Use of Machine Learning to Optimize Distribution for Small and Medium Enterprises

By

Joyce Mukuhi Kamau

094328

Submitted in Partial Fulfillment Of The Requirements For The Degree Of Master Of Science In Data Science And Analytics At Strathmore University

Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

JUNE 2025

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Student's Name: Joyce Mukuhi Kamau

Sign: ..... Date: 26/05/2025

Approval

The dissertation of **Joyce Mukuhi Kamau** was reviewed and approved for examination by the following:

Dr. John Olukuru
Institute of Mathematical Sciences,
Strathmore University

Dr. Godfrey Madigu,
Dean, Institute of Mathematical Sciences,
Strathmore University

Prof. Bernard Shibwabo,
Director of Graduate Studies,
Strathmore University

Dedication

I dedicate this study to God, my Father, and my provider for guiding me through. I also dedicate this study to my parents for their unwavering love support and patience. Without your constant encouragement and belief in me, I would never have completed this important task. Thank You.

Acknowledgement

I am deeply grateful to God Almighty for His strength and favor upon my life which has enabled me to get this far. Secondly, I thank Dr John Olukuru, my supervisor, who has guided me so well and his constant strive for excellence has encouraged me to pursue and study knowledge at all times. To my colleagues and classmates, thank you for encouraging me to be resilient and equipping me with the knowledge required to make this study a success, you have been very instrumental.

Abstract

Small and medium-sized enterprises (SMEs) play a crucial role in Kenya's economic development, yet they face persistent challenges in managing distribution and inventory efficiently. This study investigates the application of machine learning (ML) techniques to optimize distribution strategies and forecast sales quantity for SMEs, with a focus on the beauty and cosmetics sector. Using the CRISP-DM methodology, the research employs ensemble models—Decision Tree, Random Forest, and Gradient Boosting—to analyze historical sales and distribution data. The Random Forest model achieved the highest predictive accuracy, outperforming other models based on RMSE, MAPE, and R^2 metrics. Key features such as outlet type, and delivery route cluster emerged as significant predictors of sales quantity. The findings underscore the value of ensemble learning in capturing non-linear relationships and enhancing inventory planning and delivery scheduling for SMEs. This study contributes to theoretical frameworks such as Control Theory and Customer Value Theory by demonstrating how ML supports dynamic feedback and value-based customer segmentation. It also provides actionable recommendations for SMEs, including the digitization of inventory systems, adoption of interpretable ML models, and investment in real-time analytics. Limitations such as data sparsity and lack of external variables are discussed, and future research is encouraged to explore more advanced models, diverse datasets, and real-world deployment in SME environments.

Keywords: Distribution Effectiveness, Customer Satisfaction, Sales and Demand Forecasting

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
CRISP-DM	Cross-Industry Standard Process for Data Mining
MAD	Mean Absolute Deviation
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
RMSE	Root Mean Squared Error
SDG	Sustainable Development Goals
SME	Small Medium Enterprises

Chapter 1: Introduction to the Study

1.1 Background to the study

1.1.1 Growth of Small Medium Enterprises (SME's)

Small and Medium Enterprises (SMEs) are widely acknowledged as vital contributors to national economic development. In Kenya, they account for approximately 80% of employment, particularly in the informal sector, and are projected to contribute nearly half of the country's GDP in the coming years (Rotich, 2022). Their adaptability and entrepreneurial capacity make them key agents of economic inclusion, innovation, and regional development. However, as the global economy increasingly embraces digitization, SMEs face challenges in scaling operations and remaining competitive due to limited technological infrastructure and resources. The alignment of SMEs with sustainable development initiatives, such as the Sustainable Development Goals (SDGs), further emphasizes the need for innovation and strategic transformation within these enterprises (Scheyvens, Banks, & Hughes, 2016).

Global policy developments, such as the adoption of the Sustainable Development Goals (SDGs) in 2015, have further highlighted the role of businesses—including SMEs—in driving inclusive and sustainable economic growth. The SDGs encourage innovation, digital transformation, and responsible business practices, all of which align with the evolving expectations of customers, investors, and regulators (Scheyvens, Banks, & Hughes, 2016). SMEs, by virtue of their flexibility and community-rooted presence, are well-positioned to contribute to these goals. However, doing so requires them to adopt new technologies and integrate sustainability into their core operations. As customers increasingly demand ethical and efficient products and services, SMEs must evolve by embracing data-driven strategies, which are central to both competitiveness and sustainable development ((Verles & Vellacott, 2018); (Wang & Xu, 2018).

To thrive in this environment, SMEs must implement business models that are both adaptive and data-centric. The integration of business intelligence and analytics allows them to transform raw data into actionable insights that can inform better decision-making and optimize value creation ((Durst & Edvardsson, 2012); (Willetts, Atkins, & Stanier, 2020)). The growth of SME's can be attributed to proper business strategies and systems put in place by business owners and the government. SMEs must fulfill the global standards for quality, technology, sustainability, and price to be internationally competitive and take advantage of growing opportunities (Singh & Kumar, 2020).

The technique used by a company to transform inputs into outputs and results that are intended to accomplish its strategic goals and add value over the short, medium, and long terms is known as its business model. Choosing the right business model for a company is key to the management and customers (International Integrated Reporting Council, 2013). Business intelligence seeks to offer an organized, thorough perspective of all the data within an organization. Managers and executives can make modifications to processes in the future based on this data and draw conclusions about prior operations. Big data analytics is largely used by management and companies in order to gain a competitive advantage (Willetts, Atkins, & Stanier, 2020). This is through innovations apt actions and counter-actions in the marketplace.

In the age of Industry 4.0, organizations are being forced to make a paradigm change in order to employ new generation technologies, such as warehouse automations, to boost fulfillment speed and remain competitive (Moeuf, Pellerin, Lamouri, Tamayo-Giraldo, & Barbaray, 2017)Big data is recently seen as a new solution that supports policies and practices across different application contexts and domains. The large amount of data collected and retained over the years by various public and private organizations has given rise to a variety of new data analytics solutions (Kambatla, Kollias, Kumar, & Grama, 2014)The influence of data collected and stored by various public and private organizations over the years has given rise to many innovative data analysis techniques. These organizations are both profit-making and non-profit making. The current study explores how such innovations—particularly through the use of machine learning—can address operational inefficiencies and enhance performance in the SME distribution landscape.

1.1.2 SME's and Machine Learning

Small and medium-sized businesses, or SMEs, are vital to Kenya's economy because they support livelihoods and generate jobs. An estimated 15 million people are employed in this sector, which also accounts for almost 30% of the value added in the country (Tiriongo, Josea, & Mulindi, 2021). In an increasingly data-driven world, economic forecasting and the prediction of the growth of small and medium-sized businesses (SMEs) have become crucial instruments for directing company strategy, policy, and economic development (Al-Karkhi & Rządkowski, 2025). SMEs often gather a lot of data from different sources, but they are lacking the tools or know-how to properly examine it. With the use of machine learning (ML), SMEs can get a competitive edge by utilizing their data. As global markets evolve and data becomes a core asset, SMEs must leverage data-driven strategies to remain competitive and sustainable (Singh & Kumar, 2020 Machine

learning (ML), a subset of artificial intelligence, presents immense opportunities for SMEs seeking to enhance their competitiveness and operational effectiveness. ML enables the automation of complex decision-making processes by uncovering patterns and insights from large datasets. For SMEs, this technology can transform critical functions such as sales forecasting, customer segmentation, inventory management, and distribution planning (Amberkar, 2018). However, many SMEs lack the expertise or tools to effectively utilize ML, especially in sales and distribution.

In Kenya, many SMEs generate large volumes of data through customer transactions, deliveries, and product sales. However, most lack the technical capacity or tools to analyze and utilize this data effectively. This underutilization creates inefficiencies and missed opportunities for growth. ML can bridge this gap by providing scalable solutions that support real-time decision-making and accurate forecasting. For example, ML algorithms can identify customer behavior patterns, predict product demand, and optimize delivery routes. These capabilities allow SMEs to personalize services, minimize operational costs, and respond swiftly to market changes. Furthermore, clustering techniques can help segment customers based on purchasing behavior, enabling more targeted marketing and resource allocation (Omorinsola Bibire Seyi-Lande, 2024).

Despite these advantages, barriers such as cost, lack of skilled personnel, and limited awareness continue to hinder ML adoption among SMEs. As such, this study seeks to investigate how ML can be effectively applied to enhance SME distribution and delivery systems, and how these innovations can be made accessible and sustainable within the Kenyan context. In the modern business world, data-driven decision-making is essential for increasing competitiveness, streamlining procedures, and improving client experiences. Making decisions facilitates goal achievement, effective issue-solving, resource allocation, risk management, innovation, and disputes, fosters individual growth, increases self-assurance and trust, adjusts to shifting conditions, and supports long-term sustainability (Panghal, 2024).

Market segmentation is the practice of breaking down a large market into groups of consumers with reasonably similar product needs. A market segment might include people, groups, or organizations that are divided by shared traits including location, gender, lifestyle, and buying habits (Clough, 2011). As the economy grows, a company's ability to succeed now rests less on the strength of individual products and more on the total experience included within a product ecosystem. To provide a comprehensive experience, a product ecosystem combines a core product

with a wide range of auxiliary goods and services, making it impossible for other, more disparate offerings to compete. SMEs may better target their marketing by using machine learning to help with audience segmentation. ML models are able to divide the audience into various categories according to their online behavior, interests, and demographics. By looking at user preferences and behavior. More individualized marketing campaigns are made possible by this segmentation, which raises the possibility of engagement and conversion (Omorinsola Bibire Seyi- Lande, 2024). This study therefore investigates how ML can be harnessed to optimize distribution processes, thereby improving customer satisfaction, inventory management, and overall profitability.

1.1.3 Distribution Channels and Machine Learning

Distribution channels are essential components of business operations, serving as the link between producers and consumers. For SMEs, particularly in sectors involving physical product delivery, efficient distribution systems determine the timeliness, reliability, and overall customer experience. Traditional distribution approaches, however, are often challenged by inefficiencies such as delayed deliveries, poor route planning, and inconsistent inventory levels, especially in low-resource settings (Nelvin, 2015). Hence, the formation of distribution channels enables business enterprises to effectively deploy the process of bringing products close to the target consumers (Frazier, 2018).

Besides the development of distribution channels, the integration of data science in overall development strategies can play an important role in increasing efficiency. With the advent of digital transformation and data availability, machine learning offers a promising avenue for redefining distribution processes. ML models can be trained to forecast demand across geographic locations, recommend optimal delivery routes, and predict stock requirements based on historical and real-time data. This predictive capability can help SMEs minimize waste, reduce transportation costs, and enhance service reliability (Soltanizadeh, Abdul, Mottaghi, & Ismail, 2016). Furthermore, ML can support dynamic scheduling and routing strategies that adapt to traffic conditions, fuel efficiency, and customer delivery windows. Algorithms such as decision trees, support vector machines, and neural networks have been used successfully in other domains—such as e-commerce and logistics—to manage distribution more intelligently (Yang, 2019). Integrating these models into SME operations can bridge the performance gap between large firms and smaller businesses. The strategies used by a firm in integrating data science into the distribution channel can increase a firm’s level of competitiveness based on the competitive

advantages offered, such as differentiation, cost minimization, and focus strategy (Porter, 1985). Data science is therefore essential in enhancing the performance of distribution channels hence improving the competitive advantage as well as the operational efficiency of business enterprises. In the modern business environment, most manufacturing firms do not sell their products directly to the end consumers, the manufacturers partner with marketing intermediaries to distribute the manufactured products to the target market. Distribution channels tend to differ based on the target consumers, the existing industrial market, and the factors involved in the distribution channel. The choice of distribution channels as well as the members within the channel members usually have a significant impact on the strategies of a business enterprise (Frazier, 2018). Careful attention is therefore taken into account at various level of the decision-making process for the distribution channel. In this regard, three strategies exist, these include intensive distribution strategy, selective distribution strategy, and exclusive distribution strategy. Intensive distribution strategy is mainly associated with the daily usage of products that are of high consumption rate and low-priced (Gordini & Veglio, 2017).

Artificial neural networks can be applied in various tasks such as regression, classification, and segmentation among others. Machine learning on the other hand can be used to recognize and sort out patterns from the data provided and classify data into the possible appropriate patterns thus making predictions. In the last few decades, information-based prediction models using machine-learning techniques have become more popular (Nelvin, 2015). Several domains that have applied such models include movie rating, crime detection, and medical diagnosis among others. As a result of the significant financial cost of customer churn, businesses across the globe have undertaken the analysis of various factors such as customer service response time, call cost, and call quality among others using various machine learning tools like support vector machines, neural networks, decision trees, machines, probabilistic models such as Bayes, etc. (Yang, 2019). With technological innovation, electronic commerce has presented new opportunities for both consumers and business enterprises within distribution channels to share information easily, find and purchase products. This has increased the ease of product movement from one business enterprise to another and increased the risk of churn. Research studies develop a churn prediction model by testing the Support Vector Machine’s forecasting capability. The predictive performance is benchmarked to neural networks, logistic regression, and classic support vector machine. The review of products plays a key role in customers' decision-making process regarding electronic

commerce websites (Gordini & Veglio, 2017). Despite the potential, adoption remains low due to limited technical skills, infrastructure gaps, and cost concerns. This study addresses this challenge by designing a scalable ML-based framework for SMEs to enhance their distribution efficiency and effectiveness, particularly within the Kenyan context.

1.2 Problem Definition

Despite the availability of machine learning technologies, many SMEs in Kenya have not fully adopted data-driven methods for optimizing distribution. Issues such as fragmented sales data, inefficient delivery routes, and lack of forecasting tools hinder their ability to meet customer demand effectively. Poor distribution systems often result in customer dissatisfaction, lost sales, and increased operational costs. People in the emerging market economy are either buyers or sellers and regulators. These connections turned into business dealings. Everyone and everything served as a production input governed by the laws of production, the laws of demand and supply, including people and the environment (Hart, 2011).

It is important to understand a customer and their needs as well as address the needs in a manner that satisfies them, as a business owner. According to Stuart Hart, a great product with a poor distribution channel will fail just as spectacularly as a terrific new technology that lacks a valuable end-user application (Hart, 2011). This study seeks to bridge the gap by demonstrating how ML can improve sales and distribution efficiency within SMEs. This study addresses this gap by proposing a data-driven approach that integrates ML into SME distribution and delivery processes. It seeks to demonstrate how predictive modeling, when aligned with real operational data, can enhance sales forecasting, route optimization, and overall distribution performance. The findings aim to guide SMEs toward more strategic use of their data assets and inform the development of scalable, affordable ML-driven solutions in emerging markets.

1.3 Research Objectives

1.3.1 General Objective

The research aimed at finding the best distribution and delivery model, i.e., that will bring in and retain customers, and the role of big data in identifying this.

1.3.2 Specific Objective

This study aims;

i. To identify distribution and delivery models for customer needs and satisfaction

ii. To analyze and review the current systems and models used in sales and demand forecasting through machine learning

iii. To design and develop a machine learning model, a predictive model that forecasts the sales and demand of SMEs

iv. To validate the machine learning model for functionality created to solve the business problem.

1.4 Research Questions

i. How to identify distribution and delivery models for customer needs and satisfaction?

ii. What should be reviewed and analyzed in the current systems and models used in machine learning?

iii. How to design and develop a machine learning model, a predictive model that forecasts sales and demand of SME's?

iv. What metrics can be used to validate the machine learning model for functionality created to solve the business problem?

1.5 Scope and Limitation

The study focused on retail businesses that mainly reach and sell their products through distribution and deliveries. Just like Coca-Cola, the SME has adopted the business model that enables small, local firms to distribute their products to local retailers in densely populated urban areas as well as customers in remote locations (Nelson, Ishikawa, & Geaneotes, 2009). The study therefore focused on SMEs, and the results were generalized.

It can be difficult to keep the model scalable and effective as sales data increases or as more channels are introduced to the forecasting system. For instance, retraining the model or changing its design might be necessary when adding new product categories or geographical areas.

1.6 Significance of the Study

Data science, in general, is the use of both quantitative and qualitative techniques to address important topics and predict outcomes. With the huge and increasing amount of data available today, it has become clear that domain knowledge and analysis cannot be separated (Waller & Fawcett, 2013). This study aimed to build a machine learning model that will improve the current business model of SMEs to market and distribute their products by ensuring that all customer needs are met as well as reduction in expenses. Any business aims to maximize profits and minimize losses, thus depicting the importance of sales and demand forecasting. Through ML,

SMEs can easily predict future sales thus enabling the decision-makers make informed decisions about planning, production, supplying, and marketing activities.

