.48	BRASRI20360-HTS	FLAG TOMMY LOGO SIZE HEAT SEAL - 051219 (20360)
190103053	2019-07-25 11:22:03	CRYSTAL MARTIN CENTRAL (PVT) LTD	10.93	2924	31959.32	179.68	5,742,450.62	CRY51175-HTS	ATG EU 2018 HEAT STICKER (120000)
190103054	2019-07-25 11:46:38	CRYSTAL MARTIN CENTRAL (PVT) LTD	9.95	1761	17521.95	179.68	3,148,343.98	CRYEPU00091-LCR	NABAIJI HEAT SEAL - 8328710 (EPU00091) - 106073

200100851	2020-03-12 11:30:14	STAR GARMENTS LTD	12.85	1390	17861.5	185.18	3,307,592.57	STASRI21970-HTS	COSTCO KIRKLAND SIGNATURE (7585258) HEAT SEAL - MADE IN SRI LANKA
200100722	2020-03-02 13:59:20	WORKWEAR LANKA (PVT) LTD	15.93	2224	35428.32	186.12	6,593,918.92	WORSRI21982-HTS	MANIPULA SPECIALIST TL - 12 HEAT SEAL (21982)
200100900	2020-03-17 16:56:40	SHORE TO SHORE BRAND PA	15.04	1345	20228.8	185.18	3,745,969.18	SHOSRI21328-LCR	CALVIN KLEIN (CALG0397) HEAT SEAL - EMB - 21328 (INDIA ONLY)
200100904	2020-03-17 16:56:40	SHORE TO SHORE BRAND PA	15.04	2900	43616	185.18	8,076,810.88	SHOSRI21328-LCR	CALVIN KLEIN (CALG0397) HEAT SEAL - EMB - 21328 (INDIA ONLY)
200100918	2020-03-22 16:56:40	ANSELL LANKA (PVT) LTD	15.48	2796	43282.08	185.18	8,014,975.57	ANSE55939-HTS	HYFLEX HEAT SEAL (11-541/NA) 55940
200100996	2020-04-06 16:56:40	ANSELL LANKA (PVT) LTD	15.48	1860	28792.8	186.6	5,372,736.48	ANSE55939-HTS	HYFLEX HEAT SEAL (11-541/NA) 55940
200101138	2020-05-14 17:47:42	BRATEX (PVT) LTD	15.38	2607	40095.66	186.6	7,481,850.16	BRTR601-12-BK	R601/12 BLACK HEAT SEAL
210100001	2021-01-01 11:04:52	STEWARTS MANUFACTURING LTD	17.2	1358	23357.6	195.86	4,574,819.54	STESRI23545-LCR	4505 LOGO HEAT SEAL - 23545/A
210100002	2021-01-01 11:55:16	STEWARTS MANUFACTURING LTD	17.2	1817	31252.4	195.86	6,121,095.06	STESRI23545-LCR	4505 LOGO HEAT SEAL - 23545/A
210100003	2021-01-01 12:14:29	STEWARTS MANUFACTURING LTD	18.79	1148	21570.92	195.86	4,224,880.39	STESRI20591-LCR	ATG EU 2018 HEAT STICKER (120000)
210100004	2021-01-01 12:53:02	BRANDIX APPAREL LTD	19.67	1863	36645.21	195.86	7,177,330.83	BRASRI25887-LCR	LOVE PINK HEAT SEAL (TF21-VS-003/C5,6,7,8) - BACK - 25887
210100005	2021-01-01 13:36:31	BRANDIX APPAREL LTD	19.67	1560	30685.2	195.86	6,010,003.27	BRASRI25887-LCR	LOVE PINK HEAT SEAL (TF21-VS-003/C5,6,7,8) - BACK - 25887
210100006	2021-01-01 14:09:17	BRANDIX APPAREL LTD	19.67	2881	56669.27	195.86	11,099,243.22	BRASRI25887-LCR	LOVE PINK HEAT SEAL (TF21-VS-003/C5,6,7,8) - BACK - 25887