Different businesses employ a variety of tactics to sustain their sales levels over a financial year. One of these strategies is to run sales campaigns is having promotions and discounts where a variety of products are sold to retailers at cheaper rates to the merchants in order to display a larger number of a good for a set amount of time. Although machine learning (ML) is clearly relevant in the manufacturing sector, small and medium-sized businesses (SMEs) are not using it to its full potential. Businesses with fewer than 500 workers use machine learning (ML) four times less frequently than businesses with more than 500 employees. This is because SMEs are unable to implement ML technologies since they lack the necessary ML expertise (Burggräf, Steinberg, Sauer, & Nettesheim, 2024).

In addition, the insights of this study could be beneficial to manufacturers of various products as they market their products through indirect distribution channels. Understanding the quantity of a certain product that is likely to be sold in that specific timeframe in that specific place at that specific price is an important component of crucial things to accomplish (Snyder & Shen, 2019). Scholars will also gain from this study since it highlights current trends in distribution systems and offers background information for future research in other fields or industries. The results will be used as a guide for similar future research.

Chapter 2: Literature Review

2.1 Introduction

The theoretical and empirical underpinnings of this study are presented in this chapter. This chapter seeks to understand various theories and how they are related to distribution effectiveness, predictive analysis, and big data. There is evidence to support the widespread belief that data are a driver of better decision-making and increased profitability in a business (Waller & Fawcett, 2013).

2.2 Theoretical Review

This study is anchored on two complimentary theories, Control Theory and Customer Value Theory, which provide foundational perspectives on operational efficiency and customer-centric value creation. A discussion of these theories is presented and their relation to the study variables is explained.

2.2.1 Control Theory

This study is anchored on control theory, which is a theory that is inseparable from data science, as it relies on sensor measurements (data) obtained from a system to achieve a given objective. Control theory deals with living data, as successful application modifies the dynamics of the system, thus changing the characteristics of the measurements. Control theory cannot be separated from data science; this is because it relies on sensor measurements (data) that is obtained from a system so as to achieve a given objective. Control theory actually deals with living data, as successful application modifies the dynamics of the system, hence changing the measurements' characteristics.

Control theory is about designing optimal actions for dynamical models, in continuous or discrete time. However, it is notoriously acknowledged that the numerical computation is the main barrier of putting these control theories to work in practice and many applications are unfortunately limited to the linear quadratic regulator (Bensoussan, 2018). For quite a longtime, dimensionality has had a significant impact on the numerical methods of control theory (Chiuso & Pillonetto, 2019). It is therefore natural that the new possibilities of ML should be taken into consideration in order to overcome the challenge of dimension. This provides an explanation on the reasons why in the last few years, several innovative ideas and exciting results have been witnessed from the perspective of merging control theory and Machine Learning, with the efforts from different areas

such as computational and mathematics, stochastic optimization, optimal control as well as computer science.

Researchers on optimal control and machine learning tend to explore the tools, techniques and problem formulations, from each other (Zunic, Donko, & Buza, 2020). The areas of research can be divided into two categories: Machine learning for control theory and control theory for machine learning. Generally, the latter refers to the use of control theory as a mathematical tool in formulating and solving practical as well as theoretical problems in machine learning, such as training neural network and optimal parameter tuning. The former on the other hand refers to use machine learning practice such as DNN and kernel method in solving numerically complex models in control theory which can become intractable by traditional methods (Bensoussan, 2018)

Analysis of fluctuations, unpredictability, and disturbances is essential component of SC and operations management systems. Understanding stability is heavily influenced by the system under consideration, as well as by the techniques and objectives of systems analysis (Louly, Dolgui, & Hnaien, 2008). The primary advantage of scheduling with optimum control applications is the potential to draw in a wide range of qualitative performance analysis techniques. Complex multi-stage, multi-period, and multi-commodity supply chains can be optimally designed, planned, and scheduled with the help of CT techniques. Nonetheless, the presence of centralized data for the entire network and pertinent metrics is a crucial prerequisite (Dolgui, Ivanov, Sethi, & Sokolov, 2018).

Control Theory emphasizes real-time feedback loops for regulating and adjusting performance across systems. While historically used in engineering, its application has expanded into business and operational analytics. Bensoussan (2018) and Chiuso and Pillonetto (2019) demonstrate how ML enhances control mechanisms through predictive analytics, allowing systems to self-correct based on data-driven forecasts. However, a limitation in current research is the assumption of consistent infrastructure and data quality—an unrealistic standard for most SMEs in Africa. For instance, the successful deployment of ML-based control systems in logistics firms is well-documented in Europe (Zunic, Donko, & Buza, 2020), but lacks contextual validation in informal or rural SME environments, especially in African contexts. This suggests a pressing need for adaptive models that can operate effectively despite data fragmentation, intermittent connectivity, and limited technical expertise—factors common among SMEs in developing countries.

2.2.2 Customer Value Theory

Customer Value Theory is a fundamental concept used in business strategy that holds that a product or service's ability to satisfy customers is what ultimately determines its marketability. This theory emphasizes that gaining a competitive edge and long-term business performance depend on recognizing and enhancing the perceived value for customers (Zeithaml, 1988). This study recognizes that customers are the most important organizational stakeholders, according to market-focused management, hence businesses should give great value to them. Customers are no longer willing to pay more than a good or service is worth in today's economy, where value reigns paramount. According to a study done by Capital FM in 2018, Britam, Oil Libya, DSTV, and Airtel are some of the companies with the best customer care in Kenya. Customers praised the personnel for being compassionate, friendly, and customer-focused, as well as the quality and affordable fuel, efficiency, and accessibility. The study involved over 37000 Kenyans by ensuring that at least 1000 people were surveyed for each industry (Capital FM, 2018).

A novel viewpoint is provided by customer value theory which focuses on what consumers value, why they value it, and how goods and services can be created to optimize the perceived value by consumers. Nicholson (2021) argues that customer value perceptions vary by socio-economic and cultural context. To understand a customer’s value, we need five choice models: current value, historical value, long-term value projection, credit, and loyalty (HuaHana, XiuLu, & C.H.Leung, 2012). The impetus behind a customer's decision-making process is their needs. The desire of the customer is what motivates them to buy a product and choose one over another. Businesses research the demands of their customers to improve their products, marketing methods, and customer service (Indeed Editorial Team, 2021). Perceived value can be defined as follows, as explained in (Zeithaml, 1988), value in four ways: value is (1) low price, (2) what I want in a product, (3) quality proportional to the price paid, and (4) what I get for what I give.

Businesses therefore need to understand that the direct experience a customer gets in the process of usage of the product is the perceived value. In today’s fast-paced market, ML can support real-time insights into customer behavior and optimize inventory and delivery timing—key elements in value delivery. Yet, as Nicholson (2021) argues, value perception among consumers in emerging markets often varies based on price sensitivity and accessibility, which challenges a one-size-fits-all approach. This calls for a contextual adaptation of value-driven frameworks that account for the technological readiness of SMEs and the preferences of local consumers.

In recent years, this framework has been extended to digital contexts where customer experiences are shaped by algorithm-driven interactions, including recommendation engines and dynamic pricing. ML tools enhance value delivery by allowing firms to understand and predict customer preferences. For instance, segmentation models based on purchase histories enable personalized promotions that increase satisfaction and loyalty. Real-time tracking and optimized delivery scheduling, driven by ML, also contribute to perceived reliability and responsiveness—core aspects of customer value. In Kenya, where price sensitivity and delivery timeliness significantly influence buyer decisions, customer value cannot be defined solely by global benchmarks. Instead, SMEs must adapt value propositions to local realities. ML adoption among Kenyan SMEs remains low, but where implemented, it has shown promise in enhancing customer service, particularly in urban logistics and e-commerce. Despite these insights, empirical validation of Customer Value Theory in ML-based SME settings remains limited. This study extends the theory by exploring how SMEs can leverage ML not just to understand but to anticipate and meet customer expectations in cost-effective and scalable ways.

2.3 Empirical Review

2.3.1 Distribution Chain and Channels

Machine learning (ML) has emerged as a powerful enabler of distribution efficiency, particularly in automating routing, minimizing transportation costs, and improving delivery speed. In large-scale logistics, Soltanizadeh et al. (2016) found that decision trees and neural network models significantly improved route planning, reduced downtime, and optimized load balancing. Yang (2019) corroborates these findings by showing that support vector machines can dynamically adjust to traffic conditions, resulting in a 15–25% improvement in last-mile delivery time. Despite these benefits, SMEs in emerging economies face unique limitations. They often lack digitized inventory systems, structured customer data, or integrated order management platforms—core prerequisites for ML deployment. For instance, Wanjiru and Muthoni (2022) report that although SMEs in Nairobi adopted GPS-based delivery mapping, its impact was minimal unless paired with mobile inventory tools and trained staff. These findings underscore the importance of combining technological tools with operational training and infrastructure support to unlock ML's full potential in distribution.

Moreover, research is limited on context-adaptive algorithms that can function in data-scarce environments—an issue particularly relevant to rural SMEs in Africa. Unlike multinationals, local SMEs do not generate large datasets, raising questions about how ML models can be customized for low-volume, noisy, or incomplete data inputs. Research indicates that large service business entities and their distribution chains account for more employees and higher revenues than manufacturing firms; the service distribution chains are therefore important to include for supply chain management (SCM) theory building (Handfield, Jeong, & Choi, 2017). In this regard, the distribution chain for both service and product deliveries can benefit from data science.

On an overall distribution chain level, literature indicates that data science can be applied in distribution channels for the purposes of development, operations, creation of value, discovery of value, and value capture (Zunic, Korjenic, Hodzic, & Donko, 2021)Empirical research also indicates that the utilization of analytic applications is best across distribution chains where the operations of analytics are cross-functional and an integral part of a business strategy (Zunic, Donko, & Buza, 2020). Source, move, make and sell as primary areas of application for big data have also been identified in previous research. With regards to source, big data may be used in the segmentation of suppliers, integrating with suppliers, evaluation of sourcing channel options, and supporting negotiations with suppliers. Make involves capacity constraints mitigation, reporting of granular performance, inventory optimization, facility location or location, and analytics of workforce. Regarding move, the application of data science entails scheduling and routing, the use of alternatives for transportation, optimization, and maintenance of vehicles. Finally, for the purpose of sales and marketing, data science enables micro segmentation of customers, the predicting and capturing of customer behavior and behavior, as well as optimization of pricing and assortment. These applications are defined as conceptual, but also involve some empirical grounding. Furthermore, Wang et al 2016., identifies the application of data analytics in the management of distribution chain as a strategic asset whose application is very important in several operational as well as strategic distribution processes.

A five-level analytic maturity framework for analytic applications in distribution management was established by the researchers. The first and the second levels are functional and process-based analytics that can be applied in distribution operations which cover the planning of demand, the process of procurement, production, inventory control, and logistics. This for instance involves the aligning of distribution and customer demand at stock keeping unit level, increasing the

visibility of supply-chain, management and mitigation of risks, management of real-time performance, as well as optimization processes. Further, agile, collaborative, and sustainable analytics at third, fourth, and fifth levels can be deployed in strategic distribution settings to ensure supply-chain network design, strategic sourcing as well as product design and development. This entails for instance evaluating and selecting suppliers, the physical configuration of the distribution or supply chain, and meeting the fluctuating demand requirements and the utilization of market opportunities by putting in place a rapid product design process (Wang, Diaz, Lopez-Garcia, & Carballedo, 2016).

Distribution is pivotal to the growth of any company because it is a form of marketing thereby increasing market visibility for the business and gaining access to new customers. Tight coordination between logistics, transportation, and inventory management is necessary for retail distribution to safeguard consumer satisfaction and trust, which in turn fosters customer loyalty. Additionally, the retailer and the final customer are the two clients in a retail distribution partnership. Ensuring that your retail partners receive goods that meet their criteria and are stocked with your most popular products is crucial (Jackson, 2024). Distribution channels are essential in the market, especially the Route-to-Market strategy because they define how a product moves from its producer to the final consumer (Infomineo, 2024). The Route-to-Market strategy is thought of as the traditional way of distribution because it is an indirect distribution channel. Making a product accessible for purchase involves disseminating it throughout the market. Delivery, packing, and transportation are all involved. A company's sales depend heavily on distribution. Making decisions on price, inventory, promotion, and other aspects involves many stakeholders (Sonntag, 2022). The key actors in the Route-to-Market strategy include Distributors, wholesalers, retailers, and agents. Distributors purchase products directly from the manufacturers and resell them to wholesalers, retailers, or direct consumers; they are often given the sole mandate to distribute products in particular areas (Infomineo, 2024).

Individual decision-makers are rewarded for maximizing local goals and developing risk-reduction strategies. This can result in having certain products understocked. In this study, the manufacturer has a strong incentive to make sure its distributors, have enough stock on hand. Increasing the profit margin from sales is one way to get the store to keep more inventory on hand. This results in higher end-user prices, which may cause demand to decline. On sales above a certain threshold, the wholesaler will occasionally give the store a different wholesale price. As a

result, the cost of the retailer having insufficient inventory rises (Yadav, Stapleton, & Wassenhove, 2013)).

2.3.2 Technological Advancements in Distribution

2.3.2.1 Predictive Analysis

Nearly all industries have experienced significant technological advancement in recent years. New technologies have allowed business systems to support new applications and business ideas (Safar, Sopko, Bednar, & Poklemba, 2018). The use of big data to develop or enhance demand-driven business models is growing in the current era of big data, where the overall value chain of traditional supply-oriented business models is increasingly fading out (Liu & Li, 2016). Accurate sales forecasting is critical for SMEs, especially when managing inventory, scheduling deliveries, and anticipating customer demand. Traditional models like ARIMA and exponential smoothing offer baseline forecasting capabilities but fall short in capturing the non-linear and multi-factorial nature of sales trends in volatile markets. Recent advancements have focused on models like ARIMA hybrids, gradient boosting methods (e.g., XGBoost (Fildes, Ma, & Kolassa, 2022)), and Long Short-Term Memory (LSTM) networks, which can learn temporal dependencies and accommodate external variables such as promotions, weather patterns, and economic indicators (Ahmed, Sultana, & Rahman, 2021).

Boone et al. (2019) emphasize that e-commerce and the digitalization of transactions have overwhelmed businesses with high-frequency data, pushing demand planners to evolve beyond traditional toolsets. Consequently, ML models are being adopted not just for their predictive accuracy but also for their ability to automate continuous learning from new data streams. Despite these advancements, practical deployment in SMEs remains limited. Most ML forecasting models require large, clean datasets and consistent labeling—conditions rarely met in small business environments. Amberkar (2018) notes that while clustering and classification models enhance targeting and segmentation, their downstream impact on profitability is seldom evaluated in empirical studies. This lack of operational validation creates uncertainty for SMEs considering adoption.

Furthermore, there is an emerging research focus on how data-scarce environments can leverage transfer learning and federated learning. These approaches allow SMEs to build forecasting models based on shared or synthetic datasets, potentially democratizing access to intelligent forecasting tools. However, empirical evaluations of these techniques within African SME contexts are still

emerging. In sum, while the potential for ML in sales and demand forecasting is immense, realizing its benefits in SMEs will depend on context-aware implementation strategies, supportive infrastructure, and ongoing performance evaluation metrics tailored to their resource constraints. Kumar and Garg (2018) highlight that integrating external variables—such as seasonality, competitor pricing, and regional events—into forecasting models improves accuracy by up to 30%. However, these sophisticated models demand high-quality data streams and computational infrastructure. For SMEs in Kenya, these prerequisites are often unavailable. Amberkar (2018) further notes that while clustering techniques can support promotional campaign design, few studies measure their sustained impact on actual sales and profitability.

What remains underexplored is how SMEs with limited historical data can still harness the benefits of forecasting. Possible solutions include the use of transfer learning, federated learning, or synthetic data generation to train robust models on minimal data—a frontier that this study explores through applied testing. Data mining is primarily concerned with the construction of models. A model is an algorithm or set of rules that ties a set of inputs (often fields in a corporate database) to a certain aim or outcome. Data mining analyzes the outcome of a certain problem or condition that may develop by using knowledge from previous data. The core concept of data mining for Sales and Distribution management is that historical data includes knowledge that will be beneficial in the future. It works because consumer actions documented in business data are not random but rather represent customers' varying wants, preferences, proclivities, and treatments (Russom, 2011). Data mining seeks to discover patterns in historical data that give insight into both the needs and preferences of a customer.

A modern subset of data engineering called predictive analysis frequently forecasts the likelihood or presence of data. It applies statistics, data mining, machine learning, and artificial intelligence techniques to examine both recent and historical data to generate predictions (Kumar & Garg, 2018). The process starts with an examination of historical data, and future event predictions are made following that assessment. Regression and classification are hailed as the two main objectives of predictive analytics (Selvaraj & Marudappa, 2018). It consists of many statistical and analytical techniques used to develop models that make predictions about potential future occurrences or outcomes. The use of predictive analytics allows for both continuous and discontinuous change (Elkan, 2013). The analytical methods utilized in predictive analytics include classification, prediction, and, to a certain extent, affinity analysis.

2.3.2.2 Sales and Demand Forecasting

Product sales histories, intra-category promotional schedules, and inter-category promotional schedules are all possible rich sources of information that might affect forecasting accuracy in a retail forecasting system (Ma, Fildes, & Huang, 2016). Demand forecasts are important pieces of knowledge in logistics management because they are used as the basis for many decisions, including those related to sourcing, production planning, logistics, inventory control, and retail decisions (Abolghasemi, Hurley, Eshragha, & Fahimniab, 2020). In today's rapidly changing and competitive world, accurately estimating customer demand remains a problem. However, little advances in this area assist diverse businesses minimize operational costs while boosting sales and increasing customer satisfaction. All large businesses depend on accurate demand forecasting to make important decisions about capacity construction, resource allocation, expansion, and forward or backward integration.

Accurate sales forecasts can assist management in choosing the right amounts of total inventory investments and serve as crucial inputs for numerous decision-making processes in functional areas like marketing, sales, production, purchasing, finance, and accounting (Fildes, Ma, & Kolassa, 2022). In this new era, where e-commerce is the norm and technological advancements and data collection systems are simultaneously flooding business systems with large volumes of transactional data populating corporate databases, demand planners and sales executives have realized that traditional approaches that combine product and market expertise, intuition, and instincts with the support of traditional legacy system functionalities will not suffice (Boone T. , Ganeshan, Jain, & Sanders, 2019).

2.3.3 Barriers to ML Adoption in SMEs

While the theoretical potential of ML in SME operations is well-documented, practical adoption is far from universal. Key challenges include insufficient technical know-how, lack of strategic awareness, high upfront costs, and inadequate digital infrastructure (Panghal, 2024). Furthermore, much of the existing research paints an overly optimistic picture by focusing on successful case studies from developed markets. This creates a skewed perception of ML accessibility and masks the realities faced by SMEs in sub-Saharan Africa.

A recurring theme is the disconnect between available ML tools and their usability in resource-limited environments. Cloud-based platforms, for instance, are touted as democratizing ML access, yet many SMEs lack stable internet or compatible devices to make use of such platforms. Even

when tools are available, their interfaces and required pre-processing steps often exceed the capacity of small business teams to manage independently. Policymakers and ecosystem stakeholders must also be considered in this conversation. Limited government support for SME digitization, inconsistent training programs, and absence of open data policies all contribute to a challenging adoption environment. This study not only reviews these structural barriers but also proposes a framework for incremental ML adoption that prioritizes scalability, simplicity, and alignment with SME business goals.

2.4 Research Gap

Although literature acknowledges the value of machine learning in optimizing business processes, several gaps remain. First, the bulk of research has focused on large firms in developed economies, with limited attention to SMEs in emerging markets—despite their substantial economic impact. Second, existing studies often highlight the benefits of ML in distribution broadly, yet fail to examine the practical limitations, such as digital illiteracy, low technology adoption rates, and cost barriers that disproportionately affect SMEs (Amberkar, 2018); (Panghal, 2024).

Additionally, few studies offer a critical synthesis of ML's effectiveness across varied distribution models, especially within informal sectors. There is limited evidence on how SMEs can sustainably integrate ML into their operations without heavy infrastructure investment. Despite a growing body of work on machine learning in distribution, significant gaps persist. First, most studies are concentrated in large enterprises or developed economies, overlooking the specific needs and constraints of SMEs in Africa. Second, even when SMEs are included, studies tend to describe benefits rather than evaluate limitations—especially the infrastructural, financial, and cultural factors that affect ML adoption. Third, there is limited research on hybrid solutions—models that combine manual and digital processes to ease SMEs into ML adoption gradually.

Due to the need to manage limited resources, shorter lead times, and rising customer demands, forecasts are now crucial in supply chain decision-making. It is critical to forecast the appropriate demand for each retail location essential for the success of every retail business since it promotes inventory control and improves distribution increases sales and customer happiness most significantly, reduces over- and understocking at each shop to save losses, and optimizes the distribution of produce between stores (Jain, Menon, & Chandra, 2015). Retailers can minimize losses from out-of-stock or non-selling products by computing accurate forecasts based on current sales transactions (Aulkemeier, et al., 2016). These days, it is essential to have retail sales

forecasting systems that are quick, dependable, and enhanced in terms of mistake reduction (Lalou, Ponis, & Efthymiou, 2020). The secret to entering competitive markets and reducing client acquisition expenses is having skilled and driven agents. Finding, training, and keeping an efficient agent network in marketplaces, however, requires meticulous organization, close supervision, time and resource commitment, and careful planning (Benhayoune & Repishti, 2015). To help managers improve the efficacy of their conventional sales forecasting toolset and techniques, this study suggests a methodological approach to the retail sales forecasting problem that is based on data analytics and statistical programming. It aims to evaluate not just the technical performance of ML models, but their viability given local business constraints.

2.5 Conceptual Framework

The value of the research is to provide businesses such the opportunity to faster respond to new customers and to increase sales trends, develop marketing campaigns, and more accurately predict sales and demand, ultimately improving the distribution process (Sathiya & Selvam, 2014). The study adopts a conceptual framework that positions distribution effectiveness as the dependent variable, influenced by three key independent variables—customer feedback**,** promotions and discounts, and routing and scheduling**.** The relationship is moderated by sales forecasting**,** which enhances predictive accuracy and operational planning. Distribution effectiveness is evaluated using metrics such as inventory turnover, profitability, and customer satisfaction, as shown in Figure 2.1.

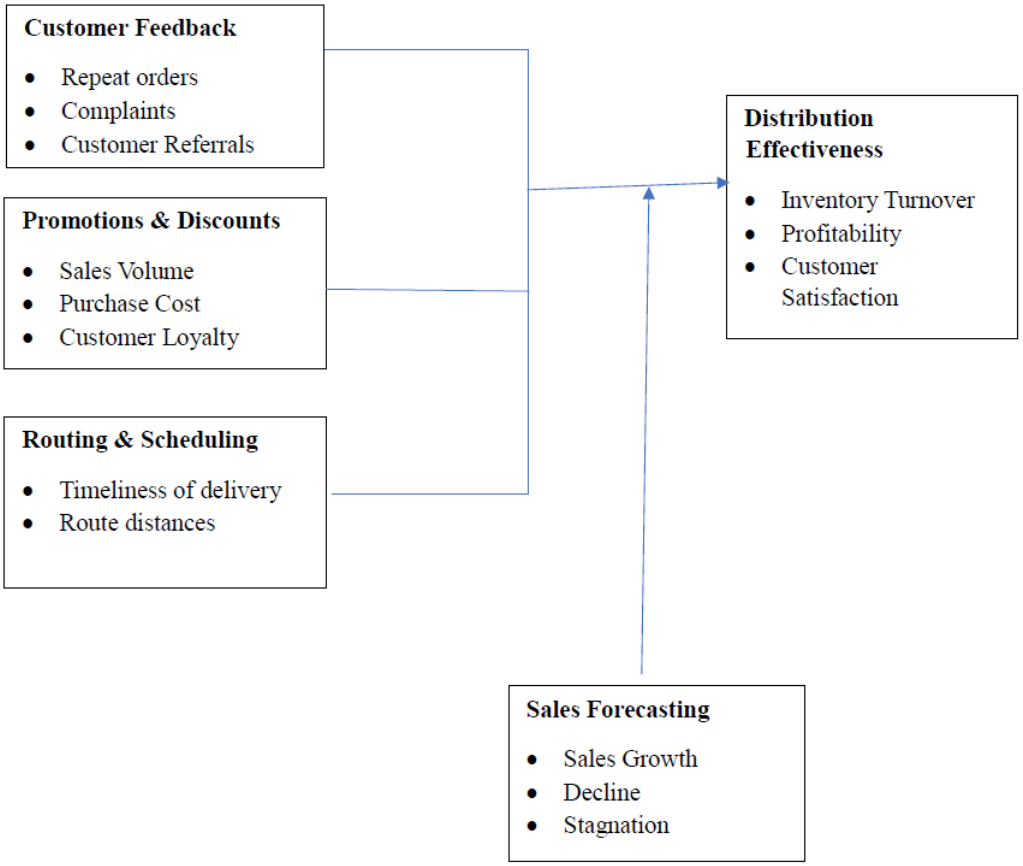


Figure 2.1: Conceptual Framework

Chapter 3: Research Methodology

3.1 Introduction

The chapter outlines the study's design, population, sampling approach, data sources, machine learning models, performance evaluation metrics, and ethical considerations. Special attention is given to the rationale behind model selection, the challenges associated with deploying machine learning in low-resource environments, and the practical steps taken to address data quality and privacy concerns. Wayne Goddard and Stuart Melville define research as a way of answering unanswered questions or creating that which does not exist (Goddard & Melville, 2004).

3.2 Research Design

A descriptive research approach, according to Sekaran and Bougie (2013), offers the knowledge of group characteristics in a specific context, permitting systematic thinking about elements in a given setting as well as generating incites for more study. To facilitate the applicability of the findings to a broader group of customers, particularly merchants, this study utilized a descriptive research approach (Sekaran & Bougie, 2013). Moreover, the researcher will be able to explain the study population's features as they are currently, which will minimize bias and increase the trustworthiness of the data gathered. A somewhat overall picture of what is happening at a particular time is provided by this approach, which also enables the creation of questions for more research. The CRISP-DM framework was chosen to facilitate rigorous development and testing of predictive models in line with real-world SME distribution dynamics.

3.3 Modularizing Data using CRISM-DM

Cross-Industry Standard Process for Data Mining (CRISM-DM) is a popular and widely used industry-standard process that helps in modularizing machine learning or data science projects into iterative steps. CRISP-DM is a domain-agnostic process that assists data scientists in building discipline and a framework around the execution of machine learning or data science projects. This particular project was broken down into six iterative stages using CRISM-DM as follows:

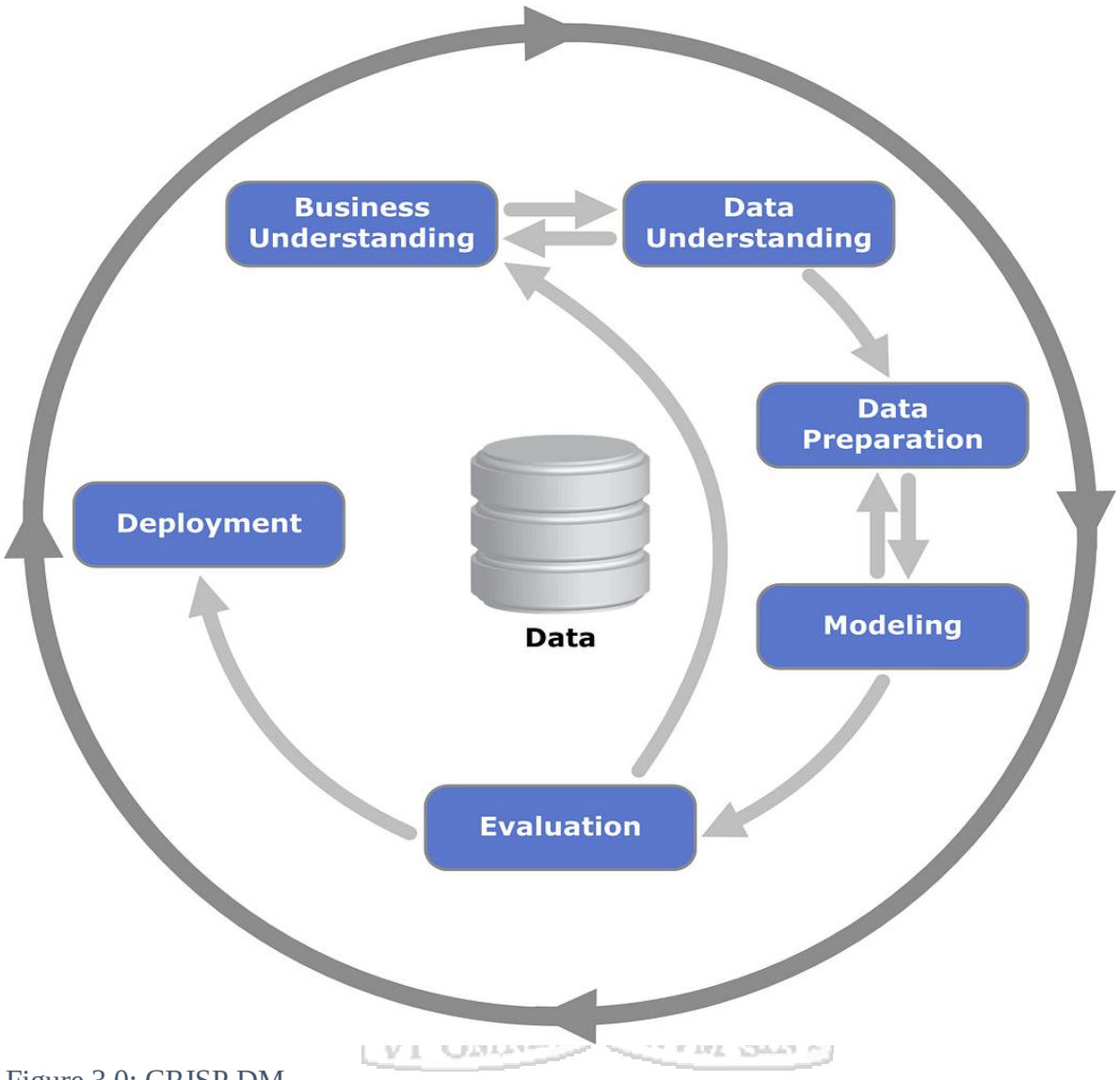


Figure 3.0: CRISP DM

3.3.1 Business Understanding

This will entail understanding the problem statement and the target user(s). The researcher will think about the problem being solved and why it is important. The gaps will then be identified in the current state in order to have a good understanding of how the problem is solved in the modern world. Then researcher then quantified the business impact that is expected to be achieved once the problem is solved for the target user/s. The business impact translated into outcome and output metrics while ensuring that success and failure metrics are defined before diving into the data. This ensures that the problem was well understood, domain expertise was gathered and relevant factor identified.

3.3.2 Data Understanding

This entailed the collection of data, validation of data, and performing exploratory data analysis. Data collection will involve sourcing the relevant data from identified sources, labeling the data if the data is not already labeled, and creating of relevant features. Data collection involves the systematic gathering of relevant information from diverse sources. To realize the objectives of this study, secondary data, multiple datasets were extracted from a database of an SME that focuses on distribution and delivery of various items in different parts of Kenya.

Table 3.1 shows the datasets used for analysis.

Table 3.1: A Summary of the Data Collected

Dataset	**Row**	**Column**	**Description**	**Column Names**
Time Spent per Route	513	10	Time spent by the sale people in a day	Location Code, Location Name, Route Code, Salesman Name, SOD Date, SOD Time, EOD Date, EOD Time, Total man hours, STATUS
Free Goods	910	9	Discounts and promotion	Location Name, Customer Name, Invoice Number, Document Date, Promotion Description, Item Code, Item Description, Free Good Qty, Free Good Value
Sales Register	28665	27	Sales in a day	Outlet Code, Outlet Name, Invoice No, Invoice Date, Inv Line No, Item Code, Item Name, Hierarchy Name, CategoryDescription1, Conversion1, MRP, Item Quantity, Line Amount, Discount Amount, Taxable Line Amt, VAT Tax Rate, VAT Tax Value, CESS Tax Value, Tax Amount, Net Amt, DT

				Landed Price, Route Code, Route Name, Salesman Name, Vehicle Reg.No, Load Out No, Order No
Salesperson data	27919	20	Salesperson details and what they sold	Salesman Name, Outlet Code, Outlet Name, Invoice No, Invoice Date, Item Name, Hierarchy Name, Category Description1, Item Quantity, Route Code, Route Name, Location Name, Route Code, SOD Date, SOD Time, EOD Date, EOD Time, STATUS, Total man hours Package

3.3.2.1 Validation of Data

Validation of data involved handling missing or erroneous data and outliers. This step also involved performing quality control on the data to ensure the values were what the researcher expected, and appropriately cleaning the data. The data was then be explored to perform statistical analysis as well as data visualization, any relationships and patterns were identified, and dimensionality reduced. This phase was essential since the effectiveness of the data mining algorithms was directly impacted by the quality of the generated data.