220109768	2022-12-30 14:08:56	BRANDIX APPAREL LTD	23.5	1279	30056.5	370	11120905	BRASRI33057-HTS	VICTORIA'S SECRET (VV-15487/C1 HEAT SEAL - EMB - 15487)
220109769	2022-12-30 16:54:27	BRANDIX APPAREL LTD	21.07	1513	31878.91	370	11795196.7	BRASRI33054-HTS	VICTORIA'S SECRET (VV-15341/C1 HEAT SEAL - EMB - 15341)
220109770	2022-12-30 17:48:44	BRANDIX APPAREL LTD	23.02	1464	33701.28	370	12469473.6	BRASRI33055-HTS	VICTORIA'S SECRET (VV-15354/C1 HEAT SEAL - EMB - 15354)
220109771	2022-12-30 17:50:29	BRANDIX APPAREL INDIA (PVT) LTD	23	2100	48300	370	17871000	BRASRI27534-LCR	PINK (PD-131/C151) HEAT SEAL - EMB - (27534)
220109772	2022-12-30 21:33:16	BRANDIX APPAREL INDIA (PVT) LTD	23	2123	48829	370	18066730	BRASRI27534-LCR	PINK (PD-131/C151) HEAT SEAL - EMB - (27534)
220109773	2022-12-31 13:23:53	BRANDIX APPAREL LTD	21.07	1572	33122.04	370	12255154.8	BRASRI33054-HTS	VICTORIA'S SECRET (VV-15341/C1 HEAT SEAL - EMB - 15341)
220109774	2022-12-31 15:22:40	BRANDIX APPAREL INDIA (PVT) LTD	23	2349	54027	370	19989990	BRASRI27534-LCR	PINK (PD-131/C151) HEAT SEAL - EMB - (27534)
230100001	2023-01-01 10:20:34	INQUEB GLOBAL (PVT) LTD	27.71	696	19286.16	331	6383718.96	INQSRI31479-LCR	NABAIJI BRANDING SILICON HEAT SEAL - 31479/B
230100002	2023-01-01 10:26:14	INQUEB GLOBAL (PVT) LTD	25.37	1085	27526.45	331	9111254.95	INQSRI31648-HTS	NABAIJI BRANDING HEAT SEAL -31648/B
230100003	2023-01-01 10:33:14	INQUEB GLOBAL (PVT) LTD	25.37	552	14004.24	331	4635403.44	INQSRI31648-HTS	NABAIJI BRANDING HEAT SEAL -31648/B
230100004	2023-01-01 10:37:50	JAY JAY MILLS LANKA (PVT) LTD	26.62	1558	41473.96	331	13727880.76	JAYSRI34277-LCR	CK LOGO HEAT SEAL - J20J222400 - FRONT - 34277
230100005	2023-01-01 10:41:26	BRANDIX APPAREL LTD	28.76	798	22950.48	331	7596608.88	BRASRI36666-HTS	PINK HEAT SEAL - YF23-014/C4 - 36666/A
230100006	2023-01-01 10:45:32	JAY JAY MILLS LANKA (PVT) LTD	26.23	1313	34439.99	331	11399636.69	JAYSRI27420-HTS	CALVIN KLEIN JEANS & CK LOGO HEAT SEAL - J30J32019/
230100007	2023-01-01 10:49:32	JAY JAY MILLS LANKA (PVT) LTD	26.23	1839	48236.97	331	15966437.07	JAYSRI27420-HTS	CALVIN KLEIN JEANS & CK LOGO HEAT SEAL - J30J32019/
230100008	2023-01-01 10:54:53	JAY JAY MILLS LANKA (PVT) LTD	27.22	799	21748.78	331	7198846.18	JAYJAY26923-HTS	CALVIN KLEIN JEANS & CK LOGO HEAT SEAL - J30J32080/
230100009	2023-01-01 10:58:44	JAY JAY MILLS LANKA (PVT) LTD	28.78	718	20660.04	331	6839797.24	JAYSRI33505-LCR	CK LOGO HEAT SEAL - J20J222343 - FRONT - 33505/A
230100010	2023-01-01 11:02:45	ATG HAND CARE (PVT) LTD	25.55	584	14921.2	331	4938917.2	ATGEU120000-HTS	ATG EU 2018 HEAT STICKER (120000)
230100011	2023-01-01 11:06:52	ATG HAND CARE (PVT) LTD	25.55	1814	46347.7	331	15341088.7	ATGEU120000-HTS	ATG EU 2018 HEAT STICKER (120000)

Appendices C

Identifying Missing values

```
import pandas as pd

file_path = "G:/python_work/sales_data.xlsx"
df = pd.read_excel(file_path)

missing_values = df.isna().sum()

total_rows = df.shape[0]

missing_percentage = (missing_values / total_rows) * 100

missing_data_info = pd.DataFrame({
    "Missing Values": missing_values,
    "Missing Percentage": missing_percentage
})

missing_data_info = missing_data_info.sort_values(by="Missing Percentage", ascending=False)

print(missing_data_info)
```

Description, ProductID and ExchangeRate Imputation

```
import pandas as pd

file_path = "G:/python_work/sales_data.xlsx"
df = pd.read_excel(file_path)

description_mode = df['Description'].mode()[0]
df['Description'].fillna(description_mode, inplace=True)

description_mode = df['ProductID'].mode()[0]
df['ProductID'].fillna(description_mode, inplace=True)

df['ExchangeRate'].ffill(inplace=True)

output_file_path = "G:/python_work/sales_data_imputed.xlsx"
df.to_excel(output_file_path, index=False)

print("Mode imputation for 'Description', 'ProductID' and forward fill imputation for 'ExchangeRate' completed.")
print("The DataFrame with imputed values is saved to", output_file_path)
```

Identifying complete missing data

```
import pandas as pd

file_path = 'G:/python_work/sales_data_imputed.xlsx'
df = pd.read_excel(file_path)

df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate']).dt.date
df.set_index('InvoiceDate', inplace=True)

full_date_range = pd.date_range(start=df.index.min(), end=df.index.max())
missing_dates = full_date_range.difference(df.index)

missing_records_count = len(missing_dates)

missing_records = []
for date in missing_dates:
    missing_record = pd.Series({'BuyerName': '',
                                'UnitPrice': pd.NA, 'Qty': pd.NA,
                                'InvoiceAmount (USD)': pd.NA, 'ExchangeRate': pd.NA,
                                'ProductID': '', 'Description': ''}, name=date)
    missing_records.append(missing_record)

missing_df = pd.DataFrame(missing_records)
df = df.append(missing_df)

df.to_excel('G:/python_work/sales_data_with_missing_records.xlsx', index=True)

print(f"{missing_records_count} missing records were added.")
```