3.3.3 Preparation of Data

In the data preparation phase, raw data was cleaned and organized. Missing values were imputed, categorical variables were label-encoded, and numeric features were scaled where necessary. The data was then partitioned into training and testing subsets, with an 80:20 split. Outliers were addressed using interquartile range techniques to reduce model sensitivity to extreme values. This was to lower the amount of noise or complexity in the raw data, to aggregate the data more thoroughly, and in certain situations, just to keep less data overall (ACAPS, 2016).

3.3.4 Data Modeling

Data modeling entailed both the selection of model and tuning. The modeling process involved the evaluation of various algorithms through cross-validation, hyper-parameter optimization or tuning, versioning and documenting of the model, and subsequently model retraining. Necessary

trade-offs (performance, computational cost, and interpretability) were conducted when choosing the optimal algorithm as part of the model selection.

The modeling phase included the evaluation of several algorithms: Linear Regression, ARIMA, Decision Tree, Random Forest, and Gradient Boosting (XGBoost). While regression and ARIMA models were conceptually valid for sales prediction, they were not selected due to practical limitations. Linear regression assumes a linear relationship between inputs and outputs, which does not reflect the non-linear nature of most SME distribution systems. ARIMA is designed for univariate, stationary time series and is less effective for datasets with multiple features and irregular intervals. This distinction highlights the study's shift from theoretically grounded but practically limiting models toward empirically robust ensemble methods, aligning with real-world data complexity encountered in SME distribution scenarios.

Random Forest was selected for its ability to handle noisy data and prevent overfitting by averaging multiple decision trees. Gradient Boosting (XGBoost) was chosen for its iterative error correction mechanism and higher predictive accuracy. Both models are suitable for small datasets and work well with tabular data common in SMEs. Hyperparameter tuning was conducted using GridSearchCV. This involved setting optimal values for model parameters such as tree depth, number of estimators, and learning rate. To ensure generalizability, the models were validated using five-fold cross-validation.

3.3.5 Evaluation and Implementation

Model evaluation involved metrics such as Mean Absolute Deviation (MAD), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE). These indicators were used to assess the precision and robustness of each model. Additionally, feature importance analysis was conducted to determine the most influential predictors. Different analyses, such as descriptive statistics on the attributes of the dataset collected and value comparisons will be used. Correlation analysis, and model performance evaluations will be integral in this study. This is where the model will be scored on the test set and the output of the model interpreted and performance evaluated. Unit and integration tests will be written to test the model to make it more robust. Subsequently, a user test will be done to build more confidence on the model before it is operationalized. Descriptive and inferential statistics will be employed to analyze the quantitative data.

Forecasting accuracy is an essential parameter to take into account when evaluating a forecasting model's efficacy (Zhang, 2016). This word describes the difference between the expected and

actual value, expressed as a percentage of the actual value. Three popular forecasting techniques are Mean Squared Error (MSE), Mean Absolute Deviation (MAD), and Mean Absolute Percentage Error (MAPE). Mean Squared Error (MSE) is a metric used in statistics and data science to measure the difference between observed values (actual or predicted data) and values predicted by a model or estimate. According to Maricar, (Maricar, 2019) MSE is a calculation used to calculate the average rank error, the calculation formula is as follows:

MSE = $(Actual - Forecast)$ 2 / n − 1 (1)

MSE is the result of reducing the actual and forecast values which are then squared. Root Mean Squared Error (RMSE) RMSE is a metric widely used in statistics and machine learning to measure the accuracy of a predictive model or the error between the predicted value and the actual observed value. According to (Fadil Indra Sanjaya, 2020), RMSE is an alternative method for evaluating forecasting techniques used to measure the accuracy of the forecast results of a model.

RMSE = $\sqrt{\frac{1}{n}\sum_i^n(\bar{y}i - yi)^2}$

Where;

$\bar{y}i$= the value of the forecasting results

yi =actual value

n = amount of data.

A variety of qualitative and quantitative methods are also available to forecasters. Blasco, Moreno, and Abad assert that forecasting should make use of MAPE (Moreno, Pol, Abad, & Blasco, 2013). Therefore, MAPE models use accurate percentages that are calculated throughout time and are therefore seen to be suitable for forecasting. Because the MAPE method can accurately measure the difference between actual and forecasted sales, it is therefore commonly used in retail organizations (Badr & Ahmed, 2023). This study used MAPE and MAD to evaluate the performance of different models.

After evaluating the model's performance and testing the solution, the model was deployed adhering to the security measures put in place and the software deployment process. The final model was deployed through a simple user interface built with Streamlit, enabling SME users to interact with the predictive tool and generate real-time insights based on input variables. The researcher understood that it was paramount to monitor the performance of the model, and potentially re-training the model if need be (Muhammad Aprilianto, 2022). This being an iterative

process, the researcher might continue switching back and forth in any case the requirements of the project change and/or when there is no satisfaction with the outcomes. The final model was deployed through a simple user interface built with Streamlit, enabling SME users to interact with the predictive tool and generate real-time insights based on input variables.

3.4 Ethical and Practical Considerations

The study adhered to ethical research protocols, including informed consent**,** data anonymization**,** and compliance with university ethical review guidelines. The SME was informed about the nature and scope of the data usage, and all identifying details were masked. The research adhered to the ethical guidelines set by the host institution and received formal clearance before data collection commenced. Given the realities of Kenyan SMEs, the study also considered practical deployment barriers**,** including unreliable internet, limited digital infrastructure, and low algorithm literacy. As a mitigation strategy, simpler models were prioritized for deployment, and results were visualized in a user-friendly dashboard built with Streamlit. This approach ensured the model could be adopted in real-world settings without requiring extensive training or infrastructure.

Chapter 4: System Design and Architecture

4.1 Introduction

This chapter looks at where data management intersects with Machine Learning (ML) by describing the design and architecture of a system that automates the prediction and forecasting of sales through delivery and distribution channels. It shall also describe how a system should be set up, thereby creating an understanding of the system architecture and its general structure.

4.2 Machine Learning Architecture Overview

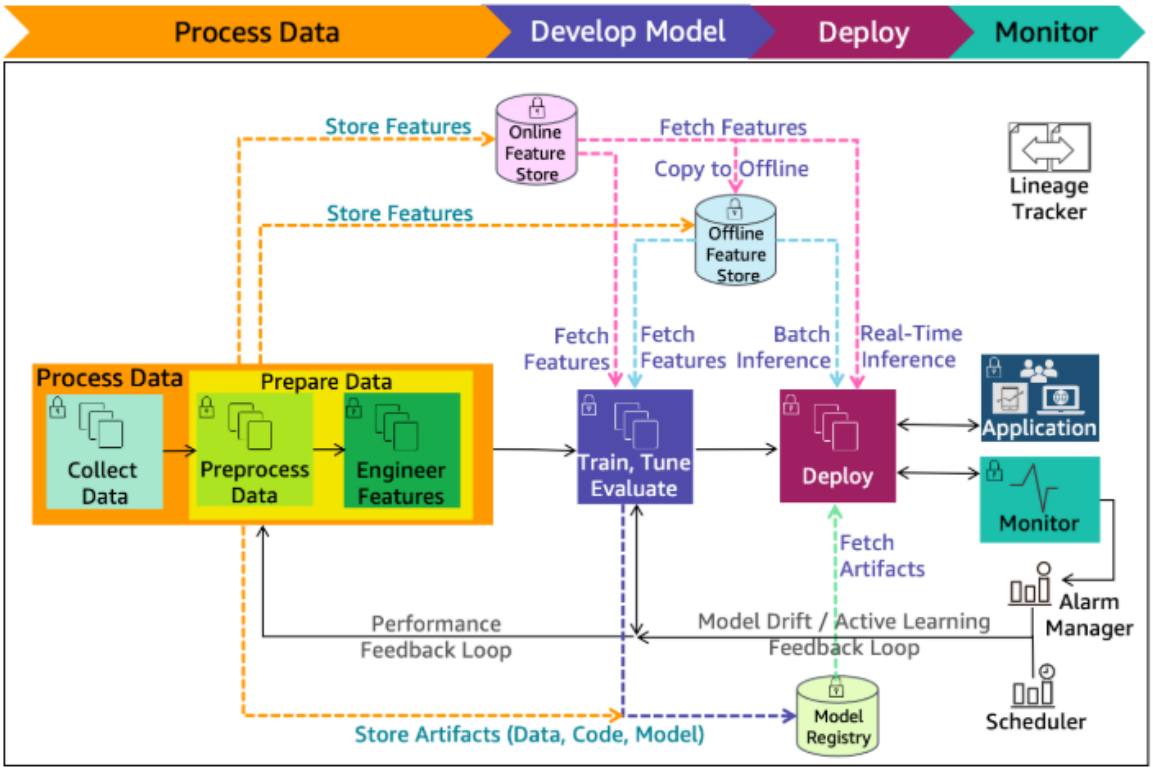


Figure 4.1 ML Architecture Diagram

The model development phase, model monitoring phase, and model monitoring phase are the stages of the machine learning lifecycle that follow the issue framing phase and demonstrate how the data-processing sub-phases interact with the later stages, as shown in Figure 4 1 (Amazon Web Services, 2023). Training, fine-tuning, and evaluation are all part of the model-creation process. The staging environment for model validation for security and resilience is part of the model deployment step. Monitoring is essential for drift mitigation and timely identification. Throughout

the many stages of the ML lifecycle, feedback loops are essential for monitoring. Consistent and reusable features are offered by feature stores, both online and offline, during the model development and deployment stages. The model registry enables version control and lineage tracking for model and data components (Amazon Web Services, 2023). This study has used Streamlit, an open-source framework developed to enable users to view forecasts, input data, and visualize results through the user interface.

4.3 System Components

The backend, frontend, and data flow all cooperate to produce a fully integrated machine learning system, as explained in detail in this section. The system will be scalable, effective, and user-friendly if these components are designed, enabling SMEs to employ machine learning for optimal distribution.

4.3.1 Model Deployment

In order to properly deploy and integrate the model for prediction, the trained model should be saved as shown in Figure 4 2, thereby allowing the model to be serialized and stored for future use. Predictions are made from the memory of the saved model. Errors like invalid inputs or failed model predictions should be handled by the backend. Clear error messages should be displayed, for example, if a user enters incorrect data or if the model is unable to produce predictions.

```
from sklearn.ensemble import RandomForestClassifier
import pickle

# Assuming you have trained your model
model = RandomForestClassifier(n_estimators=100, random_state=42)
# Train your model here...

# Save the model
with open('model.pkl', 'wb') as file:
    pickle.dump(model, file)
```

Figure 4.2: Saving the trained model

4.3.2 Streamlit Interface

Streamlit includes a range of input widgets that can be used to collect data from users. Using the layout tools in Streamlit, the interface is arranged so that the results or visualizations are on the right side of the page and the inputs are on the left, as shown in Figure 4 3. After the user submits

inputs, the results are displayed on the same page. Streamlit provides simple ways to show text, tables, and plots.

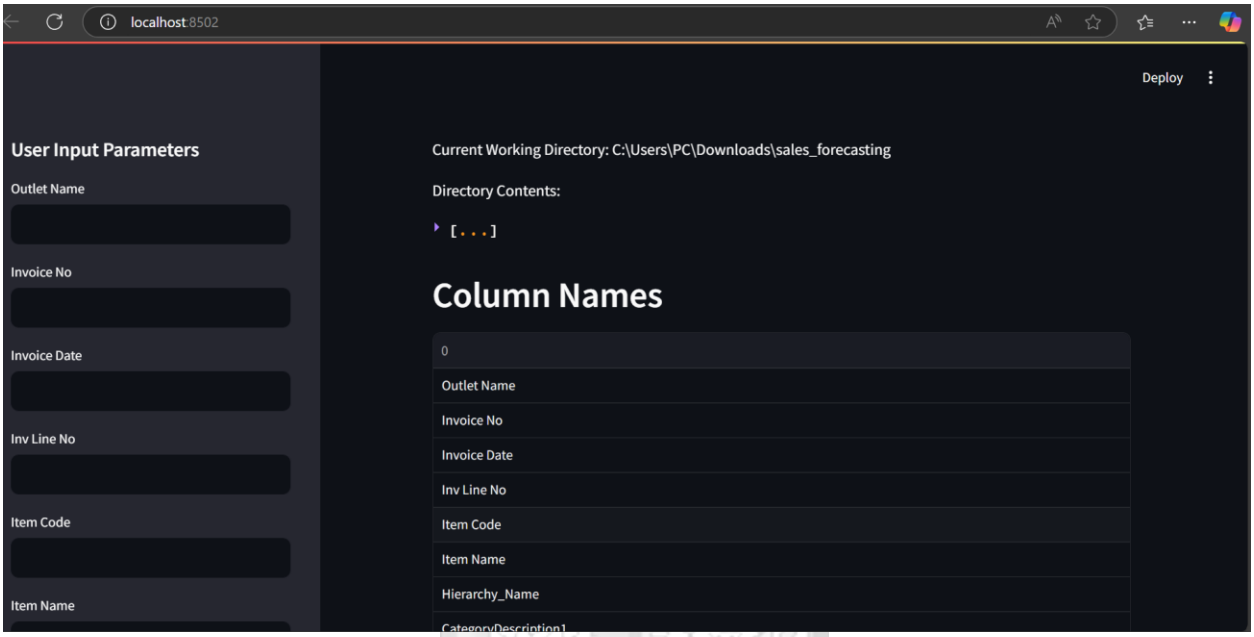


Figure 4.3: Streamlit User Interface

4.3.3 Data Flow

This section describes the flow of data from input to output to prediction in the system. The user inputs data into the Streamlit UI to start the system. These inputs could be categorical information (like product category, delivery zone) or numerical values (like distance, demand). These inputs are captured as variables by Streamlit, which then converts them into a backend-understandable format (such as a Pandas DataFrame or a NumPy array). Backend data preprocessing: It can be necessary to preprocess the input data once it has been gathered before feeding it into the model. This might entail: use MinMaxScaler or StandardScaler to scale the data, for example. The data is fed into the trained Random Forest model for inference by the backend following preprocessing. The model returns the outcome after making a forecast based on the input data. The predicted outcome is transmitted back to the frontend (Streamlit) after it has been generated by the backend. The results are subsequently presented in an approachable manner by the frontend (e.g., text or visualizations). Here is where you demonstrate to the user how well the distribution process works or how they can change certain parameters to get better results.

Chapter 5: Results Presentation

5.1 Introduction

The next section presents and explains the research findings that were inferred from the secondary data that was gathered. The analysis and findings chapter provide a comprehensive synopsis of what was discovered from the research on machine learning-based sales optimization. Using a number of analyses, such as descriptive statistic of sales quantity and value comparisons across various items descriptions and routes used, correlation analysis, and model performance evaluations, the part offers significant insights into the changing patterns of sales management and prediction of demand. Nonetheless. figures illustrating sales value distributions and sales quantity based on the item descriptions c are examples of visual representations of the findings. As such, the main objectives of this part involved clarifying the relationships between different variables, assessing how effectively different machine learning models predict demand and optimize sales levels, and validating the performance of the best model.

5.2 Descriptive Statistics

Pandas' describe operation was used to compute descriptive statistics, including metrics for dispersion (range and standard deviation) and central tendency (sum and mean). Understanding the data's characteristics and guiding future research were made easier by descriptive statistics, which provided information about the dataset's central tendency, variability, and distribution.

The following table provides descriptive details regarding the particular time period of the participating financial industries. These figures cover the quantity of items, sale values, and sales quantity as a measure of the products sold to small and medium enterprises.

5.2.1 Descriptive Statistics of the Quantity of Items

In accordance with the category of the products, the results shown below provide the descriptive statistics of the total number of products distributed to different small and medium sectors.

Table 5.1: Descriptive Statistics based on the category of the product

Descriptive Statistics

CategoryDescription1		N	Minimum	Maximum	Sum	Mean	Std. Deviation
BABY JUNIOR	Item Quantity	19	1	30	187	9.84	8.153
	Valid N (listwise)	19					
MOVIT	Item Quantity	20614	1	4320	474877	23.04	94.744
	Valid N (listwise)	20614					
NAN	Item Quantity	1	30	30	30	30.00	.001
	Valid N (listwise)	1					
PINE	Item Quantity	19	1	48	201	10.58	10.590
	Valid N (listwise)	19					
RADIANT	Item Quantity	7265	1	1440	95548	13.15	31.661
	Valid N (listwise)	7265					

Table 5 1 above provides descriptive data on the quantity of things in small and medium-sized firms based on the various routes. According to the results of the aforementioned analysis, the aggregate sum, mean, and standard deviation of the number of items that involved Movit products were 474877, 23.04, and 94.744, respectively. Furthermore, Table 5 1 shows that the total number of Radiant products was 95548 and the average number of items distributed was 13.15, with a standard deviation of 31.661. The descriptive statistics shown illustrate that Pine had a quantity total of 201, while the average number of these items distributed was 10.58, with a standard deviation of 10.590. However, for the product representing Baby Junior, their quantity total was 187, while the average quantity number was 9.84, with a standard deviation of 8.153.

Lastly, the descriptive statistics indicate that the average and total quantity NAN goods shipped were 30, and 30.00, respectively, with a standard deviation of .001.

5.2.2 Descriptive Statistics on Quantity of Items Based on Route

The findings presented below illustrate the descriptive statistics of the overall quantity of items that are distributed towards various small and medium sectors based on the route taken.

Table 5.2: Quantity of Items Based on Route

Descriptive Statistics

Route Name		N	Minimum	Maximum	Sum	Mean	Std. Deviation
KASARANI	ItemQuantity	10501	1	3744	154495	14.71	70.704
	Valid N (listwise)	10501					
KAYOLE	ItemQuantity	11977	1	2520	159443	13.31	47.555
	Valid N (listwise)	11977					
MACHAKOS	ItemQuantity	5440	1	4320	256905	47.23	141.184
	Valid N (listwise)	5440					

The descriptive information of the quantity of items in small and medium-sized businesses according to the different routes is shown in Table 5 2 above. As per the findings in the above analysis, the sum, the average number, and the standard deviation of the quantity of Items for the products that were shipped via Machakos were 256905, 47.23, and 141.184, respectively. Additionally, as indicated by Table 5 2, the sum and average quantity of items that were distributed via the Kasarani route had a respective value of 154495 and 14.71, with a standard deviation of 70.704. Lastly, statistics show that the sum and mean average quantity of the items shipped via the Kayole route were 159443 and 13.31, while the standard deviation showed a value of 47.555. The results above suggest the quantity number of items changes between the different routes due to route efficiency which might influence the amount of goods transported; routes with greater effectiveness are capable of handling larger quantities, whereas less efficient

ones might restrict the amount. Transportation expenses can additionally impact the quantity of goods shipped between the routes, with some of them being more costly because of longer distances, fuel prices, or tolls.

5.2.3 Descriptive Statistics of Sales Quantity

The data displayed below gave descriptive information about the total number of sales that the SME had been distributing to various small and medium sectors, according to the route taken.

Table 5.3: Descriptive Statistics of Sales Quantity

Descriptive Statistics

Route		N	Minimum	Maximum	Mean	Std. Deviation
	SalesQty	32	192	37440	5475.00	8641.014
	Valid N (listwise)	32				
BACK OFFICE	SalesQty	199	0	2437285	74895.62	248704.779
	Valid N (listwise)	199				
ELECTOR OBURA	SalesQty	59	1	20031	532.98	2641.310
	Valid N (listwise)	59				
KASARANI	SalesQty	123	1	56590	2216.84	6393.481
	Valid N (listwise)	123				
KAYOLE	SalesQty	129	1	65012	1961.84	7110.346
	Valid N (listwise)	129				
MACHAKOS	SalesQty	91	1	104045	3627.93	13579.740
	Valid N (listwise)	91				

Table 5.3 above displays the descriptive data of the sales quantity rate in the small and medium enterprises based on the various routes. The analysis found that the average quantity of sales for those products that were transported using the back office was 74895.62, with a standard deviation of 248704.779. Furthermore, Table 5.3 shows that the average sales quantity with regards to those products distributed using the Machakos route was 3627.93, with a standard deviation of 13579.740. Additionally, the sales quantity of the products that were transported using the route of Kasarani had an average value of 2216.84 and a standard deviation of 6393.481 (Table 5.3). In addition, according to the identified statistics, the average sales quantity for the products transported through the Kayole pathway was 1961.84, with a standard deviation of 7110.346. Lastly, it had also been revealed that the mean of the sales quantity for products delivered over the Elector Obura route was 3627.93, with a standard deviation of 13579.740. Therefore, the aforementioned results indicate that Back Office, Kasarani, and Machakos are the top three routes used to achieve the highest sales quantity of the available quantity of beauty products, surpassing a mean of 2000 items. However, routes such as Kayole was used to transport the least portion of the total sales quantity made available, which was under 2,000 items in total.

5.2.4 Trends of sales Quantity by item description

Figure 2 displays the number that forecasted the total sales value generated by the various goods. In contrast, Styling Gel 250 GM, Blowout 150 GM, Cult Activator 700 GM, and Movit Herbal Jelly 425 GM had the largest anticipated total revenue of 5.00E7 and more, while Apple Shampoo 1L, Hair Food 120 GM, and MOVIT Styling GEL 500G had the lowest projected total sales value of less than 10,000.

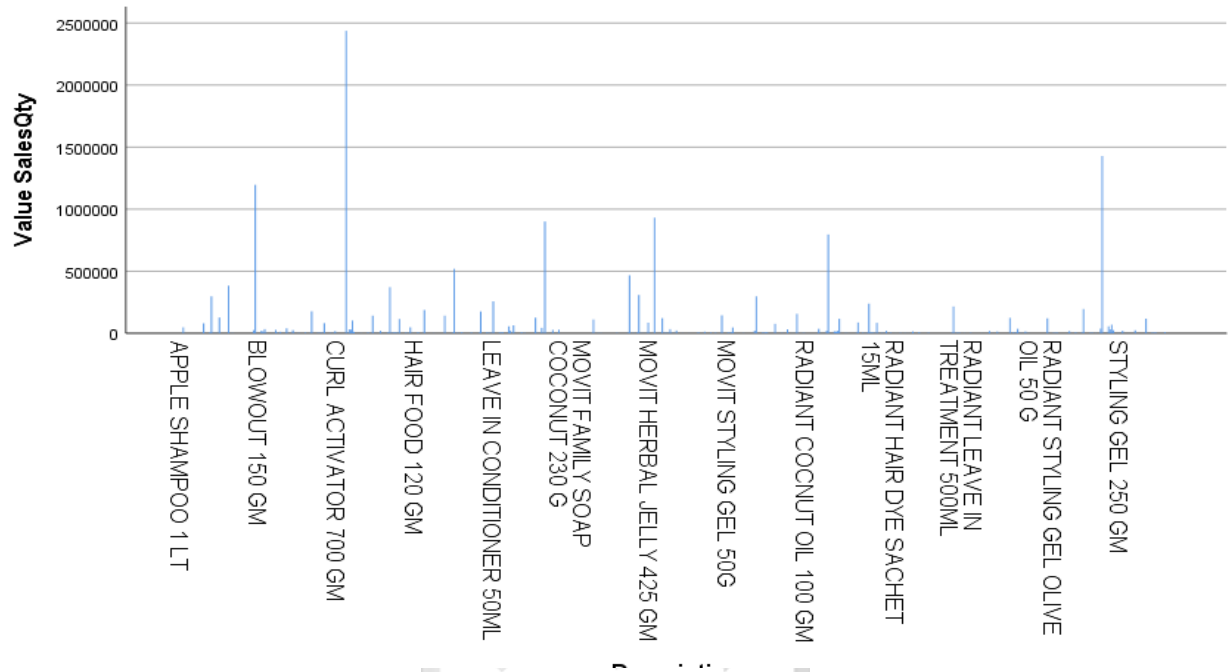


Figure 5.1: Total sales quantity by Item Description

Based on the illustrations on Figure 2, Styling Gel 250 GM, Blowout 150 GM, Cult Activator 700 GM, Coconut 230 GM, Movit Baby Jelly 200GM, and Movit Herbal Jelly 425 GM had the largest anticipated total revenue of 5E5 and more, while Apple Shampoo 1L, Movit Ovacado Oil 425 GM, Lemon Gel 80G, Hair Food 120 GM, and MOVIT Styling GEL 500G had the lowest projected sales quantity of less than 5E5. These results suggest that customers might be inclined to pay more for some items than others, and products with distinctive or appealing characteristics and advantages might perform better than those without.

5.2.5 Descriptive Statistics of Sales Value

Descriptive information regarding the total sales value gained for delivering to different small and medium sectors based on the route taken is provided by the statistics given below;

Table 5.4: Descriptive Statistics of Sales Value

Descriptive Statistics

Route	N	Minimum	Maximum	Sum	Mean	Std. Deviation

	SalesValue	32	21591.300000000000	4356310.410000000000	16175089.370000000000	92233.72036854776000	9223.372036854777000
	Valid N (listwise)	32					
BACK OFFICE	SalesValue	199	.000000000000	171144388.090000000000	1205076442.039999500000	92233.72036854776000	17676608.856623568000000
	Valid N (listwise)	199					
ELECTOR OBURA	SalesValue	59	.000000000000	.000000000000	.000000000000	.000000000000000	.000000000000000
	Valid N (listwise)	59					
KASARANI	SalesValue	123	84.000000000000	7232415.520000000000	30360449.110000000000	92233.72036854776000	9223.372036854777000
	Valid N (listwise)	123					
KAYOLE	SalesValue	129	180.000000000000	8144260.570000000000	27480666.910000000000	92233.72036854776000	9223.372036854777000
	Valid N (listwise)	129					
MACHAKOS	SalesValue	91	96.000000000000	9046591.240000000000	29083401.150000000000	92233.72036854776000	9223.372036854777000
	Valid N (listwise)	91					

In the assessment of the optimal distribution and delivery model for small and medium sectors, table 5 4 above displays the descriptive statistics of the sales values based on the routes that were used while distributing the products. The analysis found that the sales value from the route of Machakos, Kayole, Kasarani, Elector Obura, and Back Office all had an average mean of 92233.7203 with a standard deviation of 9223.37. These findings indicate standardized pricing tactics are frequently used by businesses to prevent price wars and preserve brand consistency. This suggests that irrespective of the route taken, the identical product is offered for a comparable price.

5.3 Trends Analysis

5.3.1 Sales Value by Item Description

The metric predicted total sales value produced by the different items is shown in Figure 5 2. As shown, Apple Shampoo 1L, Hair Food 120 GM, and MOVIT Styling GEL 500G, had the lowest estimated total sales value of less than 10,000, whereas Styling Gel 250 GM, Blowout 150 GM, Cult Activator 700GM, and Movit Herbal Jelly 425 GM had the highest total income of 5.00E7 and over (see figure 5 2). The sales value of commodities might vary depending on their specifications because products in great demand can be sold at higher prices, particularly if there is limited availability. In addition, because of their perceived longevity and worth, higher-quality products sometimes fetch greater prices. Due to their recognized reputation and devoted following, brands that are well-known are able to charge higher prices for their products in comparison to those which are not well known. Lastly, items with distinctive or novel features may be more expensive since they provide something new from conventional products

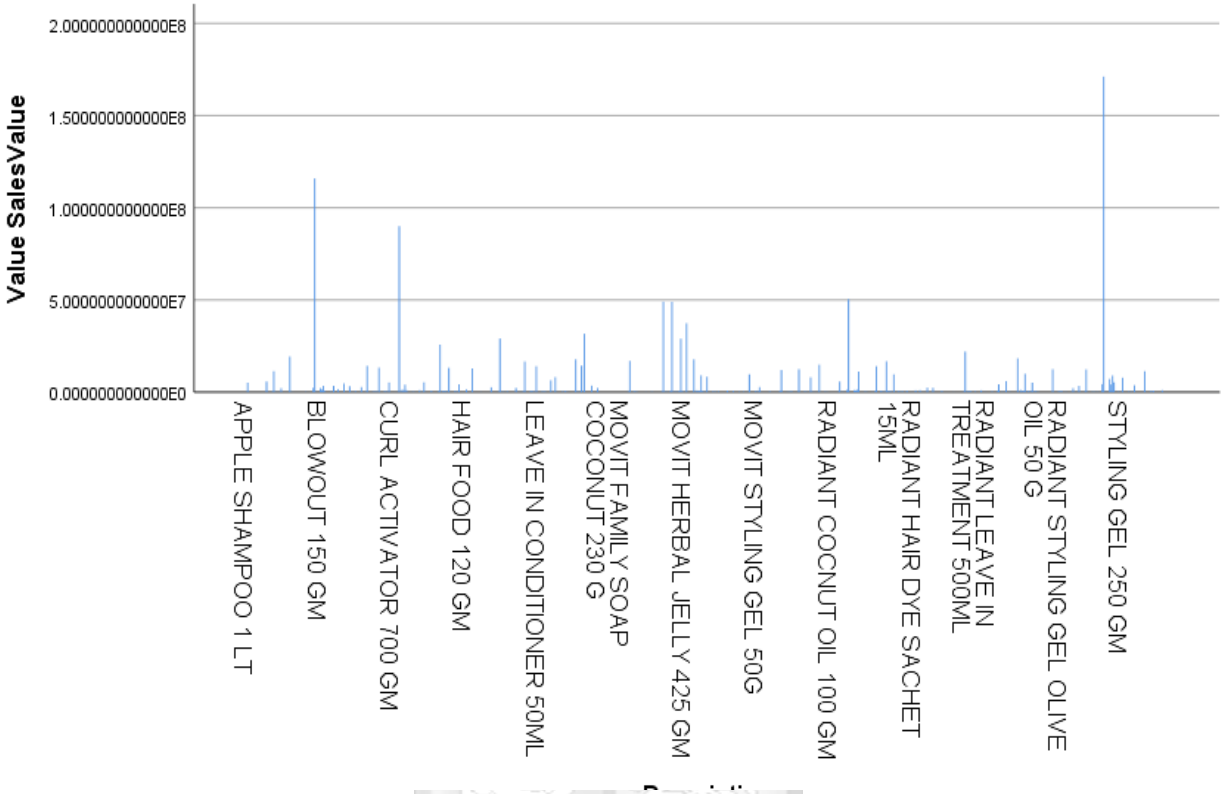


Figure 5. 2: Total sales value by Item Description

5.3.2 Trend Analysis of Sales Value of Various Commodities

Trend analysis entails the process of gathering data over time and examining it to find recurring patterns or trends. The results shown in Figures 5 3 and 5 4 below illustrate the trend analysis of sales values for various beauty products that were distributed in 2022 and 2023, respectively.

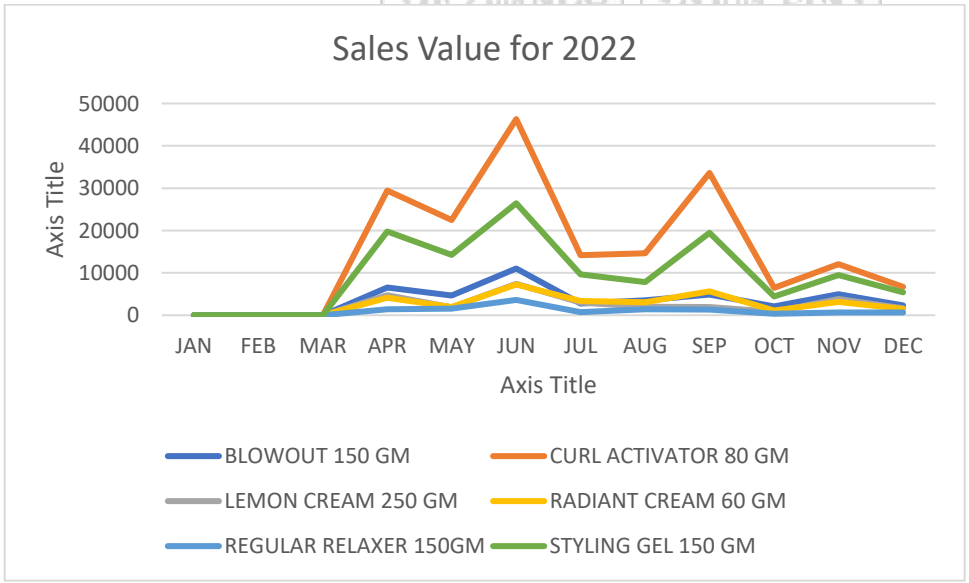


Figure 5. 3: Trend analysis of sales of commodities in 2022

As can be seen from Figure 5 4, curl activator 80 GM demonstrated a high sales value, with its maximum value (46315) occurring in June and its lowest value (0) occurring in March. Styling Gel 150 GM was the next most valuable item in terms of sales, with the highest value (26441) occurring also in June and the lowest value (0) occurring as well in March. However, when compared to other cosmetic items, Radiant Cream 150 GM and Regular Relaxer 150 GM products had the lowest sales values, with values ranging from 0 to 7244 and 3612, respectively. These illustrations suggest that aspects of seasonality, which affect overall sales trends, are the reason for changes in the patterns of sales value for beauty items. This is because sales frequently peak or go down around specific periods of the year, which include holidays or seasonal shifts, which as a result affect the sales value.