Imputing newly added missing data

```
import pandas as pd

file_path = 'G:/python_work/sales_data_with_missing_records.xlsx'
df = pd.read_excel(file_path)

df['ExchangeRate'].fillna(method='ffill', inplace=True)

df['BuyerName'].fillna(df['BuyerName'].mode().iloc[0], inplace=True)
df['ProductID'].fillna(df['ProductID'].mode().iloc[0], inplace=True)
df['Description'].fillna(df['Description'].mode().iloc[0], inplace=True)

df.to_excel('G:/python_work/sales_dataset_new.xlsx', index=False)

print("Missing values have been imputed.")
```

```

import pandas as pd

file_path = 'G:/python_work/sales_dataset_new.xlsx'
df = pd.read_excel(file_path)

df['Date'] = df['InvoiceDate'].dt.date

date_mean = df.groupby('Date')[['UnitPrice', 'Qty']].mean().shift(1)

df['UnitPrice'].fillna(df['Date'].map(date_mean['UnitPrice']), inplace=True)
df['Qty'].fillna(df['Date'].map(date_mean['Qty']), inplace=True)

df['Qty'] = df['Qty'].round(0)

df['InvoiceAmount (USD)'] = (df['UnitPrice'] * df['Qty']).round(2)

df['ProductID'].fillna(df['ProductID'].mode().iloc[0], inplace=True)
df['Description'].fillna(df['Description'].mode().iloc[0], inplace=True)

df.to_excel('G:/python_work/sales_dataset_new_imputed.xlsx', index=False)

print("Missing values have been imputed, ExchangeRate has been calculated, and 'Qty' is now rounded without decimals.")

```

Outlier Detection and removing

```

import pandas as pd

file_path = 'G:/python_work/sales_dataset_new.xlsx'
df = pd.read_excel(file_path)

target_variable = 'InvoiceAmount (USD)'

Q1 = df[target_variable].quantile(0.25)
Q3 = df[target_variable].quantile(0.75)

IQR = Q3 - Q1

threshold = 1.15

lower_bound = Q1 - threshold * IQR
upper_bound = Q3 + threshold * IQR

outliers_df = df[(df[target_variable] < lower_bound) | (df[target_variable] > upper_bound)]
cleaned_df = df[(df[target_variable] >= lower_bound) & (df[target_variable] <= upper_bound)]

cleaned_file_path = 'G:/python_work/sales_dataset_without_outliers.xlsx'
cleaned_df.to_excel(cleaned_file_path, index=False)

outliers_below = len(outliers_df[outliers_df[target_variable] < lower_bound]) / len(df) * 100
outliers_above = len(outliers_df[outliers_df[target_variable] > upper_bound]) / len(df) * 100
total_outliers = outliers_below + outliers_above

print(f"Outliers Below Lower Bound: {outliers_below:.2f}%")
print(f"Outliers Above Upper Bound: {outliers_above:.2f}%")
print(f"Total Outliers: {total_outliers:.2f}%")

print(f"dataset_without_outliers saved to {cleaned_file_path}")

```

Label Encoding

```
import pandas as pd
from sklearn.preprocessing import LabelEncoder

file_path = 'sales_dataset_without_outliers.xlsx'

df = pd.read_excel(file_path)

columns_to_encode = ['BuyerName', 'Description', 'ProductID']

print("First five records of columns before encoding:")
print(df[columns_to_encode].head())

label_encoder = LabelEncoder()

for column in columns_to_encode:
    encoded_column_name = f'Encoded_{column}'
    df[encoded_column_name] = label_encoder.fit_transform(df[column])

print("\nFirst five records of columns after encoding:")
print(df[[f'Encoded_{column}' for column in columns_to_encode]].head())

output_file_path = 'encoded_sales_data.xlsx'
df.to_excel(output_file_path, index=False)

print("\nEncoding completed and saved to", output_file_path)
```

InvoiceDate with Date and Time Transformed to Date , Month and Year

```
import pandas as pd

file_path = 'new_sales_dataset.xlsx'
df = pd.read_excel(file_path)

df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'])

df['Date'] = df['InvoiceDate'].dt.date
df['Month'] = df['InvoiceDate'].dt.month
df['Year'] = df['InvoiceDate'].dt.year

new_file_path = 'new_sales_dataset_modified.xlsx'
df.to_excel(new_file_path, index=False)

print(f"Dataset with 'Date' , Month' and 'Year' columns saved to {new_file_path}")
```

Feature Selection

```
import pandas as pd
from sklearn.feature_selection import SelectKBest, f_regression

file_path = 'new_sales_dataset_modified.xlsx'
df = pd.read_excel(file_path)

target_variable = 'InvoiceAmount (USD)'

features = df[['UnitPrice', 'Qty', 'ExchangeRate', 'Encoded_BuyerName', 'Encoded_Description',
               'Encoded_ProductID', 'InvoiceValue (LKR)', 'Month', 'Year']]

k_best = SelectKBest(score_func=f_regression, k=5)

X_new = k_best.fit_transform(features, df[target_variable])

selected_feature_indices = k_best.get_support(indices=True)

selected_features = features.columns[selected_feature_indices]

print("Selected Features:", selected_features)
```