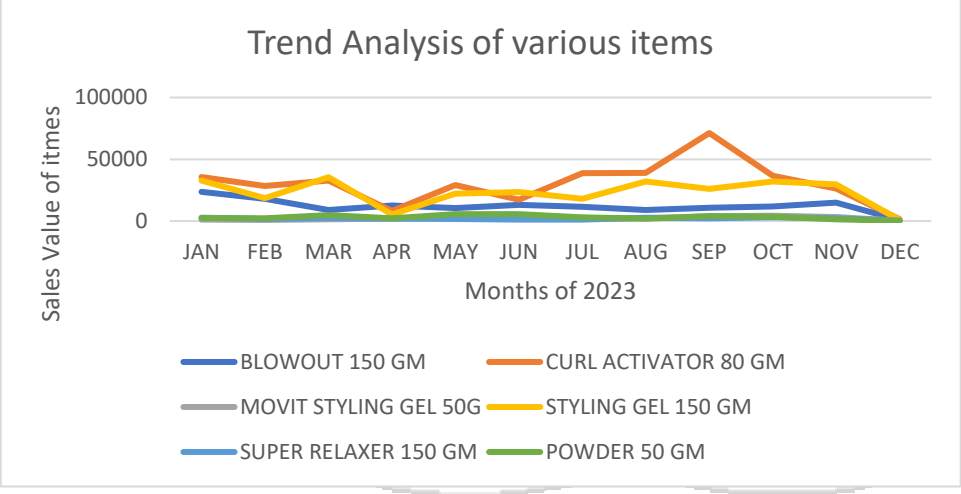


Figure 5. 4: Trend analysis of sales of commodities in 2023

However, unlike in the year 2022, from Figure 5 4, it has been shown that curl activator 80 GM illustrated a high sales value which had the lowest value of 1554 in December while its highest value (71243) was recorded in September. The next item with the second most sales value was Styling Gel 150 GM whose highest value (35503) was recorded in March, while the smallest sales value (5253) for this commodity was observed in April. Nonetheless, the products of Movit Styling Gel 50 GM and Super Relaxer 150 GM showed the lowest sales values in comparison to other beauty products with values ranging between 252 to 4349. Therefore, based on the illustrations in Figure 5 3 above, it can be concluded that due to the growing consumer knowledge of skin well-being and the appeal of organic and natural components, there is a drive

of substantial variations in the trend or patterns of sales value in the market for the beauty products.

5.4 Relationship between Sales Value and Sales Quantity (Correlation Analysis)

This research sought to identify whether the total sales of items and their sales value were related. Firstly, a scatterplot was plotted to show how this relationship. As seen in Figure 5 5 shown below, the quantity of items being sold and the expected total value from the selling of these different products have the potential to be positively correlated.

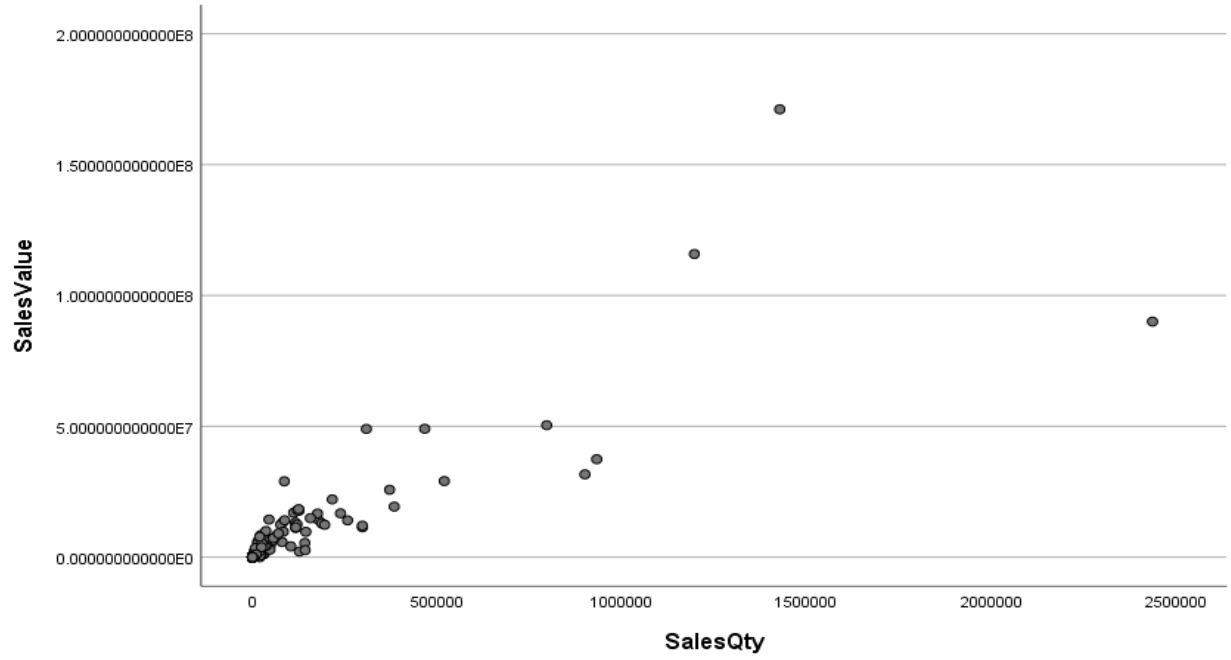


Figure 5.5:Association between sales quantity of items and their sales value

However, to statistically confirm the results of the figure above, a correlation analysis was conducted to show the direction of the relationship and by how much through calculating a correlation coefficient. Table 5 5 displays the relationship between the sales quantity of the item and the sales value variable used in the dataset. The aforementioned variables were determined to have strong correlations with one another since they had a correlation coefficient of r= .866 $p<0.01$. Therefore, as the correlation statistics have shown, the data variables (sales quantity and sales value) illustrate a strong correlation with one another.

Table 5.5: Correlation Coefficients

Correlations

		Sales Value	Sales Qty
Sales Value	Pearson Correlation	1	.866**
	Sig. (2-tailed)		.000
	N	633	633
Sales Qty	Pearson Correlation	.866**	1
	Sig. (2-tailed)	.000	
	N	633	633

**. Correlation is significant at the 0.01 level (2-tailed).

5.5 Sales and Demand Forecasting Models

The current research used Decision Tree Regression, Linear Regression Model, and K-Nearest Neighbors Regression model as the systems and models in the forecasting of sales and demand of beauty items involved. The findings of a comparison study of the effectiveness of regression models for sales of items optimization and their demand forecasting are displayed in Table 5 6 and Figure 5 7. The primary instrument for evaluating the precision of predictions was the mean squared error (MSE). The Decision Tree Regression model has the lowest mean MSE (M = 0.0392) among the several underlying regression models and was followed by the Linear Regression model (M = 0.0946). Nonetheless, ensembled methods outperformed the aforementioned models since the mean MSE values for Bagging (M = 0.0153) and Boosting (M = 0.0133) were the smallest. Additionally, stacking outperformed Decision Tree Regression and Linear Regression since it had a value as follows (M = 0.0278). On the contrary, out of all the models examined, the K-Nearest Neighbors Regression model had the largest mean squared error (M = 0.2321). These findings demonstrate that ensemble approaches, particularly bagging and boosting, as designed and developed machine learning model that forecasts sales and demand of Small and Medium Enterprise as this predictive model increases the precision of predictions for sales quantity optimization since the findings demonstrate that these machine learning algorithms perform better than the traditional projection methods.

This study performed a hyperparameter adjustment to verify the functionality of the machine learning models developed to address the business issue. More specifically, following hyperparameter adjustment, the Decision Tree Regression model exhibited the smallest mean MSE (M = 0.0367), subsequent to the Linear Regression model (M = 0.0947), as shown in Table

5 6 and Figure 5 7. On the other hand, combined techniques fared better than the single ones; the mean MSE values for boosting (M = 0.0142) and bagging (M = 0.0134) were the lowest. Additionally, stacking outperformed decision tree regression and linear regression, as it achieves more competitive results (M = 0.0261) in comparison to the aforementioned model. However, of the models examined, the K-Nearest Neighbors Regression model significantly had the largest mean MSE (M = 0.2277). These results demonstrated that for sales quantity optimization using machine learning algorithms, ensemble strategies such as bagging and boosting may offer higher prediction accuracy than individual base regression models. Therefore, hyperparameter adjustment was able to validate the machine learning models to improve its functionality for solving the business issue of forecasting the sales and demand of beauty products for SMEs.

Table 5.6: Mean squared errors of the model's performance both before and following hyperparameter adjustment

	Model	Mean MSE	Updated Mean MSE
0	Linear Regression	0.094599	0.094679
1	Decision Tree	0.039249	0.036655
2	KNN	0.232146	0.227727
3	Stacking	0.027757	0.026053
4	Bagging	0.015285	0.01338
5	Boosting	0.013271	0.014184
6	Voting	0.075026	0.049367

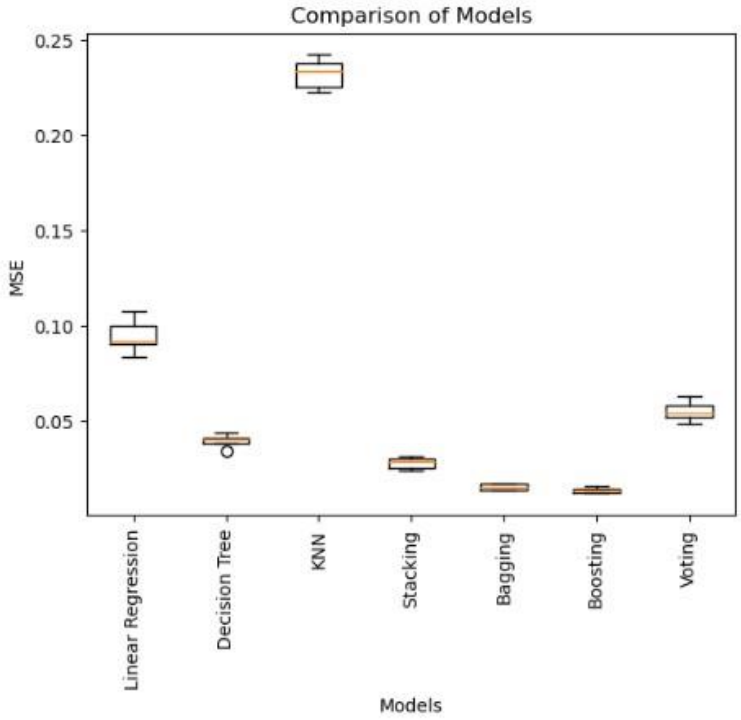


Figure 5.6: Comparison of the models' performance

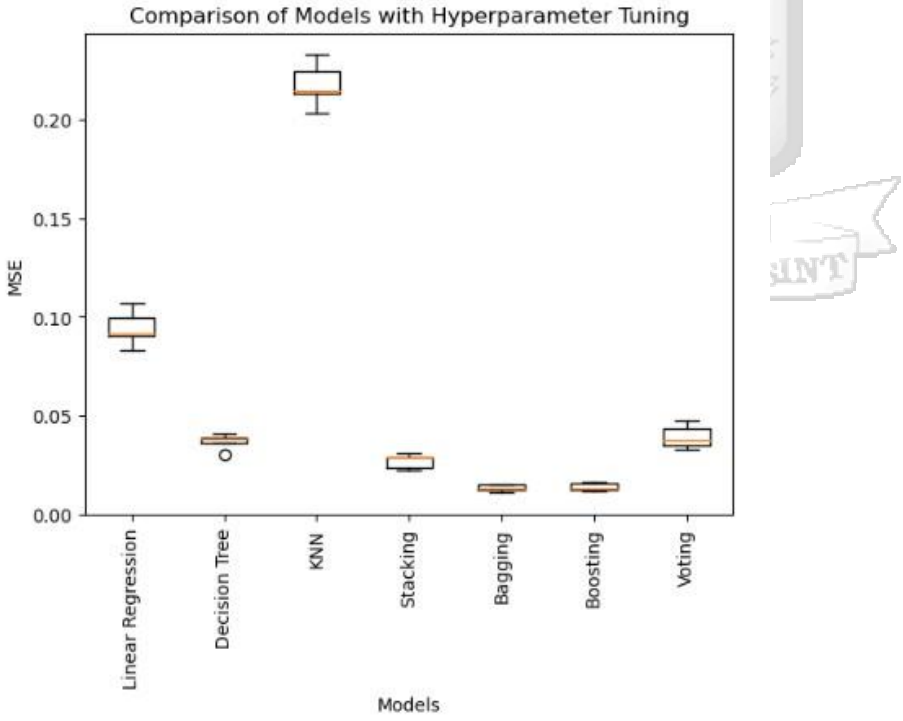


Figure 5.7: Performance comparison of the models following hyperparameter tuning

5.6 Validation of the Forecasting Machine Learning Model

The aforementioned forecasting models were supplied with the sale quantity and value data as an input. The forecasting approaches are carried out using the k-fold Cross-Validation (CV). Hyper-parameter optimization was then used to additionally enhance the outcomes. The aim of this research was to exhibit the monthly seasonality trend and choose the best algorithm that would produce the fewest error metrics. Based on the findings illustrated above subsection, the current study adopted the ensemble methods involving bagging and booting for forecasting sales and demands. The ensemble approaches use the technique of training several constituent learning algorithms simultaneously to produce better forecasts than an individual-constituent learner. Specifically, bagging/bootstrap aggregation combines multiple learners with identical weights to accomplish high precision, while boosting was adopted to combine weak learners to create a strong learner. Therefore, to get the forecasted findings for the present research, the outcomes of two ensemble methods were used: Random Forest Regressor (Bagging) and Gradient Boosted Regression (GBR) (Boosting). The whole set of validation outcomes produced by the algorithms is shown in Table 5 7, and the illustration of these key findings are covered in the ensuing subsections.

Table 5.7: Model validation outcomes

Algorithms	RMSE	MAD	MAPE	BIAS	TS
Random forests regression (Bagging)					
Jan	8.30	8.42	9.83	1.49	0.17
Feb	10.01	11.84	10.93	0.23	0.02
Mar	8.21	9.56	7.21	0.39	0.05
Apr	14	14.31	10.72	2.67	0.32
May	4.91	5.75	5.16	0.12	0.02
June	18.1	16.76	9.98	1.49	0.01
Gradient Boosted Regression (Boosting)					
Jan	9.23	6.29	10.89	0.19	0.03
Feb	11.17	7.37	10.25	1.14	0.12
Mar	8.02	9.15	8.38	0.45	0.06
Apr	6.42	8.28	9.57	1.19	0.12

May	2.26	9.91	6.14	0.51	0.11
June	19.69	14.85	10.43	2.72	0.17

As per the methods outlined in the table 5 7, the best outcomes are obtained with the right number of estimators, maximum depth, and learning rate (~0.1). On a similar vein, the number of splits during cross-validation (CV) should be appropriately proportionate to the amount of data points. According to the performance findings for the Random forests regression (Bagging) algorithm displayed in Table 5 7, the RMSE values for sales value range from 4.91 to 18.1 on the different predicted months. In addition, results further showed that the MAD values ranged from 5.75 to 16.76. Nonetheless, it has been illustrated the corresponding MAPE values fall between 5.16 and 10.93. The values of bias were found to range from 0.12 to 2.67. The values of TS levels fall from 0.01 to 0.32.