Removing Unwanted fields, consider means values of multiple transactions on same day

```
import pandas as pd

file_path = 'G:/python_work/new_sale_dataset2.xlsx'
df = pd.read_excel(file_path)

columns_to_drop = [
    'InvoiceID',
    'BuyerName',
    'ProductID',
    'Description',
    'Encoded_BuyerName',
    'Encoded_Description',
    'Encoded_ProductID',
    'InvoiceDate',
    'Month'
]

df.drop(columns=columns_to_drop, inplace=True)

output_file_path = 'G:/python_work/new_sale_dataset2.xlsx'
df.to_excel(output_file_path, index=False)

print("Dataset with specified columns dropped saved to:", output_file_path)
```



```

import pandas as pd

file_path = 'G:/python_work/new_sale_dataset2.xlsx'
df = pd.read_excel(file_path)

grouped = df.groupby('InvoiceDate').agg({
    'UnitPrice': lambda x: round(x.mean(), 2),
    'Qty': 'mean',
    'ExchangeRate': 'mean'
})

grouped['Qty'] = grouped['Qty'].round().astype(int)

grouped['InvoiceAmount (USD)'] = (grouped['UnitPrice'] * grouped['Qty']).round(2)

grouped['InvoiceValue (LKR)'] = (grouped['ExchangeRate'] * grouped['InvoiceAmount (USD)']).round(2)

grouped.reset_index(inplace=True)

output_file_path = 'G:/python_work/new_sale_dataset3.xlsx'
grouped.to_excel(output_file_path, index=False)

print("New dataset with mean values, rounded Qty, and calculated columns saved to:", output_file_path)

```

Defining heat maps

```

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

file_path = 'new_sales_dataset_modified.xlsx'
df = pd.read_excel(file_path)

selected_columns = [
    'UnitPrice', 'Qty', 'InvoiceAmount (USD)', 'ExchangeRate', 'InvoiceValue (LKR)',
    'ExchangeRate', 'Year'
]

selected_df = df[selected_columns]

correlation_matrix = selected_df.corr()

plt.figure(figsize=(8,6))
sns.heatmap(correlation_matrix, annot=True, cmap='viridis', linewidths=0.5)
plt.title('Correlation Matrix Heatmap')
plt.show()

```

Rolling Mean and Standard Deviation

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm
from statsmodels.tsa.stattools import adfuller

file_path = 'final_sale_dataset.xlsx'
df = pd.read_excel(file_path)

df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'])

df.set_index('InvoiceDate', inplace=True)

column_name = 'InvoiceAmount (USD)'

def check_stationarity(data):
    rolling_mean = data.rolling(window=7).mean()
    rolling_std = data.rolling(window=7).std()

    plt.figure(figsize=(12, 6))
    plt.plot(data, label='Original Data')
    plt.plot(rolling_mean, label='Rolling Mean')
    plt.plot(rolling_std, label='Rolling Std')
    plt.legend()
    plt.title('Original Data vs. Rolling Mean & Std')
    plt.show()

    result = adfuller(data)
    print('ADF Statistic:', result[0])
    print('p-value:', result[1])
    print('Critical Values:')
    for key, value in result[4].items():
        print(f'{key}: {value}')

check_stationarity(df[column_name])
```

First Order Differencing

```
import pandas as pd
import matplotlib.pyplot as plt

file_path = 'final_sale_dataset.xlsx'
df = pd.read_excel(file_path)

plt.figure(figsize=(12, 6))
plt.subplot(2, 1, 1)
plt.plot(df['InvoiceDate'], df['InvoiceAmount (USD)'], label='Original Data')
plt.xlabel('Invoice Date')
plt.ylabel('Invoice Amount (USD)')
plt.title('Original Invoice Amount (USD)')
plt.grid(True)
plt.legend()

df['InvoiceAmount (USD) Diff'] = df['InvoiceAmount (USD)'].diff()

plt.subplot(2, 1, 2)
plt.plot(df['InvoiceDate'], df['InvoiceAmount (USD) Diff'], label='First Order Differencing')
plt.xlabel('Invoice Date')
plt.ylabel('Differenced Invoice Amount (USD)')
plt.title('First Order Differencing of Invoice Amount (USD)')
plt.grid(True)
plt.legend()

plt.tight_layout()
plt.show()
```

Autocorrelation and Partial Autocorrelation

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm
from statsmodels.tsa.stattools import adfuller

file_path = 'new_sales_dataset_modified.xlsx'
df = pd.read_excel(file_path)

df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'])

df.set_index('InvoiceDate', inplace=True)

column_name = 'InvoiceAmount (USD)'

def check_stationarity(data):
    rolling_mean = data.rolling(window=7).mean()
    rolling_std = data.rolling(window=7).std()

    plt.figure(figsize=(12, 6))
    plt.plot(data, label='Original Data')
    plt.plot(rolling_mean, label='Rolling Mean')
    plt.plot(rolling_std, label='Rolling Std')
    plt.legend()
    plt.title('Original Data vs. Rolling Mean & Std')
    plt.show()

    result = adfuller(data)
    print('ADF Statistic:', result[0])
    print('p-value:', result[1])
    print('Critical Values:')
    for key, value in result[4].items():
        print(f'{key}: {value}')

check_stationarity(df[column_name])
```

ARIMA Model Evaluation

```
import pandas as pd
import numpy as np
from pmdarima.arima import auto_arima
import matplotlib.pyplot as plt
from sklearn.metrics import mean_absolute_error, mean_squared_error

file_path = 'final_sale_dataset.xlsx'
df = pd.read_excel(file_path)

df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'])

df.set_index('InvoiceDate', inplace=True)

daily_sales_data = df['InvoiceAmount (USD)']

daily_sales_data_diff = daily_sales_data.diff(1).dropna()

train_size = int(len(daily_sales_data_diff) * 0.8)
train_data_diff, test_data_diff = daily_sales_data_diff[:train_size], daily_sales_data_diff[train_size:]

n_forecast_test = len(test_data_diff)
n_forecast_year = 730
```

```

model = auto_arima(train_data_diff,
                  start_p=1, start_q=1,
                  test='ocsb',
                  max_p=5, max_q=5, max_d=2,
                  m=12,
                  d=1,
                  seasonal=True,
                  stationary=False,
                  start_P=0,
                  D=None,
                  trace=True,
                  error_action='ignore',
                  suppress_warnings=True,
                  stepwise=True)

print(model.summary())

forecast_test, conf_int_test = model.predict(n_periods=n_forecast_test, return_conf_int=True)
forecast_index_test = pd.date_range(start=df.index[train_size], periods=n_forecast_test, freq='D')
forecast_series_test = pd.Series(forecast_test, index=forecast_index_test)

forecast_year, conf_int_year = model.predict(n_periods=n_forecast_year, return_conf_int=True)
forecast_index_year = pd.date_range(start=forecast_index_test[-1] + pd.Timedelta(days=1), periods=n_forecast_year, freq='D')
forecast_series_year = pd.Series(forecast_year, index=forecast_index_year)

actual_values_test = test_data_diff
mae_test = mean_absolute_error(actual_values_test, forecast_test)
mse_test = mean_squared_error(actual_values_test, forecast_test)
rmse_test = np.sqrt(mse_test)
mape_test = np.mean(np.abs((actual_values_test - forecast_test) / actual_values_test)) * 100

print("Test Set Metrics:")
print(f"Mean Absolute Error (MAE): {mae_test:.2f}")
print(f"Mean Squared Error (MSE): {mse_test:.2f}")
print(f"Root Mean Square Error (RMSE): {rmse_test:.2f}")
print(f"Mean Absolute Percentage Error (MAPE): {mape_test:.2f}%")

plt.figure(figsize=(12, 6))
plt.plot(daily_sales_data_diff, label='Original Data (Differenced)', color='blue')
plt.plot(train_data_diff, label='Training Data (Differenced)', color='purple')
plt.plot(test_data_diff, label='Test Data (Differenced)', color='orange')
plt.plot(forecast_series_test, label='Forecasted Test Data', color='green')
plt.plot(forecast_series_year, label='Forecasted Next Year', color='red')
plt.legend()
plt.title('ARIMA Forecasting for Daily Sales Data (with Seasonality and Differencing)')
plt.show()

```

SARIMA model Evaluation

```

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm
import pmdarima as pm
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from sklearn.metrics import mean_absolute_error, mean_squared_error

file_path = "final_sale_dataset.xlsx"
data = pd.read_excel(file_path)

data['InvoiceDate'] = pd.to_datetime(data['InvoiceDate'])

data.set_index('InvoiceDate', inplace=True)
daily_sales_data = data['InvoiceAmount (USD)']

train_size = int(len(daily_sales_data) * 0.8)
train_data, test_data = daily_sales_data[:train_size], daily_sales_data[train_size:]

```

```

seasonal_period = 12
decomposition = seasonal_decompose(train_data, model='additive', period=seasonal_period)
trend = decomposition.trend
seasonal = decomposition.seasonal
residual = decomposition.resid

plt.figure(figsize=(12, 8))
plt.subplot(411)
plt.plot(train_data, label='Original Data (Train)')
plt.plot(test_data, label='Original Data (Test)')
plt.legend(loc='upper left')
plt.title('Original Time Series Data')

plt.subplot(412)
plt.plot(trend, label='Trend')
plt.legend(loc='upper left')
plt.title('Trend Component')

plt.subplot(413)
plt.plot(seasonal, label='Seasonal')
plt.legend(loc='upper left')
plt.title('Seasonal Component')

plt.subplot(414)
plt.plot(residual, label='Residuals')
plt.legend(loc='upper left')
plt.title('Residual Component')

plt.tight_layout()
plt.show()

auto_sarima = pm.auto_arima(train_data, seasonal=True, m=seasonal_period,
                           stepwise=True, suppress_warnings=True,
                           error_action="ignore", max_order=None,
                           trace=True)

```

```

p, d, q = auto_sarima.order
P, D, Q, S = auto_sarima.seasonal_order

sarima_model = sm.tsa.SARIMAX(train_data, order=(p, d, q), seasonal_order=(P, D, Q, S))
sarima_results = sarima_model.fit()

forecast_steps = len(test_data)
forecast = sarima_results.get_forecast(steps=forecast_steps)

forecasted_values = forecast.predicted_mean
confidence_intervals = forecast.conf_int()

future_forecast_steps = len(daily_sales_data) - len(train_data)
future_forecast = sarima_results.get_forecast(steps=future_forecast_steps)

future_forecasted_values = future_forecast.predicted_mean
future_confidence_intervals = future_forecast.conf_int()

future_forecast_index = pd.date_range(start=test_data.index[-1], periods=future_forecast_steps, freq='D')

plt.figure(figsize=(12, 6))
plt.plot(daily_sales_data.index, daily_sales_data, label='Original Data', color='blue')
plt.plot(train_data.index, train_data, label='Train Data', color='purple')
plt.plot(test_data.index, test_data, label='Test Data', color='orange')
plt.plot(test_data.index, forecasted_values, label='Forecast (Test)', color='green')
plt.plot(future_forecast_index, future_forecasted_values, label='Future Forecast', color='red')
plt.legend()
plt.title('SARIMA Forecasting')
plt.show()

```

```

mae = mean_absolute_error(test_data, forecasted_values)
mse = mean_squared_error(test_data, forecasted_values)
rmse = np.sqrt(mse)

mape = np.mean(np.abs((test_data - forecasted_values) / test_data)) * 100

print(f"Mean Absolute Error (MAE) for Test Data: {mae:.2f}")
print(f"Mean Squared Error (MSE) for Test Data: {mse:.2f}")
print(f"Root Mean Squared Error (RMSE) for Test Data: {rmse:.2f}")
print(f"Mean Absolute Percentage Error (MAPE) for Test Data: {mape:.2f}%")

```

SARIMAX Model Evaluation

```

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm
import pmdarima as pm
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from sklearn.metrics import mean_absolute_error, mean_squared_error

file_path = "final_sale_dataset.xlsx"
data = pd.read_excel(file_path)

data['InvoiceDate'] = pd.to_datetime(data['InvoiceDate'])

data.set_index('InvoiceDate', inplace=True)
daily_sales_data = data['InvoiceAmount (USD)']

exog_vars = data[['UnitPrice', 'Qty', 'ExchangeRate']]

train_size = int(len(daily_sales_data) * 0.8)
train_data, test_data = daily_sales_data[:train_size], daily_sales_data[train_size:]
exog_train = exog_vars[:train_size]
exog_test = exog_vars[train_size:]

auto_sarima = pm.auto_arima(train_data, exogenous=exog_train, seasonal=True, m=12,
                             stepwise=True, suppress_warnings=True,
                             error_action="ignore", max_order=None,
                             trace=True)

p, d, q = auto_sarima.order
P, D, Q, S = auto_sarima.seasonal_order

sarima_model = sm.tsa.SARIMAX(train_data, exog=exog_train, order=(p, d, q), seasonal_order=(P, D, Q, S))
sarima_results = sarima_model.fit()

forecast_steps = len(test_data)
forecast = sarima_results.get_forecast(steps=forecast_steps, exog=exog_test)

```

```

forecasted_values = forecast.predicted_mean
confidence_intervals = forecast.conf_int()

future_forecast_steps = len(daily_sales_data) - len(train_data)
future_forecast = sarima_results.get_forecast(steps=future_forecast_steps, exog=exog_vars[train_size:])

future_forecasted_values = future_forecast.predicted_mean
future_confidence_intervals = future_forecast.conf_int()

future_forecast_index = pd.date_range(start=test_data.index[-1], periods=future_forecast_steps, freq='D')

plt.figure(figsize=(12, 6))
plt.plot(daily_sales_data.index, daily_sales_data, label='Original Data', color='blue')
plt.plot(train_data.index, train_data, label='Train Data', color='purple')
plt.plot(test_data.index, test_data, label='Test Data', color='orange')
plt.plot(test_data.index, forecasted_values, label='Forecast (Test)', color='green')
plt.plot(future_forecast_index, future_forecasted_values, label='Future Forecast', color='red')
plt.legend()
plt.title('SARIMAX Forecasting with Exogenous Variables')
plt.show()

mae = mean_absolute_error(test_data, forecasted_values)
mse = mean_squared_error(test_data, forecasted_values)
rmse = np.sqrt(mse)

mape = np.mean(np.abs((test_data - forecasted_values) / test_data)) * 100

print(f"Mean Absolute Error (MAE) for Test Data: {mae:.2f}")
print(f"Mean Squared Error (MSE) for Test Data: {mse:.2f}")
print(f"Root Mean Squared Error (RMSE) for Test Data: {rmse:.2f}")
print(f"Mean Absolute Percentage Error (MAPE) for Test Data: {mape:.2f}%")

```

Random Forest Regression

```

import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score, explained_variance_score
import matplotlib.pyplot as plt

file_path = 'final_sale_dataset.xlsx'
df = pd.read_excel(file_path)

df['Year'] = pd.to_datetime(df['InvoiceDate']).dt.year

X = df.drop(columns=['InvoiceAmount (USD)', 'InvoiceDate'])
y = df['InvoiceAmount (USD)']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

rf_regressor = RandomForestRegressor(random_state=42)
rf_regressor.fit(X_train, y_train)

y_pred = rf_regressor.predict(X_test)

```



```

mse = mean_squared_error(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
explained_var = explained_variance_score(y_test, y_pred)

print(f'Mean Squared Error (MSE): {mse}')
print(f'Mean Absolute Error (MAE): {mae}')
print(f'R-squared (R2): {r2}')
print(f'Explained Variance Score: {explained_var}')

residuals = y_test - y_pred
plt.scatter(y_pred, residuals)
plt.xlabel("Predicted Values")
plt.ylabel("Residuals")
plt.axhline(y=0, color='r')
plt.title("Residual Plot")
plt.show()

plt.scatter(y_test, y_pred)
plt.xlabel("Actual Values")
plt.ylabel("Predicted Values")
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color='r')
plt.title("Predicted vs. Actual Plot")
plt.show()

feature_names = X.columns
importances = rf_regressor.feature_importances_
indices = importances.argsort()[::-1]
top_k = min(10, len(importances))

```

```

plt.figure(figsize=(10, 6))
plt.title("Top {} Feature Importances".format(top_k))
plt.bar(range(top_k), importances[indices][:top_k], align="center")
plt.xticks(range(top_k), feature_names[indices][:top_k], rotation=90)
plt.xlabel("Feature")
plt.ylabel("Importance")
plt.tight_layout()
plt.show()

forecast = rf_regressor.predict(X)
forecast = pd.Series(forecast, name='Forecasted Invoice Amount (USD)')

plt.plot(df['Year'], df['InvoiceAmount (USD)'], label='Actual', color='blue')
plt.plot(df['Year'], forecast, label='Forecast', color='red')
plt.xlabel('Year')
plt.ylabel('Invoice Amount (USD)')
plt.title('Actual vs. Forecasted Sales')
plt.legend()
plt.show()

```


XGBoost Model Evaluation

```
import pandas as pd
from sklearn.model_selection import train_test_split
import xgboost as xgb
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error, explained_variance_score
import matplotlib.pyplot as plt

file_path = 'final_sale_dataset.xlsx'
df = pd.read_excel(file_path)

X = df.drop(columns=['InvoiceAmount (USD)', 'InvoiceDate'])
y = df['InvoiceAmount (USD)']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

xgb_regressor = xgb.XGBRegressor(objective="reg:squarederror", random_state=42)

xgb_regressor.fit(X_train, y_train)

y_pred = xgb_regressor.predict(X_test)

mse = mean_squared_error(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
explained_var = explained_variance_score(y_test, y_pred)

print(f'Mean Squared Error (MSE): {mse}')
print(f'Mean Absolute Error (MAE): {mae}')
print(f'R-squared (R2): {r2}')
print(f'Explained Variance Score: {explained_var}')
```

```
residuals = y_test - y_pred
plt.scatter(y_pred, residuals)
plt.xlabel("Predicted Values")
plt.ylabel("Residuals")
plt.axhline(y=0, color='r')
plt.title("Residual Plot")
plt.show()

plt.scatter(y_test, y_pred)
plt.xlabel("Actual Values")
plt.ylabel("Predicted Values")
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color='r')
plt.title("Predicted vs. Actual Plot")
plt.show()

xgb.plot_importance(xgb_regressor, importance_type='weight', max_num_features=10)
plt.show()
```

LSTM Model

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dense
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score, explained_variance_score
import matplotlib.pyplot as plt

file_path = 'final_sale_dataset.xlsx'
df = pd.read_excel(file_path)

target_column = 'InvoiceAmount (USD)'
data = df[target_column].values.reshape(-1, 1)

scaler = MinMaxScaler()
data = scaler.fit_transform(data)

train_size = int(len(data) * 0.8)
train_data = data[:train_size]
test_data = data[train_size:]

def create_sequences(data, look_back=1):
    X, y = [], []
    for i in range(len(data) - look_back):
        X.append(data[i:(i + look_back), 0])
        y.append(data[i + look_back, 0])
    return np.array(X), np.array(y)

look_back = 10
X_train, y_train = create_sequences(train_data, look_back)
X_test, y_test = create_sequences(test_data, look_back)

model = Sequential()
model.add(LSTM(50, input_shape=(look_back, 1)))
model.add(Dense(1))
model.compile(loss='mean_squared_error', optimizer='adam')

model.fit(X_train, y_train, epochs=100, batch_size=64)

y_pred = model.predict(X_test)

y_pred = scaler.inverse_transform(y_pred)
y_test = scaler.inverse_transform(y_test.reshape(-1, 1))

mse = mean_squared_error(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
explained_var = explained_variance_score(y_test, y_pred)

model.summary()

print(f'Mean Squared Error (MSE): {mse}')
print(f'Mean Absolute Error (MAE): {mae}')
print(f'R-squared (R2): {r2}')
print(f'Explained Variance Score: {explained_var}')

years = range(len(y_test))
```

```
plt.plot(years, y_test, label='Actual')
plt.plot(years, y_pred, label='Predicted')
plt.legend()
plt.xlabel('Year')
plt.ylabel('Invoice Amount (USD)')
plt.title('Time Series Forecasting with LSTM')
plt.show()
```

References

- Amrutkar, P. & Mahadik, S., 2022. Sales Prediction Using Machine Learning Techniques. *International Journal of Research Publication and Reviews*, 3(8), pp. 1887-1890.
- Anil, G. A. et al., 2023. Sales Forecasting Using Machine Learning Techniques. *International Research Journal of Modernization in Engineering Technology and Science*, 5(3).
- Anon., 2023. *Apparel Industry in Sri Lanka – Moving up the Supply Value Chain*. [Online] Available at: <https://blog.bizvibe.com/blog/textiles-and-garments/apparel-industry-in-sri-lanka> [Accessed 12 March 2023].
- Arif, I. A., Sany, I. S., Nahin, I. F. & Rabby, A. S., 2019. *Comparison Study: Product Demand Forecasting with Machine Learning for Shop*. India, s.n.
- Bohanec, M., Borštnar, M. K. & Robnik-Šikonja, M., 2017. Explaining machine learning models in sales predictions. *Expert Systems with Applications*, Volume 71, pp. 416-428.
- Bohanec, M., Robnik-Šikonja, M. & Kljaj, M., 2017. Organizational Learning Supported by Machine Learning Models Coupled with General Explanation Methods: A Case of B2B Sales Forecasting. *Organizacija*, Volume 50, pp. 217-233.
- Dias, S., 2022. *Sri Lanka's apparel workers face bleak future*. [Online] Available at: <https://www.sundaytimes.lk/221218/business-times/sri-lankas-apparel-workers-face-bleak-future-505595.html> [Accessed 12 March 2023].
- Ensafia, Y., Amin, S. H., Zhang, G. & Shah, B., 2022. Time-series forecasting of seasonal items sales using machine learning –A comparative analysis. *International Journal of Information Management Data Insights*, Volume 2.
- Giri, C., Jain, S., Zeng, X. & Pascal, B., 2019. A Detailed Review of Artificial Intelligence Applied in the Fashion and Apparel Industry. *Digital Object Identifier*, Volume 7, pp. 95376-95396.
- Haselbeck, F. et al., 2022. Machine Learning Outperforms Classical Forecasting on Horticultural Sales. *Machine Learning with Applications*, Volume 7.
- Linh, K., 2022. *Sri Lankan Apparel Exports Under Threat Amid Economic Crisis*. [Online] Available at: <https://www.businessoffashion.com/news/global-markets/sri-lankan-apparel-exports-under-threat-amid-economic-crisis/> [Accessed 12 March 2023].
- Lu, C.-J. & Kao, L.-J., 2016. A clustering-based sales forecasting scheme by using extreme learning machine and ensembling linkage methods with applications to computer server. *Engineering Applications of Artificial Intelligence*, Volume 55, pp. 231-238.
- Makridakis, S., Spiliotis, E. & Assimakopoulos, V., 2020. The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1), pp. 54-57.

- Makridakis, S., Spiliotis, E. & Assimakopoulos, V., 2018. The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4), pp. 802-808.
- Mirza, J. & Ensign, P., 2021. New direction for a Sri Lankan apparel venture: chasing a capitalist or cooperative dream. *Small Enterprise Research*, 28(1), pp. 83-94.
- Mortensen, S. et al., 2019. *Predicting and Defining B2B Sales Success with Machine Learning*. Charlottesville, VA, USA, IEEE, pp. 1-5.
- Pavlyshenko, B., 2019. *Machine-Learning Models for Sales Time Series Forecasting*. Lviv, Ukraine, s.n., pp. 21-25.
- Raizada, S. & Jatinderkumar, S., 2021. Comparative Analysis of Supervised Machine Learning Techniques for Sales Forecasting. *International Journal of Advanced Computer Science and Applications*, Volume 12, pp. 102-110.
- Rammandala, C., 2022. *A Close Examination of Sri Lankan Apparel Industry*. [Online] Available at: <https://www.onlineclothingstudy.com/2022/08/sri-lankan-apparel-industry.html> [Accessed 12 March 2023].
- Rezazadeh, A., 2020. A Generalized Flow for B2B Sales Predictive Modeling: An Azure Machine-Learning Approach. *Forecasting*, Volume 2, pp. 267-283.
- Samanthi, S. A. G., 2022. *Industry Capability Report, SriLankan Apparel Sector*, Sri Lanka: Export Development Board (EDB).
- Stephan, K., 2022. Commentary on the M5 forecasting competition. *International Journal of Forecasting*, 38(4), pp. 1562-1568.
- Varshney, N., 2022. *Meanwhile in Sri Lanka!*. [Online] Available at: <https://apparelresources.com/business-news/manufacturing/meanwhile-sri-lanka/> [Accessed 24 March 2023].
- Wickramasingha, S., 2023. *Crisis mode in the Sri Lankan apparel industry: a closer look at the implications for firms and workers*. [Online] Available at: <https://cbds.cbs.dk/crisis-mode-in-the-sri-lankan-apparel-industry-a-closer-look-at-the-implications-for-firms-and-workers/> [Accessed 12 March 2023].
- Wisesa, O., Adriansyah, A. & Khalaf, O. I., 2020. Prediction Analysis for Business To Business (B2B) Sales of Telecommunication Services using Machine Learning Techniques. *Majlesi Journal of Electrical Engineering*, 14(4), pp. 145-153.
- Wisesa, O., Adriansyah, A. & Khalaf, O. I., 2020. Prediction Analysis for Business To Business (B2B) Sales of Telecommunication Services using Machine Learning Techniques. *Majlesi Journal of Electrical Engineering*, 14(4), pp. 145-153.
- Yan, J. et al., 2015. *Sales pipeline win propensity prediction: a regression approach*. Ottawa, ON, Canada, IEEE, pp. 854-857.