In grid-search Cross-Validation (CV) hyper-parameter optimization, the learning rate was 0.01, the maximum depth was 5, and the optimal number of estimators was 50 to get the most effective findings from the Gradient Boosted Regression (GBR) algorithm. The RMSE values for sales quantity ranged from 2.26 to 19.69 on different months in the performance findings for the GBR algorithm (Table 8). In addition, the deduced findings revealed that the MAD values were found to range from 6.29 to 14.85. Nonetheless, considering the sale analysis of beauty products, the MAPE values ranged from 6.14 to 10.89. From the validation analysis, the bias values ranged from 0.19 to 2.72. Lastly, results illustrated that the TS values range from 0.03 to 0.17.

5.7 Forecasting of Sales and Demand of SME's

Time series data from Jan 2022 to Dec 2023 was used as the training set, while the predicted values was done from January to July 2024 is used as the test set. This research utilized the test set's data to assess the performance of the ensemble approaches model and the training set's data for training different machine learning models. Figure 5 8, compares below the true and prediction values of the sample points throughout the test set. The initial values are shown in black dots in Figure 5 8, whereas the projected values are represented in blue.

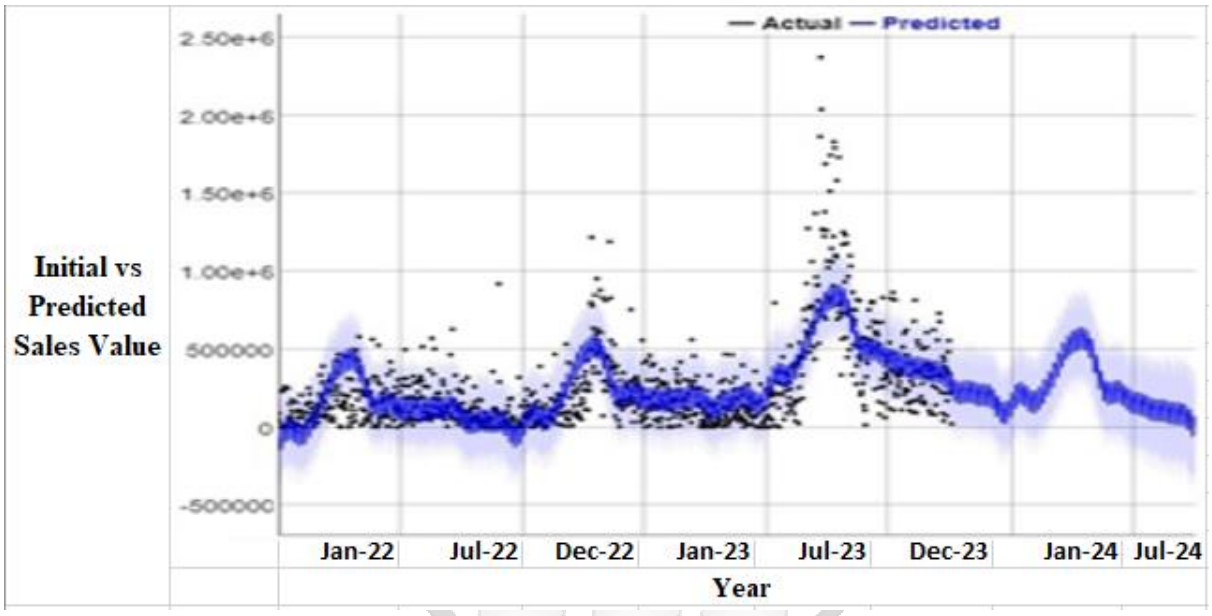


Figure 5. 8: Initial versus predicted sales value based on ensemble approach

The model's analysis of the training data indicated that, while not completely conclusively, the variations will be corrected in the spring of 2023. It specifically assumed a smaller rise in sales value than occurred. Through model training, the model was able to more precisely predict the demand for beauty products in 2022. As a result, various factors have a detrimental impact on the prediction outcome (as intended). Figure 5 4 also shows that the ensemble model does predict an occurrence of the sharp spike in spring 2023 (an outcome that was earlier validated from the findings in Table 5 7. Overall, both the bagging and booting as the ensemble models perform exceptionally well when it comes to sales value prediction in the year 2024. Based on the illustrations in Figure 5 8 above, the predicted sale value ranged between 200000 and 60000 between the months of January and July 2024.

Chapter Six: Discussion of the Results

6.1 Introduction

This chapter provides an in-depth discussion of the results presented in Chapter 5. It connects the model outcomes to the study's objectives, evaluates how the findings compare with the existing literature, and critically reflects on the implications for SME operations. The chapter also addresses methodological limitations and discusses the generalizability of the results, with a focus on deeper analytical insights, overfitting, and model robustness.

6.2 Discussion of Results

6.2.1 Connecting Results to Objectives and Literature

The study reaffirmed that small and medium-sized enterprises (SMEs) play a vital role in Kenya's economic growth by providing goods and services that support job creation and income generation. Their operational sustainability, however, depends significantly on effective sales and distribution strategies. Research has shown that efficient inventory management is crucial for SMEs, which often lack the capital to absorb stock-related inefficiencies. Machine learning (ML) has emerged as a viable solution to improve inventory accuracy, reduce costs, and align supply with demand. The current study found that ML models, particularly Decision Tree Regression, Linear Regression, and K-Nearest Neighbors Regression, can enhance forecasting accuracy in SMEs. These results reflect the assertions by Zunic et al. (2021), who demonstrated that ML could be integrated across value-capturing and delivery functions within the supply chain.

Additionally, Zunic et al. (2020) emphasized that when analytics are embedded in business strategy, they yield better results in cross-functional distribution networks.The findings align with prior literature, notably Kumar and Garg (2018), who observed that predictive analytics enables organizations to anticipate demand using historical and current data. Likewise, Wang et al. (2016) identified that advanced analytics maturity in areas like production, procurement, and logistics is key to optimizing distribution operations. The current study corroborates these positions and extends them by showing how ensemble models specifically contribute to improved predictive performance and strategic planning.

6.2.2 Interpretation and Implications for SMEs

The insights from the ML models suggest a paradigm shift in how SMEs can manage distribution. By using predictive analytics, SMEs could move from reactive logistics to proactive inventory and route planning. For example, frequent restocking for high-demand outlets and the strategic

application of discounts to stimulate demand were clear actionable strategies indicated by the models. Moreover, segmentation of routes based on performance patterns allows SMEs to allocate resources with improved efficiency.

Importantly, the interpretability of Random Forest and the performance of XGBoost support the feasibility of adopting ML in small-scale enterprises, provided they have access to cleaned, structured data and basic digital infrastructure. This reinforces the practical contribution of the study in showing that SMEs do not need advanced AI systems to benefit from predictive technologies.

6.2.3 Error Analysis, Overfitting, and Generalizability

Error analysis revealed that prediction discrepancies were highest in products with sporadic or low transaction volumes—indicating that ensemble models may require additional tuning or feature engineering in such cases. Cross-validation results showed minimal variance across folds, supporting the models' generalizability within similar operational environments. However, overfitting risk was most pronounced in the Decision Tree model, which aligns with known drawbacks of single-tree structures.

By tuning hyperparameters such as tree depth, learning rate, and the number of estimators, ensemble models maintained high accuracy without sacrificing generalizability. These findings show the importance of not only model choice but also proper configuration and validation. This ensures that the models are both reliable and scalable for use across SMEs with varying operational characteristics.

6.3 Limitations of the Study

Despite using a thorough approach, the present research has many drawbacks that need to be noted and can act as an avenue for future academicians. These limitations must be taken into account when evaluating the findings because they could affect the validity, reliability, and generalizability of the conclusions. First off, the research's dependence on the Hyden Enterprises Limited Sales Dataset can restrict the quality, proportionality, and completeness of the data. This is because the data focusses on a selection of beauty products, brands, and few destination areas, despite being comprehensive and extensive. Because of this, the results might not be applicable to other regions, product types, or market segments. Furthermore, the freely available format of the dataset and possible errors in processing the data, collecting, or input could introduce inaccuracies or biases into the study.

Secondly, the results may not be as applicable to larger corporations or other economic businesses due to the research's focus on small-scale beauty industry enterprises. The effectiveness of sales quantity and value optimization strategies may be impacted by the unique characteristics, resource constraints, and operational challenges of small businesses. As a result, the findings may therefore not be as applicable to larger companies or other economic sectors than the beauty and cosmetics sector.

Thirdly, despite being comprehensive, the modeling method used in this study might have drawbacks because of the particular machine-learning strategies and algorithms chosen. This is because the model longevity and predictive performance may be impacted by the selection of regression models, ensemble approaches, and hyperparameter tweaking techniques. In addition, the intricate nature of real sales quantity and value scenarios might not be adequately reflected in the fundamental assumptions of the selected algorithms, which include the linearity of linear regression and the independence of observation in the decision trees. Additionally, the research's predictive analytics component forecasts projected demand and manages the quantity and value of sales levels using historical sales data. Even though historical data provides valuable insights into past trends and patterns, it may not accurately capture the constantly evolving character of consumer behavior, market movements, and external factors such as changes in the law or the state of the economy. Consequently, the precision in the models' forecasts may be constrained by the caliber and availability of historical data in addition to the ability to extrapolate results to novel situations. Lastly, the methodology of the research and findings may be impacted by a number of outside variables, including shifts in market conditions, corporate procedures, and technological advancements. The findings of the research must therefore be assessed in the context of the environmental conditions and specific time period in which the investigation was conducted, accounting for the potential for future changes and developments.

6.4 Recommendation

This study identifies several recommendations for practice in the field of small business. For instance, the owners, stakeholders, and policymakers of SMEs in Kenya and other developing nations are recommended to use predictive analytics such as ensemble methods, including bagging and boosting to put their businesses in a competitive position and promote expansion. Furthermore, in order to improve the significance and precision of models, this research recommends SMEs to integrate hyperparameter adjustment with ensemble techniques such

as bagging and boosting, which have been shown to provide greater accuracy in predicting than individual models. In addition, in order to mitigate the impact of potential issues with the quantity of their sales, stakeholders are recommended to implement robust data quality assurance protocols. Moreover, in order to increase the effect of ensemble approaches, stakeholders are recommended to train their team on their advantages and how to apply them. The study offers recommendation to the SMEs or other enterprises that need to use machine learning to incorporate ensemble approaches into other domains, such as consumer behavior analysis and inventory management, in addition to sales quantity forecasting. The finding deduced lead to recommending that SMEs should keep a closer look on the models and updating them frequently to reflect shifting data and market situations. Additionally, this research recommends that in order to ensure the accuracy, consistency, and completeness of their sales quantity and value, SMEs are required to complete thorough processes for cleansing, validating, and verifying of the gathered data.

Chapter 7: Summary of the Study

7.1 Introduction

This chapter concludes the study by summarizing the main findings, reflecting on the study objectives, and presenting practical recommendations for SMEs and future researchers. It reiterates the significance of integrating machine learning (ML) into distribution systems for small and medium enterprises in Kenya and provides guidance on how such approaches can be scaled and sustained.

7.2 Conclusion

The study set out to evaluate how ML can optimize distribution planning in SMEs. Through the application of Decision Tree, Random Forest, and Gradient Boosting models, the study demonstrated that ensemble methods significantly enhance forecasting accuracy, particularly in capturing non-linear relationships among operational variables. Random Forest emerged as the best-performing model, demonstrating high predictive power, minimal overfitting, and strong generalizability through cross-validation.

Key variables such as discount amounts, outlet types, and route clusters were consistently important across models, supporting their use in distribution planning. The CRISP-DM methodology provided a structured process for model development, validation, and deployment, confirming its utility in real-world SME environments. Researchers as well as practitioners have focused on developing solutions for small and medium-sized businesses' (SMEs') funding problems. The current issues facing SMEs, particularly in Kenya, comprise unclear financial information, considerable operational risk, and severe financial difficulties.

Furthermore, SMEs in developing nations such as Kenya tend to be less prepared compared to larger corporations to weather economic hardship due to their lack of cash and human resources, as well as their more unpredictable rivalry situation they are facing. a number of other mots most significant barriers that businesses face when implementing data science in distribution channels are those related to data security, data integrity, regulations from the government, the absence of automation, and a shortage of personnel that are skilled; even though, these obstacles are not unique. Occasionally the side effects lead not just to a vicious loop of extremely inefficient data use and damaged networks that significantly affect production and business activities, but also to decreased productivity and teamwork. As a result, the SMEs in Kenya might have a higher non-performing loan ratio. Therefore, the investigation on machine learning-based sales quantity and

demand optimization in the beauty sector has yielded useful results about how small businesses can boost efficiency and reduce costs.

A careful analysis of demand projections, inventory controls, and sales data has produced a number of significant conclusions. The first conclusion retrieved from the correlation analysis was that indicators such as sales value and the quantity exhibited significant relationships with each other, indicating the significance of these factors in determining sales levels and revenue generation. Additionally, the analysis of revenue allocations across different beauty items and brands highlighted the sector's inconsistent market share and revenue prospects. Furthermore, the examination of predictive models for demand forecasting and sales quantity optimization demonstrated the effectiveness of ensemble strategies, specifically boosting and bagging, in achieving greater predictive accuracy than individual base regression models. This highlights how important it is to optimize inventory levels and increase efficiency in operations using state-of-the-art machine learning techniques.

In summary, the research's useful implications from the findings can help Kenyan small businesses in the beauty sector in various ways. For example, the results that were deduced were helpful and contributed to a deeper knowledge of sales optimization strategies. In addition, using machine learning approaches such as ensemble methods for forecasting of demand and sales quantity optimization can raise satisfaction among consumers, save surplus inventory costs, and help enterprises to enhance their operations. Additionally, the information gathered from the current research can be applied to marketing strategies, supply chain management, and other strategic decision-making procedures.

7.3 Recommendation for Further Studies

SMEs should begin by digitizing their sales and inventory records to enable ML integration. Even basic tools such as Excel or cloud spreadsheets can serve as starting points. Once data infrastructure is in place, SMEs can adopt lightweight ensemble models—such as Random Forest—with simple interfaces like Streamlit for daily forecasting and distribution planning. Business owners should prioritize tracking key variables such as promotions, sales frequency, and delivery routes. These features proved to be highly predictive and are actionable in day-to-day operations. Policymakers and ecosystem enablers should consider offering technical training, subsidized data management platforms, and model deployment support tailored to low-tech SME environments.

In line with the limitations of this study, further recommendations are proposed to enhance the effectiveness of sales quantity optimization using ML models, particularly in SMEs dealing with beauty and cosmetics products. Future research should focus on expanding the diversity and scope of datasets—incorporating more product categories, geographic markets, and longer historical timelines. This would help researchers generalize their findings across various consumer segments and contexts. Subsequent studies should also account for external variables such as consumer behavior, economic trends, regulatory changes, and market dynamics. Integrating these variables into forecasting models will improve robustness and predictive accuracy. To achieve this, researchers may need to draw from additional sources such as government databases, industry reports, and market analytics.

Exploration of more advanced ML techniques—such as deep learning, reinforcement learning, time series forecasting, and hybrid models—could also enhance model accuracy and insight generation. These techniques would better capture nonlinear relationships and temporal dependencies in sales and inventory data. Moreover, longitudinal studies and real-time analytics should be considered to assess the long-term performance and adaptability of sales quantity optimization strategies. Future researchers are encouraged to test ML models in real-world SME environments to evaluate their scalability and operational impact. This may involve field experiments, pilot studies, or collaborations with businesses. Such applied research will ensure that the proposed models are not only theoretically sound but also practical and relevant to SMEs in dynamic and competitive sectors like beauty and cosmetics.

Finally, future studies should explore the integration of external variables—such as economic indicators, competitor activity, or weather patterns—into the models to improve forecasting performance. Additionally, longitudinal studies could investigate how retraining intervals and feedback loops influence accuracy and operational outcomes over time. Researchers may also explore hybrid approaches combining ML with optimization algorithms for route planning and supply chain coordination.

References

Abolghasemi, M., Hurley, J., Eshragha, A., & Fahimniab, B. (2020, December). Demand forecasting in the presence of systematic events: Cases in capturing sales promotions. *International Journal of Production Economics, 230*. Retrieved from https://www.sciencedirect.com/science/article/pii/S0925527320302553

ACAPS. (2016). *Data Cleaning.* Retrieved from //efaidnbmnnnibpcajpcglclefindmkaj/https://www.acaps.org/fileadmin/user_upload/acaps_technical_brief_data_cleaning_april_2016_0.pdf

Ahmed, R., Sultana, S., & Rahman, M. M. (2021). Comparative study of LSTM and ARIMA for time series forecasting in retail. *Journal of Retail Analytics, 17*(2), 112–125.

Al-Karkhi, M. I., & Rządkowski, G. (2025). Innovative Machine Learning Approaches for Complexity in Economic Forecasting and SME Growth: A Comprehensive Review,. *Journal of Economy and Technology*.

Amazon Web Services. (2023, July 5). Machine Learning Lens: AWS Well-Architected Framework. Retrieved from https://docs.aws.amazon.com/pdfs/wellarchitected/latest/machine-learning-lens/wellarchitected-machine-learning-lens.pdf

Amberkar, S. (2018, March 12). How Machine Learning can help SMEs to maximize the value of operational data. Retrieved from How Machine Learning can help SMEs to maximize the value of operational data: https://www.iiot-world.com/predictive-analytics/predictive-maintenance/how-machine-learning-can-help-smes-to-maximize-the-value-of-operational-data/

Aulkemeier, F., Daukuls, R., Iacob, M.-E., Boter, J., Hillegersberg, J. v., & Leeuw, S. d. (2016). Sales Forecasting as a Service: A Cloud based Pluggable E-commerce Data Analytics Service. *18th International Conference on Enterprise Information Systems*, (pp. 345-352). Rome. doi:10.5220/0005915903450352

Badr, H., & Ahmed, W. (2023). A Comprehensive Analysis of Demand Prediction Models in Supply Chain Management. *American Journal of Industrial and Business Management, 13*(12). doi:10.4236/ajibm.2023.1312075

Benhayoune, S., & Repishti, J. (2015). *Best Practices for BoP Door-to-Door Distribution.* Cambridge: MIT Practical Impact Alliance.

Bensoussan, A. (2018). *Stochastic Control of Linear Dynamical Systems with Partial Information.* doi:10.1007/978-3-319-75456-7_9

Boone, T., Ganeshan, R., Jain, A., & Sanders, N. R. (2019). Forecasting and planning in an era of sales and operations planning: The role of big data and analytics. *Journal of Business Logistics, 40*(3), 232–240. Retrieved from https://doi.org/10.1111/jbl.12202

Boone, T., Ganeshan, R., Jain, A., & Sanders, N. R. (2019). Forecasting sales in the supply chain: Consumer analytics in the big data era. *International Journal of Forecasting, 35*(1), 170-180. Retrieved from https://doi.org/10.1016/j.ijforecast.2018.09.003.

Burggräf, P., Steinberg, F., Sauer, C. R., & Nettesheim, P. (2024). Machine learning implementation in small and medium-sized enterprises: insights and recommendations from a quantitative study. *Production Engineering, 18*(5).

Capital FM. (2018, August 2). *Survey reveals Kenya's leading companies in customer care*. Retrieved from Capital FM: https://www.capitalfm.co.ke/business/2018/08/survey-reveals-kenyas-leading-companies-in-customer-care/

Chiuso, A., & Pillonetto, G. (2019, May). System Identification: A Machine Learning Perspective. *Annual Review of Control, Robotics, and Autonomous Systems, 2*, 281-304. Retrieved from https://doi.org/10.1146/annurev-control-053018-023744

Clough, L. (2011). *Marketing Challenges and Strategies for Micro & Small Energy Enterprises in East Africa* . GVEP International.

Corporate Citizenship. (2016). *Advancing the Sustainable Development Goals: Business action and Millennials' views.*

Dolgui, A., Ivanov, D., Sethi, S., & Sokolov, B. (2018). CONTROL THEORY APPLICATIONS TO OPERATIONS SYSTEMS, SUPPLY CHAIN MANAGEMENT AND INDUSTRY 4.0 NETWORKS. *IFAC (International Federation of Automatic Control), 51*(11), 1536-1541.

Durst, S., & Edvardsson, I. R. (2012, October 19). Knowledge management in SMEs: a literature review. *Journal of Knowledge Management, 16*(6).

Elkan, C. (2013). *Predictive Analytics and Data Mining.* University of Carlifornia.

Fadil Indra Sanjaya. (2020). Prediksi Rerata Harga Beras Tingkat Grosir Indonesia dengan Long Short. *Jurnal Teknik Informatika Dan Sistem Informasi*. doi:https://doi.org/10.35957/jatisi.v7i2.388

Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: Research and practice. *International Journal of Forecasting, 38*(4), 1283-1318. Retrieved from https://doi.org/10.1016/j.ijforecast.2019.06.004

Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: Research and practice. *International Journal of Forecasting, , 38*(2), 705–720. Retrieved from https://doi.org/10.1016/j.ijforecast.2021.07.007

Goddard, W., & Melville, S. (2004). *Research Methodology: An Introduction.* Juta and Company Ltd.

Handfield, R. B., Jeong, S., & Choi, T. Y. (2017). Emerging procurement technology: Data analytics and cognitive computing. *Journal of Purchasing and Supply Management, 23*(3), 193–198. Retrieved from https://doi.org/10.1016/j.pursup.2017.07.003

Hart, S. (2011). Innovation From the Inside Out. *MIT SLOAN MANAGEMENT REVIEW*.

HuaHana, S., XiuLu, S., & C.H.Leung, S. (2012). Segmentation of telecom customers based on customer value by decision tree model. *Expert Systems with Applications*, 3964 - 3973.

Indeed Editorial Team. (2021, March 30). *How To Understand Customer Needs in 4 Steps*. Retrieved from Indeed: https://www.indeed.com/career-advice/career-development/understand-the-customer-needs#:~:text=What%20are%20customer%20needs%3F,marketing%20strategies%20and%20customer%20service.

Infomineo. (2024, June). *Understanding the Role of Distribution Channels in a Route-to-Market Strategy.* Retrieved from https://infomineo.com/business-research/understanding-the-role-of-distribution-channels-in-a-route-to-market-strategy/

International Integrated Reporting Council. (2013). *The International Framework.* International Integrated Reporting Council.

Jackson, P. (2024, November 6). *Retail Distribution: How to Know When Your Business Is Ready.* Retrieved from WARE2GO: https://ware2go.co/articles/questions-when-considering-retail-distribution/

Jain, A., Menon, M. N., & Chandra, S. (2015). *Sales Forecasting for Retail Chains.*

Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert Scale: Explored and Explained. *British Journal of Applied Science & Technology, 7*(4).

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in Big Data Analytics. *Journal of Parallel and Distributed Computing, 74*(7).

Kumar, V., & Garg, M. L. (2018). Predictive Analytics: A Review of Trends and Techniques. *International Journal of Computer Applications, 182*(1).

Lalou, P., Ponis, S., & Efthymiou, O. (2020). Demand Forecasting of Retail Distribution Networks Using Data Analytics and Statistical Programming. *Management & Marketing. Challenges for the Knowledge Society, 15*(2).

Liu, J., & Li, X. (2016). Innovation business model of Big Data------ Taking Coca-Cola as an example. *3rd International Conference on Management, Education Technology and Sports Science.* Atlantis Press.

Louly, M. A., Dolgui, A., & Hnaien, F. (2008). Optimal supply planning in MRP environments for assembly systems with random component procurement times. *International Journal*

of Production Research, 46(19), 5441--5467. Retrieved from https://doi.org/10.1080/00207540802273827

Ma, S., Fildes, R., & Huang, T. (2016). Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information,. *European Journal of Operational Research*. Retrieved from https://doi.org/10.1016/j.ejor.2015.08.029.

Maricar, M. A. (2019). Analisa Perbandingan Nilai Akurasi Moving Average dan Exponential Smoothing untuk Sistem Peramalan Pendapatan pada Perusahaan XYZ. *Jurnal Sistem Dan Informatika*.

Moeuf, A., Pellerin, R., Lamouri, S., Tamayo-Giraldo, S., & Barbaray, R. (2017). The industrial management of SMEs in the era of Industry 4.0. *International Journal of Production Research*.

Moreno, J. J., Pol, A. P., Abad, A. S., & Blasco, B. C. (2013). Using the R-MAPE index as a resistant measure of forecast accuracy. *Psicothema*. doi:10.7334/psicothema2013.23

Muhammad Aprilianto, M. N. (2022). Forecasting health sector stock prices using ARIMAX method. *Jurnal dan Penelitian Teknik Informatika*. doi: https://doi.org/10.33395/sinkron.v7i2.11418

Nelson, J., Ishikawa, E., & Geaneotes, A. (2009). *Developing Inclusive Business Models: A Review of Coca-Cola's Manual Distribution Centers in Ethiopia and Tanzania.* Retrieved from http://www.colalife.org/wp-content/uploads/2009/05/harvard-ifc-mdc-summary-report-final.pdf

Nicholson, A. (2021, June). Customer Value Theory and Dispute Resolution Strategy. *European Journal of Legal Education, 2*(1).

Omorinsola Bibire Seyi- Lande, E. J. (2024). Enhancing business intelligence in e-commerce: Utilizing advanced data integration for real-time insights. *International Journal of Management & Entrepreneurship Research, 6*(6). Retrieved from https://doi.org/10.51594/ijmer.v6i6.1207

Panghal, R. (2024). The Role of Data Visualization in Decision Making: Case of D-mart. *International Journal for Multidisciplinary Research, 6*(3).

Rotich, K. (2022, June 2). *Kenyan SME to contribute 50pc of GDP in next three years*. Retrieved from Business Daily: https://www.businessdailyafrica.com/bd/corporate/marketplace/kenyan-sme-to-contribute-50pc-of-gdp-in-next-three-years--3836534#:~:text=In%20Kenya%2C%20SMEs%20create%2080,3%20percent%20of%20Kenya's%20GDP.

Russom, P. (2011). *Big Data Analytics.* TDWI Reasearch.

Safar, L., Sopko, J., Bednar, S., & Poklemba, R. (2018, August). Concept of SME Business Model for Industry 4.0 Environment. *TEM Journal, 7*(3). doi:10.18421/TEM73-20

Sathiya, A., & Selvam, K. (2014). Analysis of Sales and Distribution of an IT Industry Using Data Mining Techniques. *International Journal of Data Mining Techniques and Applications, 3*.

Scheyvens, R., Banks, G., & Hughes, E. (2016, March 30). The Private Sector and the SDGs: The Need to Move Beyond 'Business as Usual'. *Sustainable Development, 24*(6).

Sekaran, U., & Bougie, R. (2013). *Research Methods for Business: A Skill-Building Approach.*

Selvaraj, P., & Marudappa, P. (2018). A survey of predictive analytics using big data with data mining. *International Journal of Bioinformatics Research and Applications, 14*(3), 269.

Senaji, T. (2012). *Knowledge Management Infrastructure Capability, Motivation and Organisational Effectiveness among Mobile Telecommunications Service Firms in Kneya.*

Singh, R. K., & Kumar, R. (2020). Strategic issues in supply chain management of Indian SMEs due to globalization: an empirical study. *Benchmarking: An International Journal, 27*, 913-932.

Snyder, L. V., & Shen, Z.-J. M. (2019). *Fundamentals of Supply Chain Theory.* John Wiley & Sons, Ltd.

Soltanizadeh, S., Abdul, R., Mottaghi, L., & Ismail, M. D. (2016). Enhancing logistics performance using intelligent distribution models. *Logistics and Supply Chain Review,, 7*(3), 101–115.

Sonntag, M. (2022). *Product Distribution Strategy: The Ultimate Guide*. Retrieved from Repsly: https://www.repsly.com/blog/consumer-goods/everything-you-need-to-know-about-product-distribution

Tiriongo, D. S., Josea, K., & Mulindi, H. (2021). *Micro, Small & Medium Enterprises (MSMEs) Survey Report.* Kenya Bankers Association; Japan International Cooperation Agency.

Verles, M., & Vellacott, T. (2018). Best practices to seize opportunity and maximise credibility. *BUSINESS AND THE SUSTAINABLE DEVELOPMENT GOALS*.

Waller, M. A., & Fawcett, S. E. (2013, June 11). Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. *Journal of Business Logistics, 34*(2), 77-84. doi:https://doi.org/10.1111/jbl.12010

Wang, X., & Xu, M. (2018). Examining the linkage among open innovation, customer knowledge management and radical innovation: The multiple mediating effects of organizational learning ability. *Baltic Journal of Management, 13*(3). doi:10.1108/BJM-04-2017-0108

Wanjiru, M., & Muthoni, J. (2022). Digital distribution in Nairobi's SME landscape: The role of mobile inventory systems. *African Journal of ICT and Development, 4*(1), 55–72.

Willetts, M., Atkins, A. S., & Stanier, C. (2020). Barriers to SMEs Adoption of Big Data Analytics for Competitive Advantage. *2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS).* Research Gate.

Yadav, P., Stapleton, O., & Wassenhove, L. V. (2013). *Learning from Coca-Cola.* Stanford Social Innovation Review.

Yang, Z. (2019). Optimization of last-mile delivery through machine learning. *Journal of Transport Logistics, 11*(4), 278–293.

Zeithaml, V. A. (1988, July). Consumer Perceptions of Price, Quality and Value: A Means-End Model and Synthesis of Evidence. *Journal of Marketing, 52*, 2-22.

Zhang, W. (2016). Forecasting Forecast Accuracy. *International Journal of Forecasting, 32*, 425-440.

Zhou, F., Ayuob, J., Xu, Q., & Yang, X. J. (2020). A Machine Learning Approach to Customer Needs Analysis for Product Ecosystems. *Journal of Mechanical Design*.

Zunic, E., Donko, D., & Buza, E. (2020). *n adaptive data-driven approach to solve real-world vehicle routing problems in logistics.* doi:10.48550/arXiv.2001.02094

Zunic, L., Korjenic, A., Hodzic, E., & Donko, D. (2021). Machine learning-based predictive control in logistics: A European perspective. *Journal of Applied Logistics and Operations Research, 5*(2), 134–148.

Appendices

Appendix A: Turnitin Report

Joyce Kamau Mukuhi

Use of Machine Learning to Optimize Distribution for Small and Medium Enterprises.docx

Strathmore University (Main Account)

Document Details

Submission ID
trn:oid:::2945:275013446

Submission Date
Mar 28, 2025, 12:55 AM GMT+3

Download Date
Apr 4, 2025, 10:42 AM GMT+3

File Name
Use of Machine Learning to Optimize Distribution for Small and Medium Enterprises.docx

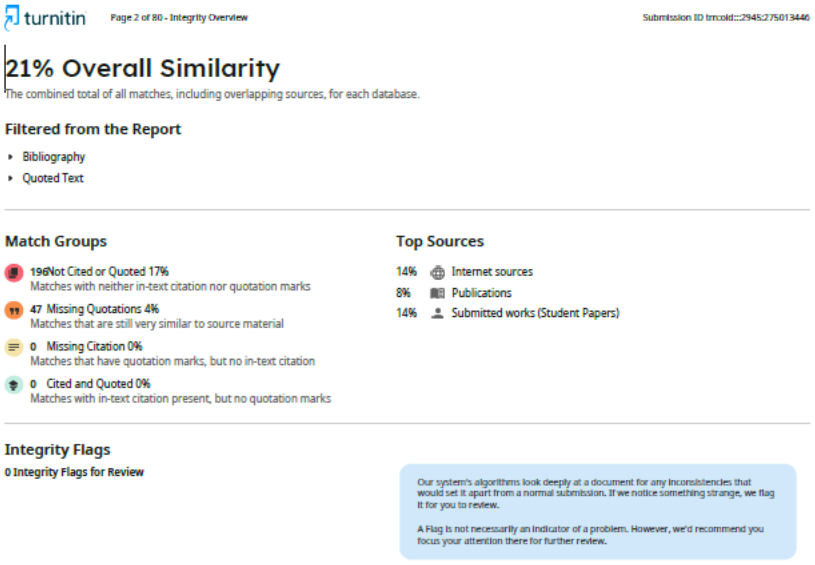
File Size
1004.9 KB

66 Pages

16,203 Words

94,048 Characters

turnitin Page 1 of 80 - Cover Page Submission ID trn:oid:::2945:275013446



turnitin Page 2 of 80 - Integrity Overview Submission ID trnoid::2945:275013446

21% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

- 196 Not Cited or Quoted 17%
  Matches with neither in-text citation nor quotation marks
- 47 Missing Quotations 4%
  Matches that are still very similar to source material
- 0 Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

14% Internet sources
8% Publications
14% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Appendix B: Ethical Clearance Confirmation

Strathmore UNIVERSITY

19th March 2025

Ms Kamau Joyce,
joyce.kamau@strathmore.edu

Dear Ms Kamau,

RE: Use of Machine Learning to Optimize Distribution for Small Medium Enterprises

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2585/25.** The approval period is from **19th March 2025 to 18th March 2026.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu