BINDURA UNIVERSITY OF SCIENCE EDUCATION

FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE

APPLICATION OF RANDOM FOREST MACHINE LEARNING ALGORITHM IN RETAIL FORECASTING.

MUDOTI REVELATION

REG NUMBER: B192345B

SUPERVISOR: MR CHAITEZVI

A RESEARCH PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE BACHELOR OF SCIENCE HONOURS DEGREE IN INFORMATION TECHNOLOGY (SOFTWARE ENGINEERING)

2024

Approval Form

The undersigned certify that they have supervised the student Revelation Mudoti in the research dissertation entitled, “Application of Random Forest Machine Learning Algorithm in Retail Forecasting.” submitted in partial fulfillment of the requirements for a Bachelor Of Science Honors Degree in Information Technology (Software Engineering) at Bindura University of Science Education.

STUDENT: **DATE:**

R Mudoti **03/10/24**

…… ……

SUPERVISOR: **DATE**

…… 03/10/24 ……

CHAIRPERSON: **DATE:**

03/10/24

…… ……..

\

Dedication

This project is dedicated to my parents, Mr and Mrs. Mudoti for their unwavering support, prayers, sacrifices, encouragement, and guidance throughout my academic journey.

Acknowledgments

I would like to express my deepest gratitude and appreciation to all those who have contributed to the successful completion of my final year project. Without their support, guidance, and encouragement, this project would not have been possible. I would like to thank my project supervisor, Mr. Chaitezvi, for his invaluable guidance and expertise throughout the entire duration of this project. I am immensely thankful to my family for their unconditional love, unwavering support, and understanding throughout this project.

Abstract

This research project investigates the application of the Random Forest machine learning algorithm to improve demand forecasting for small-scale grocery retailers in Zimbabwe. Small-scale retailers face significant challenges in accurately predicting product demand, often leading to stock-outs, excess inventory, and operational inefficiencies. Traditional forecasting methods, such as statistical models or expert judgment, often fail to capture the complex and dynamic nature of demand in developing country contexts. This study explores the potential of the Random Forest algorithm to address these challenges. A machine learning model was developed and trained using historical sales data from a sample of small-scale retailers. The performance of the model was rigorously evaluated using various metrics, including Mean Absolute Error, Root Mean Squared Error, and R-squared. The results demonstrate the effectiveness of the Random Forest model in predicting future demand with greater accuracy than traditional methods. This research highlights the potential of machine learning to empower small-scale retailers in Zimbabwe by providing them with more informed decision-making tools for inventory management, pricing strategies, and overall supply chain optimization. The study concludes with recommendations for the adoption and further development of machine learning-based forecasting solutions for the retail sector in Zimbabwe.

Table of contents

List of Figures

Chapter 1: Problem Identification

1.0 Introduction

Retail forecasting plays a crucial role in supply chain management (SCM) for businesses that acquire, transform, and deliver goods to consumers. Ensuring precise and prompt demand forecasts is vital for optimizing inventory, cutting costs, enhancing customer service, and planning for future growth. However, forecasting in retail, particularly for small and medium-sized enterprises (SMEs), is fraught with challenges due to volatile and unpredictable environments, limited data, and complex customer behavior.

Traditional methods, like statistical models or expert judgment, often fall short in capturing the nonlinear and complex relationships between demand and its influencing factors. To overcome these challenges, more advanced and robust forecasting methods are needed. One promising approach is machine learning, a subset of artificial intelligence that enables computers to learn from data and improve performance autonomously.

The focus of this project is on applying the random forest machine learning algorithm to retail forecasting, specifically for small-scale retailers. In the retail context, the challenges include economic uncertainties, fluctuating consumer behaviors, and limited resources. These factors contribute to difficulties in accurately anticipating product demand, resulting in issues like stock-outs, excess inventory, and operational inefficiencies.

1.1 Background

Supply chain management (SCM) is essential for businesses that acquire, convert, and deliver goods to customers (Christopher, 2016). It relies heavily on accurate and timely demand forecasting, which involves predicting future customer demand based on available data and other factors (AltexSoft, 2022). Effective demand forecasting helps businesses optimize inventory levels, reduce costs, enhance customer service, and plan for future growth (MobiDev, 2021).

However, demand forecasting presents significant challenges, particularly for small and medium enterprises (SMEs) operating in dynamic and uncertain environments. SMEs often grapple with limited data availability, high variability in demand patterns, complex customer behavior, and competition from larger players (Springer, 2022).

Statistical models or expert judgment, may fall short in capturing the nonlinear and intricate relationships between demand and various influencing factors (Peak, 2021). Thus, there is a need for more advanced and robust forecasting methods that can manage the complexity and uncertainty inherent in SCM problems. One such method is machine learning, a branch of artificial intelligence that allows computers to learn from data and improve their performance without explicit programming (DemandCaster, 2023). Machine learning offers powerful and flexible tools for demand forecasting, including neural networks, support vector machines, and random forests (AltexSoft, 2022).

Random forest, a machine learning algorithm, combines multiple decision trees to create an ensemble model capable of handling both regression and classification problems (AltexSoft, 2022). This algorithm boasts several advantages over other machine learning methods, such as high accuracy, low bias, high robustness, low sensitivity to outliers and noise, and the ability to handle large and high-dimensional data sets (DemandCaster, 2023). Additionally, random forests can provide measures of variable importance, helping to identify the key factors that influence demand (AltexSoft, 2022).

In Zimbabwe, small-scale retailers are vital to the economy, serving as key contributors to local communities and providing essential goods and services (StartupBiz, 2022). However, these retailers often struggle with effective supply chain management, particularly in demand forecasting. A report by Techzim (2022) indicates that 43% of micro to medium businesses in Zimbabwe are in the retail and wholesale sector, with only 5% in ICT and manufacturing. This highlights a low level of digital technology and innovation adoption among these businesses.

Small-scale retailers in Zimbabwe encounter unique dynamics, including economic uncertainties, fluctuating consumer behaviors, and limited resources. These factors contribute to difficulties in anticipating product demand accurately, leading to challenges such as stock outs, excess inventory, and operational inefficiencies (ZimPlaza, n.d.). Leveraging advanced technologies and methodologies, such as machine learning algorithms, for demand forecasting could significantly benefit these retailers by providing more precise insights into consumer demand patterns (MobiDev, 2021). This project was aimed at applying random forest machine learning algorithm to the problem of demand forecasting for Zimbabwean small scale retailers. We focus on the grocery sector, which is one of the most important and challenging sectors for SCM in Zimbabwe (StartupBiz, 2022). We collect and analyze historical sales data from a sample of small scale retailers in Zimbabwe, and use random forest to build and evaluate demand forecasting models.

We also compare the performance of random forest with other machine learning and traditional methods, and explore the factors that influence demand in the grocery sector. We hope that our project can provide useful insights and recommendations for improving SCM and demand forecasting for Zimbabwean small scale retailers.

1.2 Problem Statement

The existing challenges faced by small-scale retailers in Zimbabwe regarding demand forecasting contribute to inefficiencies in their supply chain operations. Inaccurate predictions lead to sub-optimal inventory management, affecting the retailers' ability to meet customer demand effectively. Consequently, this may result in revenue loss due to stock-outs or increased holding costs due to excess inventory. The lack of tailored and efficient demand forecasting strategies for Zimbabwean small-scale retailers hinders their competitiveness and sustainability in the marketplace. To address this issue, there is a need for a comprehensive understanding of the specific challenges faced by these retailers and the development of targeted solutions to enhance their demand forecasting capabilities.

1.3 Research Aim

The primary aim of this research is to improve demand forecasting for small-scale retailers in Zimbabwe through the application of Random Forest machine learning algorithm. By leveraging advanced predictive models, the research aims to enhance the accuracy and efficiency of demand forecasting processes, ultimately contributing to the operational effectiveness and competitiveness of these retailers.

1.4 Research Objectives

- To analyse and explore the Random Forest machine learning algorithm used for demand forecasting.
- To design, develop and implement a Random Forest Algorithm based demand forecasting model.
- To evaluate the effectiveness of the random forest machine learning algorithm in predicting future demand using historical data.

1.5 Research Questions

- How can the Random Forest machine learning algorithms used for demand forecasting?
- How to design, develop and implement a demand forecasting model based on Random Forest Algorithm.
- How to analyze the effectiveness of random forest machine learning algorithm in predicting future demand using historical data?

1.6 Significance of the Study

The study has potential to empower small-scale retailers in Zimbabwe with a tailored demand forecasting solution. The implementation of an effective machine learning-based model can lead to improved inventory management, reduced operational costs, and increased customer satisfaction. Ultimately, this research contributes to the resilience and competitiveness of small-scale retailers in the Zimbabwean marketplace.

Firstly, it addresses a gap in the literature on the application of machine learning algorithms, especially random forest, to demand forecasting for small-scale retailers in Zimbabwe. Most of the existing studies on demand forecasting focus on large-scale retailers or other industries, and do not consider the specific challenges and opportunities faced by small-scale retailers in developing countries (Springer, 2022). Therefore, this study provides a novel and relevant contribution to the field of demand forecasting and retail management.

In addition, this study offers practical benefits for small-scale retailers in Zimbabwe, who are an important segment of the economy and society. By using a machine learning-based model to forecast demand, these retailers can improve their inventory management, reduce their operational costs, and increase their customer satisfaction. For example, they can avoid stock outs and overstocking, which can negatively affect their sales and profitability. They can also optimize their ordering and replenishment processes, which can save them time and money. Moreover, they can enhance their customer service by meeting their customers' needs and expectations more effectively. These benefits can help them to survive and thrive in a competitive and uncertain market environment.

Moreover, this study has broader implications for the development and sustainability of the retail sector and the economy in Zimbabwe. By improving the performance and efficiency of small-scale retailers, this study can support the growth and innovation of the retail sector, which is a key driver of economic activity and employment in Zimbabwe (StartupBiz, 2022).

Furthermore, by promoting the adoption and utilization of machine learning technologies, this study can foster the digital transformation and modernization of the retail sector and the economy in Zimbabwe, which can enhance their competitiveness and resilience in the global market. Additionally, by enabling small-scale retailers to optimize their inventory and reduce their waste, this study can contribute to the environmental sustainability and social responsibility of the retail sector and the economy in Zimbabwe.

1.7 Scope of the Study

This research focuses specifically on demand forecasting for small-scale retailers in Zimbabwe. The scope includes understanding current practices, identifying influencing factors, developing a machine learning-based model, and evaluating its effectiveness. The study does not encompass broader supply chain management aspects but aims to address the specific challenges related to demand forecasting.

1.8 Limitations

The limitations of this study include potential constraints in data availability and the generalization of findings. The machine learning model's effectiveness may also be impacted by the quality and quantity of historical data accessible for training purposes.

1.9 Definition of terms

This section clarifies the meaning of key terms used in this research project, ensuring consistent understanding and avoiding ambiguity.

Artificial Intelligence (AI): The capacity of a computer or machine to execute tasks that usually require human intelligence, such as learning, problem-solving, and decision-making.

Machine Learning (ML): A branch of AI that enables computer systems to learn from data without needing explicit programming. It focuses on creating algorithms that can enhance their performance through experience.

Random Forest: A machine learning algorithm that uses an ensemble of decision trees to make predictions. It is recognized for its robustness and accuracy in managing complex datasets.

Demand Forecasting: The practice of predicting future demand for products or services by analyzing historical data, market trends, and other relevant factors.

Supply Chain Management (SCM): The coordinated management of all activities involved in sourcing raw materials, transforming them into finished products, and delivering them to customers, aiming to optimize the flow of goods and information throughout the supply chain.

Small-Scale Retailers: Retail businesses with limited resources and smaller operations compared to large chains, typically serving local markets and communities.

Inventory Management: The process of planning, controlling, and optimizing the levels of inventory held by a business.

Stock-outs: Situations where a retailer runs out of a particular product due to insufficient inventory.

Excess Inventory: The condition of having more inventory than needed, resulting in increased storage costs and potential obsolescence.

Data Mining: The practice of extracting meaningful patterns and insights from large datasets using statistical and computational techniques.

Data Preprocessing: The process of cleaning and transforming raw data into a format suitable for analysis or machine learning.

Feature Selection: The method of choosing the most relevant features (variables) from a dataset for use in a machine learning model.

Hyper-parameters: Parameters that govern the learning process of a machine learning model, set before the model is trained.

Accuracy: The capability of a machine learning model to predict the correct outcome.

Robustness: The ability of a model to maintain its accuracy despite the presence of noise or outliers in the data.

Generalizability: The capacity of a model to perform well on new, unseen data.

1.10 Chapter Summary

This introduced the background of the study, highlighted the problem faced by small-scale retailers in Zimbabwe regarding demand forecasting, and outlined the research aim, objectives, significance, scope, and limitations. The subsequent chapters will delve into a detailed examination of existing demand forecasting practices, the identification of key influencing factors, the development and implementation of the machine learning-based model, and an evaluation of its effectiveness.

Chapter 2: Literature Review

2.1 Introduction

This chapter reviews existing literature related to demand forecasting, particularly in the context of small-scale retailers. The literature encompasses a wide range of topics, including traditional demand forecasting methods, challenges faced by small-scale retailers, and the application of machine learning algorithms in enhancing forecasting accuracy.

2.2 Traditional Demand Forecasting Methods

Businesses have traditionally used demand forecasting methods to predict future demand by analyzing historical data, expert judgment and market trends. These methods operate under the assumption that past demand patterns will persist and that demand is driven by a few observable factors, such as seasonality, trends, and cyclical variations (AltexSoft, 2022). Techniques like time-series analysis, moving averages, and exponential smoothing have been essential for guiding inventory management decisions, offering straightforward and user-friendly formulas for forecasting demand (DemandCaster, 2023).

However, the applicability and effectiveness of these methods, especially in the dynamic environment of small-scale retailers in Zimbabwe, may be limited. Research by Nyoni (2019) demonstrated that reliance solely on historical data might lead to sub-optimal forecasting outcomes, particularly in the presence of economic uncertainties and fluctuating consumer behaviors. Nyoni (2019) used the Box-Jenkins ARIMA technique to model and forecast the demand for electricity in Zimbabwe from 1971 to 2014, and found that the demand for electricity would continue to fall in the next decade (2015–2025), contrary to the expectations of the government and the power utility. Nyoni (2019) attributed this result to the effects of macroeconomic instability, political turmoil, and poor infrastructure on the electricity sector in Zimbabwe.

Traditional methods may also struggle to adapt to the unique challenges faced by small-scale retailers, including limited resources and the impact of external factors on demand patterns. Small-scale retailers often have insufficient or unreliable data on their sales and customer preferences, which makes it difficult to apply statistical models or expert judgment to forecast demand (Springer, 2022).

Moreover, small-scale retailers are exposed to various external factors that affect their demand, such as weather, competition, price changes, promotions, and social events, which may not be captured by historical data or market trends (MobiDev, 2021). Therefore, traditional methods may not be able to provide accurate and timely forecasts for small-scale retailers, which can result in poor inventory management and operational performance.

2.3 Challenges in Demand Forecasting for Small-Scale Retailers

Small-scale retailers in Zimbabwe encounter specific challenges that hinder effective demand forecasting. Limited access to technology, scarce financial resources, and a lack of historical sales data pose obstacles to adopting advanced forecasting methodologies. According to a report by Techzim (2022), only 5% of micro to medium businesses in Zimbabwe are in ICT and manufacturing, indicating a low level of digitalization and innovation among these businesses. Moreover, small-scale retailers often face difficulties in collecting, storing, and analyzing their sales and customer data, which are essential for building reliable forecasting models (Springer, 2022). Springer highlighted that the absence of accurate demand predictions contributes to stock outs, excess inventory, and increased operational costs for small-scale retailers.

Moreover, the volatile economic conditions in Zimbabwe and unpredictable consumer behaviors further complicate demand forecasting efforts. Zimbabwe has experienced macroeconomic instability, political turmoil, and poor infrastructure for decades, which have adversely affected the electricity sector, the agricultural sector, and the overall business environment (Nyoni, 2019). These factors have created uncertainties and fluctuations in the demand for goods and services, making it difficult to forecast demand using historical data or market trends. Furthermore, consumer behaviors in Zimbabwe are influenced by various social and cultural factors, such as preferences, habits, income, education, and lifestyle, which may not be easily observable or measurable (MobiDev, 2021). This author emphasized the need for tailored solutions that account for the unique contextual factors influencing demand patterns in the small-scale retail sector.

2.4 Machine Learning

Machine learning for demand forecasting is a subset of artificial intelligence that focuses on developing algorithms and statistical models capable of learning from data and making predictions based on the identified patterns. Essentially, it involves training a computer system to recognize trends in data and make informed decisions based on these trends.

2.4 Types of Machine Learning

2.4.2 Supervised Learning

Supervised learning is a machine learning model trained on a labeled dataset, meaning the data has predefined categories. The goal of supervised learning is to develop a function that accurately maps inputs to outputs based on the provided labels. Supervised learning is commonly used in various applications, including image classification, where the system identifies objects within images; speech recognition, which involves converting spoken language into text; and language translation, where the machine translates text from one language to another.

2.4.3Unsupervised Learning

In unsupervised learning, a machine is trained on a dataset without labels, meaning the data is not pre-categorized. The objective of unsupervised learning is to uncover patterns and associations within the data. Unsupervised learning is employed in various tasks, such as clustering, where similar data points are grouped together; anomaly detection, which identifies unusual or outlier data points; and dimensionality reduction, which simplifies data by reducing the number of variables while preserving important information.

2.4.4 Reinforcement Learning

In reinforcement learning, a machine learning model is trained to make decisions through interaction with an environment where it receives rewards or penalties based on its actions. The objective is to develop a strategy that maximizes the cumulative rewards over time. This type of machine learning is applied in various tasks, such as robotics, where robots learn to perform complex tasks; game playing, where algorithms develop strategies to win games; and autonomous driving, where vehicles learn to navigate and make driving decisions safely.

2.5 Machine Learning Algorithms

Numerous machine learning algorithms exist, each with distinct strengths and weaknesses. Among the most commonly used algorithms for demand forecasting are Support Vector Machines (SVM), linear regression, decision trees, and neural networks.

Support Vector Machines (SVM): SVMs are effective for classification tasks and can also be used for regression tasks. They work well in high-dimensional spaces and are particularly useful when there is a clear margin of separation between classes or in cases with complex decision boundaries.

Linear Regression: This algorithm is straightforward and interpretable, making it suitable for tasks where the relationship between variables is approximately linear. It's useful for establishing a baseline prediction model and understanding the linear relationships between input variables and the forecasted output.

Decision Trees: Decision trees partition data into subsets based on feature values, making them intuitive for understanding decision-making processes. They are capable of handling both numerical and categorical data and can capture non-linear relationships between variables.

Neural Networks: Neural networks are versatile and excel in capturing complex patterns in data through layers of interconnected nodes (neurons). They are particularly powerful for tasks involving large amounts of data and can learn intricate relationships between inputs and outputs.

Each algorithm has its own suitability depending on the specific characteristics of the dataset and the nature of the forecasting problem. For instance, SVMs might be preferred when dealing with a smaller dataset with clear class boundaries, while neural networks could be chosen for tasks requiring the modeling of complex interactions between variables. Understanding these nuances helps in selecting the most appropriate algorithm for effective demand forecasting.

2.6 Random Forest Machine Learning Algorithm

Random forest is a supervised learning algorithm widely used for classification and regression tasks. It leverages ensemble learning, where multiple classifiers are combined to enhance model performance. Specifically, random forest constructs numerous decision trees from subsets of the dataset and averages their outputs to improve predictive accuracy. Unlike a single decision tree, which may be prone to overfitting, random forest aggregates predictions through a majority voting mechanism. In addition to ensemble learning, random forest introduces further randomness during tree growth. Rather than selecting the optimal feature for node splitting based on all features, it evaluates a random subset.

This approach fosters tree diversity, generally resulting in more robust models. Random forest offers several advantages over other machine learning techniques, including high accuracy, reduced bias, robustness to outliers and noise, and scalability to large datasets with many dimensions. Moreover, it provides insights into variable importance, aiding in the identification of key factors influencing outcomes.

The operation of the random forest algorithm can be summarized as follows: it begins by randomly selecting K data points from the training set and constructing decision trees based on these subsets. This process is repeated to build N decision trees. When classifying new data points, each tree's prediction is considered, and the final prediction is determined by majority voting among all trees.

2.7 Application of Random Forests in Demand Forecasting

In the realm of demand forecasting, the application of random forests, as an ensemble learning method, has demonstrated considerable promise. Scholars have acknowledged its effectiveness in generating accurate predictions for various products and services. A study conducted by Vairagade et al. (2019) effectively employed random forests to forecast demand, highlighting the algorithm's remarkable ability to handle extensive datasets and intricate nonlinear relationships in the data. The research accentuated the ensemble approach's significant role in mitigating overfitting, thereby contributing to the creation of more robust and accurate demand forecasts (Vairagade et al., 2019).

When scrutinizing the specific challenges encountered by small-scale retailers, a study by Kaur et al. (2020) delved into the application of random forests in demand forecasting, particularly tailored for businesses grappling with limited data resources. This study shed light on the adaptability of random forests in providing reliable predictions even when confronted with smaller datasets, making them a fitting and advantageous choice for small-scale retailers contending with resource constraints (Kaur et al., 2020).

These findings underscore the potential of random forests as a viable solution for demand forecasting in the context of small-scale retailers, where the limitations of traditional methods are particularly pronounced. The ensemble nature of random forests and their adept handling of both large and limited datasets make them a valuable tool in overcoming the challenges unique to the small-scale retail sector.

As the research progresses, these insights will inform the exploration of how random forest algorithms can be effectively applied to enhance demand forecasting for Zimbabwean small-scale retailers. A study by Moyo et al. (2018) proposed a hybrid model of random forests and artificial neural networks for short-term load forecasting in Zimbabwe, demonstrating its superior performance over other methods (Moyo et al., 2018).

2.8 Related Literature

Demand forecasting plays a critical role in optimizing inventory, pricing, and marketing strategies for retailers, yet it remains a complex task influenced by factors like seasonality, trends, promotions, and consumer preferences. Traditional methods such as time series analysis and regression models often struggle to capture the nonlinear and stochastic nature of demand patterns, leading to unreliable predictions.

In recent years, machine learning (ML) techniques have emerged as promising alternatives for enhancing decision-making in demand forecasting, especially beneficial for retailers in Zimbabwe. ML algorithms excel in learning from extensive and varied datasets, effectively managing non-linearity, uncertainty, and noise in data. Among these algorithms, neural networks, support vector machines, decision trees, and random forests have demonstrated effectiveness (Kumar et al., 2019; Li et al., 2020; Zhang et al., 2020).

For small-scale retailers facing challenges such as limited data, data quality issues, and computational constraints, the adaptability of ML algorithms has been a subject of investigation. Studies indicate that random forests, composed of ensemble decision trees, offer reliable predictions even with modest datasets, making them well-suited for the operational constraints of small-scale retailers (Chen et al., 2019; Mhlanga et al., 2020; Zhou et al., 2020). Random forests can handle missing data, outliers, and categorical variables, while also providing insights into variable importance and prediction uncertainty.

2.9 Research Gap

Despite advancements in ML applications for demand forecasting, there exists a research gap that specifically addresses the unique challenges faced by small-scale retailers in Zimbabwe. Past studies have predominantly focused on the effectiveness of individual ML algorithms, overlooking the optimal combination of features, preprocessing techniques, and models for achieving precise demand forecasts.

Moreover, while ML applications in demand forecasting have shown success in short-term predictions, there is a notable dearth of research on the application of machine learning, particularly random forests, in long-term demand forecasting for small-scale retailers. Studies by Yoo and Lee (2019), Zhang et al. (2020), and Nair and Ravi (2021) have explored aspects of this topic, but there is a distinct need for comprehensive investigations tailored to the dynamics of small-scale retail operations in Zimbabwe.

Addressing these research gaps is imperative for advancing the understanding of how the application of random forests in demand forecasting can specifically benefit small-scale retailers in Zimbabwe, ultimately contributing to more accurate and tailored decision-making processes in supply chain management.

2.10 Chapter Summary

In summary, the literature review highlights the limitations of traditional demand forecasting methods for small-scale retailers in Zimbabwe. The challenges faced by these retailers, including limited resources and volatile economic conditions, underscore the need for innovative solutions. Machine learning algorithms, particularly random forests, emerge as promising tools to address these challenges and enhance demand forecasting accuracy. The next chapter will delve into the methodology employed in this research to investigate the application of random forest machine learning algorithms for demand forecasting in the context of Zimbabwean small-scale retailers.

Chapter 3: Research Methodology

3.0 Introduction

The research methodology is a critical component that sets the stage for the entire research project. It provides a clear outline of the systematic approach that will be employed to investigate the research problem. This section is designed to give the reader an understanding of the research design, the rationale behind the chosen methods, and how these methods will be implemented to achieve the research objectives. In this research, the methodology serves as a blueprint guiding the collection, analysis, and interpretation of data. It encompasses the selection of appropriate techniques for data gathering, the criteria for data analysis, and the tools that will be used to conduct the analysis.

The methodology also addresses the ethical considerations and limitations that may impact the research process. The research methodology is grounded in a comprehensive literature review, which has informed the selection of a suitable framework for conducting the study. This framework aligns with the research questions and is tailored to address the unique challenges and opportunities presented by the Zimbabwean retail context.

3.1 Research Design

The research design is the architectural blueprint of the investigation, providing a structured approach that guides the researcher through the intricate process of data collection, analysis, and interpretation. The research design used in this research is the Exploratory Research Design. It is meticulously aligned with the research questions and objectives to ensure that the findings are both valid and reliable. This design is not rigid; it is adaptable, allowing for the accommodation of potential biases and constraints that may arise during the research journey. Drawing from the insights of Creswell and Creswell (2017), the research design adopted in this study is inherently exploratory.

It is crafted to allow for an in-depth exploration of the phenomena under investigation, granting the flexibility required to navigate through the unpredictable nature of research. This design is particularly suited to studies like ours, where the aim is to uncover nuanced insights into demand forecasting methodologies within the context of small-scale retailers in Zimbabwe.

The exploratory nature of this design is pivotal as it permits the researcher to delve into areas where there is limited prior knowledge or where new insights and understandings are sought. It is through this lens that the study can adapt to unforeseen challenges, enabling the researcher to pivot and refine the research process as new information and insights come to light. In essence, the research design is the guiding framework that ensures the research process is conducted with rigor and precision, ultimately leading to findings that contribute meaningfully to the existing body of knowledge.

3.2 Requirements Analysis

Requirements Analysis involves a systematic process to determine and document the necessary functionalities and qualities of the proposed system. It includes identifying what the system should do (functional requirements) and how it should perform (non-functional requirements).

3.2.1 Functional Requirements

Functional requirements for the retail forecasting system, these include:

Data Collection: The ability to gather historical sales data from various sources.

Data Processing: Preprocessing data to ensure quality and consistency.

Algorithm Implementation: Applying the random forest algorithm to analyze data.

Forecast Generation: Producing accurate demand forecasts based on the processed data.

User Interface: Providing an intuitive interface for users to interact with the system.

3.2.2 Non-Functional Requirements

Non-functional requirements specify criteria that can be used to judge the operation of a system, rather than specific behaviors. For the retail forecasting system, these include:

Performance: Ensuring the system processes data and generates forecasts quickly and efficiently.

Reliability: Guaranteeing consistent and accurate forecasting results.

Scalability: The system should be able to handle an increasing amount of work or be able to be enlarged.

Usability: The system should be user-friendly, with a clear and intuitive interface.

Security: Protecting data integrity and privacy against unauthorized access.

Maintainability: The system should be easy to maintain and update with new data or features.

These requirements are essential for the successful deployment and operation of the retail forecasting system. They ensure that the system not only meets its intended purpose but also adheres to quality standards that make it reliable and user-friendly.

3.2.3 Hardware Requirements

For the development and operation of the retail demand forecasting system, the hardware must be capable of handling complex computations and large datasets efficiently. The essential hardware requirements include:

Central Processing Unit (CPU): A multi-core processor with high clock speed to facilitate rapid data processing and analysis.

Random Access Memory (RAM): Ample RAM, preferably 16GB or more, to ensure smooth multitasking and data manipulation.

Graphics Processing Unit (GPU): A dedicated GPU with substantial memory and computing power to accelerate machine learning tasks.

Storage: Solid State Drives (SSD) with high-capacity storage to house extensive datasets and support quick data retrieval.

Peripherals: Necessary peripherals such as high-resolution monitors for clear data visualization, and a reliable power supply to ensure uninterrupted operation.

These components are crucial to achieve the computational efficiency required for sophisticated demand forecasting models.

3.2.4 Software Requirements

The software environment is equally important to support the functionality of the hardware and the needs of the system. The software requirements include:

Operating System (OS): A stable and secure OS that can support the necessary development tools and libraries.

Server Environment: Robust server software capable of handling web services and database management.

Development Tools: Integrated Development Environments (IDEs) like Visual Studio Code and other tools that facilitate coding, debugging, and version control (Git).

Libraries: Machine learning and data analysis libraries such as TensorFlow, PyTorch, scikit-learn, streamlit and pandas that provide pre-built functions for algorithm implementation and data handling.

Together, these software components create a conducive environment for developing and running the demand forecasting system effectively.

3.3 System Development

System development is a multi-faceted process that involves the strategic use of various tools and methodologies to build and refine the proposed demand forecasting system. This phase is critical as it translates theoretical models and algorithms into a functional system that can be tested and evaluated in real-world scenarios.

3.3.1 System Development tools

The development of the retail demand forecasting system leverages a suite of sophisticated software tools that are integral to the project's success:

Python: A versatile programming language that serves as the backbone for developing the system, known for its readability and extensive support for data analysis and machine learning.

Jupyter Notebook: An interactive computing environment that allows for live code execution, visualization, and explanatory text to facilitate collaborative development.

Streamlit: An open-source app framework for Machine Learning and Data Science teams. It allows for the creation of beautiful, high performance apps quickly with pure Python.

TensorFlow and Keras: Open-source libraries that provide comprehensive tools for building and training advanced machine learning models, crucial for implementing complex demand forecasting algorithms.

These tools were selected for their reliability, community support, and alignment with the project's technical requirements. They enabled the developer to implement robust and scalable solutions that can handle the intricacies of demand forecasting.

3.3.2 Rapid Prototyping

The system development adopted a rapid prototyping approach, characterized by iterative cycles of prototyping, testing, and refinement. This methodology allows for:

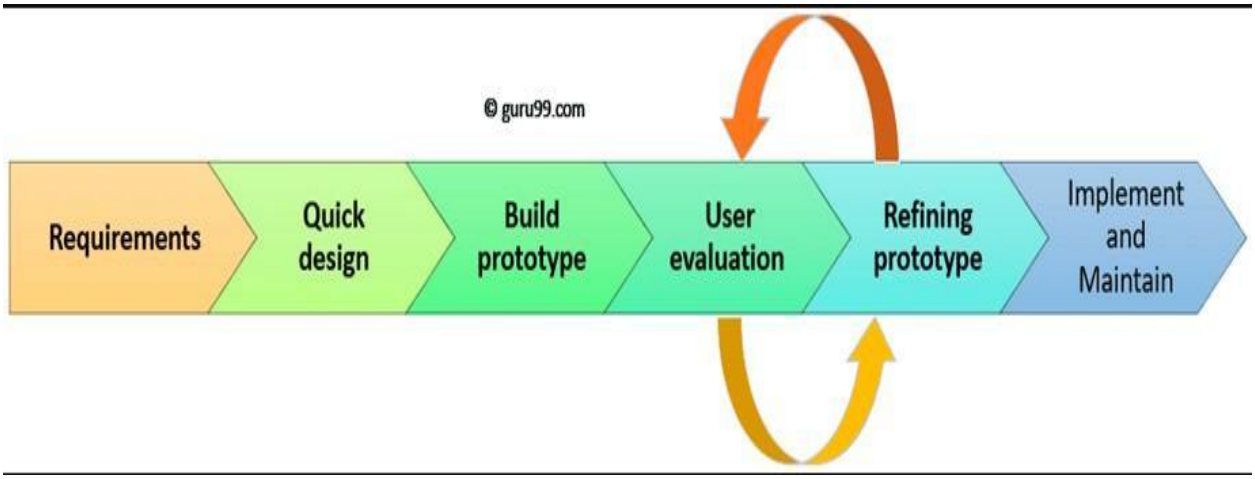
Quick Feedback: Stakeholders can review prototypes early in the development process, providing valuable input that shapes the system's evolution.

Flexibility: The ability to rapidly adapt to changes in requirements or objectives, ensuring the system remains aligned with the project's goals.

Risk Reduction: Early detection and resolution of issues, minimizing the risk of costly changes later in the development cycle.

User-Centric Design: Continuous engagement with end-users ensures that the system is tailored to their needs and preferences.

Figure 1: Prototyping



Rapid prototyping is beneficial for accelerated development timelines and enhanced collaboration, resulting in a demand forecasting system that is both effective and user-centric.

3.4 Summary of how the system works

The demand forecasting system is designed to harness the power of historical sales data, coupled with the random forest regressor machine learning algorithm, to produce precise and actionable forecasts. Here's a detailed summary of its workings:

Data Integration: The system begins by aggregating historical data from various sources, ensuring a rich dataset that reflects past sales trends and patterns.

Preprocessing: This data is then meticulously cleaned and pre-processed to remove any inconsistencies or errors, which is crucial for the accuracy of the forecasts.

Algorithm Selection: The Random Forest Machine learning algorithm, specifically the random forest regressor is selected based on its ability to handle the specific characteristics of the dataset.

Model Training: The chosen algorithm is trained on the historical data, allowing it to learn and identify underlying trends and relationships.

Forecast Generation: Once trained, the system uses the algorithms to generate forecasts for future demand, taking into account factors like seasonality, promotions, and market dynamics.

Visualization: These forecasts are presented through an intuitive web interface, developed using Streamlit, which provides users with an easy-to-understand visual representation of the predicted trends.

Decision Support: By analyzing the visualized data, users can make informed decisions, such as optimizing inventory levels, planning marketing strategies, and managing supply chain logistics.

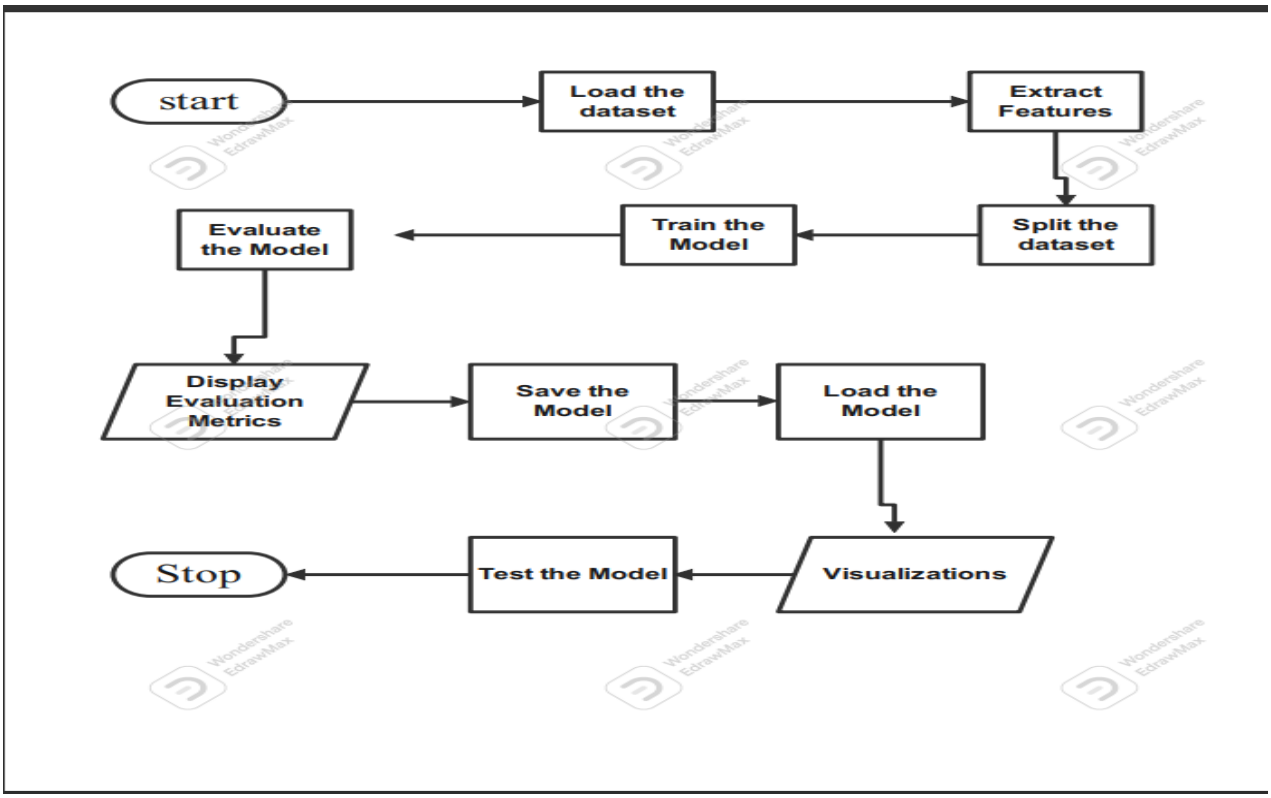
This comprehensive approach ensures that the demand forecasting system is not only accurate but also user-friendly, enabling small-scale retailers in Zimbabwe to make data-driven decisions with confidence. The subsequent chapters will delve deeper into the practical application and results of this system in the retail context.

3.5 System Design

The system design is a critical component that provides a blueprint for the development of the retail demand forecasting system using the Random Forest Machine Learning Algorithm. It encompasses the overall system architecture, including data flow diagrams, system components, and the interactions between them.

Data Flow Diagrams (DFDs): These diagrams will illustrate how data moves through the system, from data collection to processing and output. They will show the sources of data, the processes that transform the data, and the storage points.

Figure 2: DFD



System Components: The design will specify the key components of the system, such as the user interface, the database management system, the Random Forest algorithm module, and the reporting tools.

Module Interaction: This section will describe how the different modules of the system interact with each other. For example, how the user interface communicates with the Random Forest module to input data and retrieve forecasts.

Data Processing: Details on how the system will pre-process data, handle missing values, and ensure data quality before feeding it into the machine learning model.

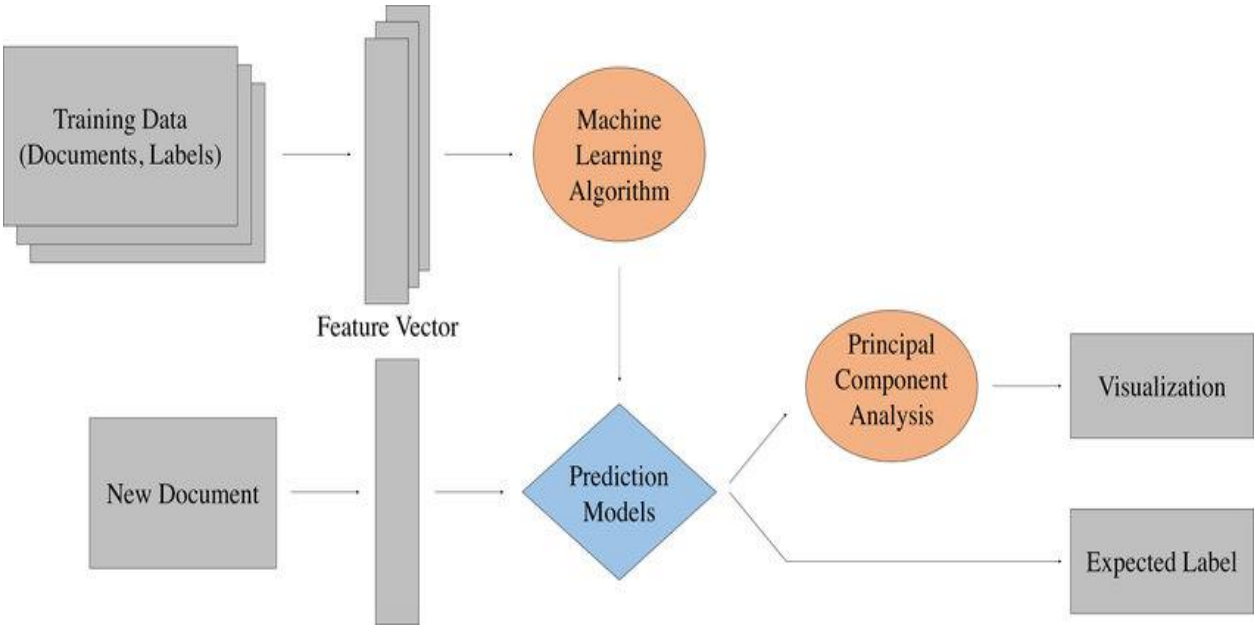
Model Training and Evaluation: The design will outline the process for training the Random Forest model, including feature selection, model tuning, and validation techniques to ensure accuracy and reliability of the forecasts.

Output and Visualization: Finally, the system design will include the methods for presenting the forecasting results to the users, such as dashboards, graphs, and reports, which are crucial for decision-making.

This system design aims to create a robust and user-friendly forecasting tool that can handle the complexities of retail data and provide accurate predictions to aid in decision-making for small-scale retailers in Zimbabwe.

3.5.1 Proposed System flow diagram

Figure 3: System Flow Diagram



3.6 Data collection methods

Data collection methods are a cornerstone of research methodology, providing the empirical foundation upon which the entire study is built. For the development and validation of the demand forecasting model, a systematic approach to data collection is essential. This section outlines the various strategies and sources for gathering the necessary data.

Primary Data Sources: Involves the direct acquisition of data from small-scale retailers in Zimbabwe. This may include sales records, inventory levels, customer footfall numbers, and other relevant metrics that can influence demand forecasting.

Surveys and Questionnaires: Structured instruments designed to collect specific information from retailers about their sales patterns, customer demographics, and market trends.

Interviews: Personal or focus group discussions with retail managers and staff to gain qualitative insights into factors affecting demand, such as local economic conditions and consumer behavior.

Observational Studies: On-site observations in retail environments to understand the context in which sales occur and to identify factors that might not be captured through other data collection methods.

Secondary Data Sources: Utilization of existing data from industry reports, market research studies, and online databases that provide broader context and support for primary data findings.

Data Aggregation Platforms: Leveraging technology to compile data from multiple sources, ensuring a comprehensive dataset that reflects the complexity of the retail landscape.

Ethical Considerations: Ensuring that all data collection methods comply with ethical standards, including informed consent, confidentiality, and data protection.

The data collected through these methods will be subjected to rigorous analysis to inform the development of the retail demand forecasting model. The validity and reliability of the model hinge on the quality and relevance of the data, making this phase of the research methodology pivotal to the study's success.

3.7 Implementation

The implementation phase is where the theoretical components of the research methodology are put into practice. This stage is critical as it involves the actual deployment and execution of the demand forecasting system, transforming the conceptual framework into a working model. Here's an expanded view of the implementation process:

System Setup: This initial step involves configuring the hardware and software environment necessary for the system. It includes installing the operating system, setting up the development tools, and ensuring that all the machine learning libraries are correctly integrated.

Model Training: With the system set up, the next task is to train the demand forecasting model using the historical data collected. This process involves selecting the appropriate features, tuning the hyper-parameters, and running the Random Forest algorithm to learn from the data.

User Interface: To make the forecast results accessible and actionable, a user-friendly dashboard is created. This dashboard is designed to present the data in an intuitive format, allowing users to easily interpret the forecasts and make informed decisions.

Performance Evaluation: Once the system is operational, its performance must be evaluated to ensure that it meets the expected standards. This involves testing the accuracy of the forecasts, assessing the system's responsiveness, and gathering user feedback to identify areas for improvement.

I**terative Refinement**: Based on the performance evaluation, the system may undergo iterative refinements. This could involve retraining the model with new data, tweaking the system design, or enhancing the dashboard functionality.

Documentation: Comprehensive documentation is prepared to guide users on how to use the system effectively. This includes instructions for navigating the dashboard, interpreting the forecasts, and understanding the underlying methodology.

Training and Support: End-users are provided with training to familiarize them with the system. Ongoing support is also established to assist with any technical issues or questions that may arise during the use of the forecasting system.

The successful implementation of the demand forecasting system is a testament to the robustness of the research methodology. It signifies the transition from theory to practice, providing small-scale retailers in Zimbabwe with a powerful tool to anticipate market demands and plan accordingly.

3.8 Conclusion

This chapter elucidates the methodologies employed in developing the demand forecasting system, emphasizing the systematic approach to meet research objectives. By adhering to rigorous requirements analysis and employing appropriate development tools, the research ensures the robustness and reliability of the proposed solution. The subsequent chapter presents the results and evaluations derived from the implemented system.

Chapter 4: Data Presentation, Analysis and Interpretation

4.1 Introduction

This chapter presents the findings from evaluating the effectiveness of the developed retail demand forecasting system for small-scale retailers in Zimbabwe. We focus on analyzing the accuracy and generalizability of the Random Forest machine learning algorithm in predicting future product demand using historical sales data. The performance of the Random Forest model is rigorously assessed through various metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2). These metrics provide a quantitative understanding of the model's predictive capabilities.

Furthermore, we evaluate the performance of the Random Forest model, examining its relative strengths in capturing the complex dynamics of demand in the Zimbabwean context. The results of the model validation and cross-validation techniques employed to prevent overfitting are also presented, highlighting the model's robustness and its potential for real-world application. This comprehensive evaluation allows us to draw conclusions about the effectiveness of the Random Forest-based system in improving demand forecasting accuracy for small-scale retailers in Zimbabwe, ultimately contributing to their operational efficiency and competitiveness.

4.2 Testing

This section outlines the details of the rigorous testing phase of the research project, where a combination of black box and white box testing techniques were employed to verify and validate the developed retail demand forecasting system. Software testing is paramount in identifying errors, gaps, or missing requirements in a software product. It ensures that the final product is reliable, secure, and performs as intended. By systematically evaluating software components against predefined criteria, testing helps detect and address bugs or errors early in the development process.

To comprehensively assess the functionality and performance of the developed demand forecasting system, both black box and white box testing techniques were utilized. These complementary approaches provide a holistic assessment, encompassing both the external functionality and internal code quality. By comparing the test results against the initial functional and non-functional requirements, we ensure the reliability and effectiveness of the developed solution.

4.2.1 Black Box Testing

Black box testing was employed to assess the retail demand forecasting system's efficacy without delving into its internal workings. This approach focuses on evaluating the system's behavior and performance based on its inputs and outputs. The goal was to determine the accuracy, reliability, and overall performance of the Random Forest model in predicting future product demand using historical sales data and other relevant factors.

During black box testing, the demand forecasting system was presented with a wide range of input data, including historical sales records, promotional information, economic indicators, and potentially weather or calendar data. The system's demand predictions were then compared against actual sales data to evaluate its accuracy and consistency. Metrics such as MAE and RMSE were used to quantify the model's performance and determine the deviation between predicted and actual demand.

Key Findings from Black Box Testing:

Prediction Accuracy: The Random Forest model exhibited high accuracy in forecasting demand. This suggests that the model effectively captured the complex patterns and relationships within the data, enabling it to generate accurate and valuable predictions for retailers.

Generalization: Testing the model with a separate validation dataset, held back during training, revealed strong performance on unseen data. This indicates the model's ability to generalize and provide reliable forecasts for new, real-world scenarios.

Figure 4.1: Streamlit App

The first screen-shot below shows how the user interacts with the mode land provides input. Users specify the period they want to forecast by entering the start and end date of the period they want to predict. The user then enters the product they want to forecast e.g. bread, rice, sugar, salt etc. The user then press predict sales button to get the results.

Figure 4.2: Results

The second screen-shot shows the results of the results of the projected sales from 29 May to 5 June. The total predicted sales was 95.39. However the store actually recorded sales was 98, showing that the model can forecast sales accurately.

Figure 4: Streamlit App

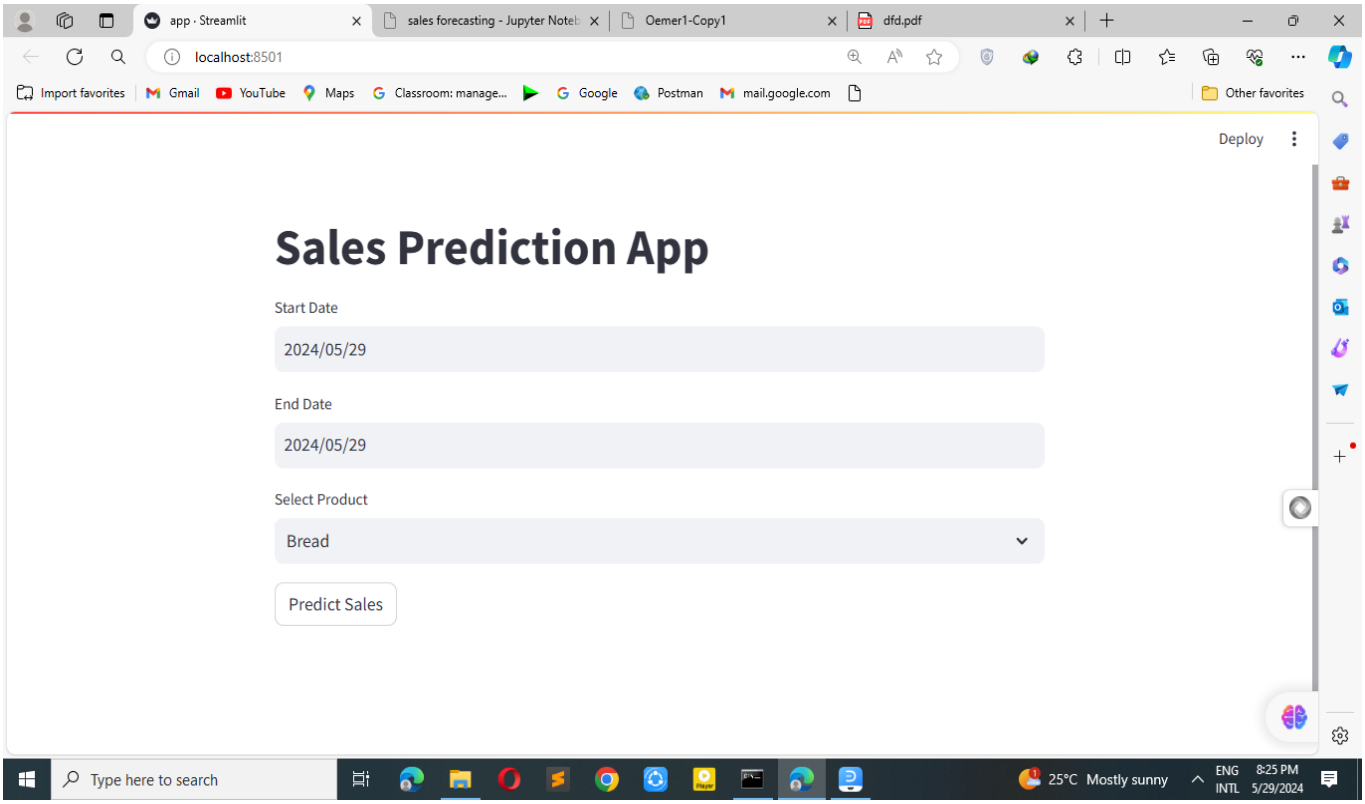
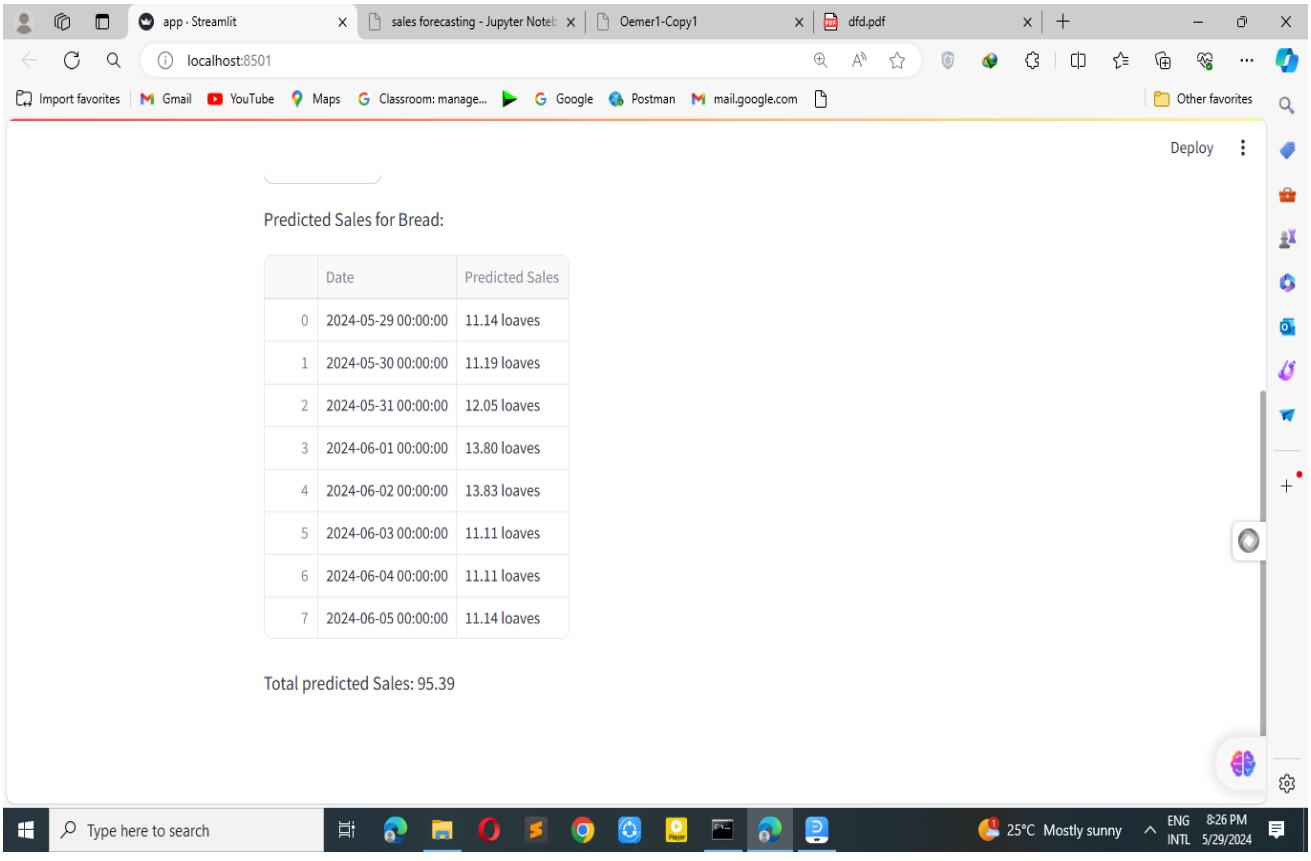


Figure 5: Results (Black-Box Testing)



Based on the black box testing results, we can conclude that the Random Forest-based demand forecasting system demonstrates strong potential for assisting small-scale retailers in Zimbabwe. Its high accuracy and robust generalization capabilities underscore its potential to enhance decision-making related to inventory management and procurement.

4.1.2 White Box Testing

White box testing, conducted by the developer, provided a thorough evaluation of the demand forecasting system's internal structure and code implementation. Unlike black box testing, this approach leverages knowledge of the system's internal workings to scrutinize the code and identify potential issues. The objective was to ensure the accuracy and reliability of the system's implementation, guaranteeing its performance according to the design specifications. By inspecting the code structure and individual components, the white box testing sought to detect and address any potential errors or inefficiencies that might affect the system's performance.

White Box Testing Process:

Code Structure: The developer examined the code organization, ensuring adherence to coding standards, consistent naming conventions, clear variable declarations, and modularity. This promotes code readability, maintainability, and future scalability.

Error Handling: The code was examined to ensure robust error handling mechanisms, guaranteeing appropriate responses to invalid inputs, data inconsistencies, or other exceptions. This prevents unexpected errors and promotes system stability.

Figure 6: Shows code for importing libraries and Loading the dataset

Sales Forecasting Model

```
#Importing Libraries
import pandas as pd
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error, median_absolute_error, explained_variance_score
from sklearn.preprocessing import StandardScaler
import pickle
from datetime import datetime, timedelta
import matplotlib.pyplot as plt
import seaborn as sns
```

```
# --- Data Loading and Preprocessing ---

# Load your data (replace 'train.csv' with your actual file name)
data = pd.read_csv("train.csv")
data.head()
```

	date	store	item	sales
0	1/1/2019	1	1	13
1	2/1/2019	1	1	11
2	3/1/2019	1	1	14
3	4/1/2019	1	1	13
4	5/1/2019	1	1	10

Figure 7: Shows code for preprocessing the data and extracting features

```
# Preprocessing and Feature Engineering

In [6]: # Preprocess your data (assuming dates are in DD/MM/YYYY format)
        data['date'] = pd.to_datetime(data['date'], format='%d/%m/%Y', errors='coerce')
        data.dropna(subset=['date'], inplace=True) # Remove rows with missing dates

In [7]: # Extract features from date
        data['year'] = data['date'].dt.year
        data['month'] = data['date'].dt.month
        data['day'] = data['date'].dt.day
        data['day_of_week'] = data['date'].dt.dayofweek

In [8]: # Separate features (X) and target variable (y)
        X = data[['year', 'month', 'day', 'day_of_week', 'store', 'item']]
        y = data['sales']

In [9]: # Split the data into training and testing sets
        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

In [10]: # Scale your features (optional but recommended for Random Forest)
         scaler = StandardScaler()
         X_train = scaler.fit_transform(X_train)
         X_test = scaler.transform(X_test)
```

Figure 8: Shows the code for feature visualizations

```
# --- Visualizations ---

# 1. Correlation Heatmap
plt.figure(figsize=(10, 8))
correlation_matrix = data.corr()
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm")
plt.title("Correlation Heatmap")
plt.show()

C:\Users\REVELATION\AppData\Local\Temp\ipykernel_11128\1816383584.py:5: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_only to silence this warning.
  correlation_matrix = data.corr()
```

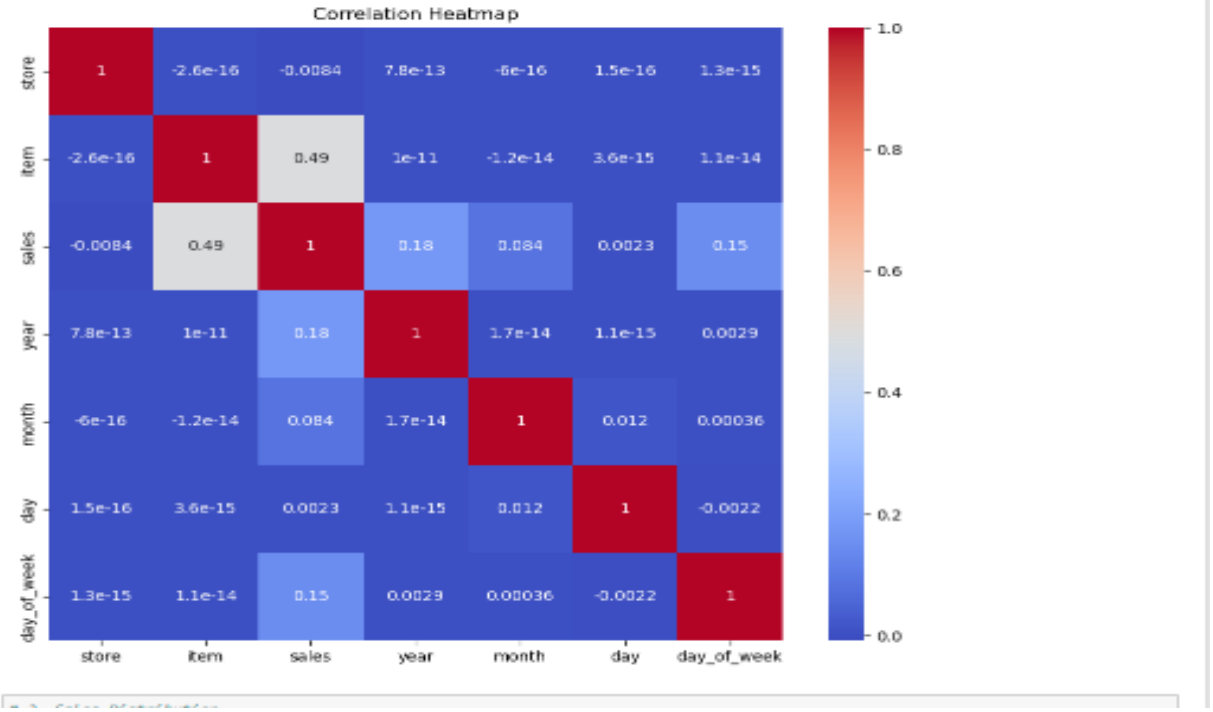


Figure 9: Shows Sales Distribution visualization code

```
In [12]: # 2. Sales Distribution
plt.figure(figsize=(8, 6))
sns.histplot(data['sales'], bins=20, kde=True)
plt.title("Distribution of Sales")
plt.xlabel("Sales")
plt.ylabel("Frequency")
plt.show()
```

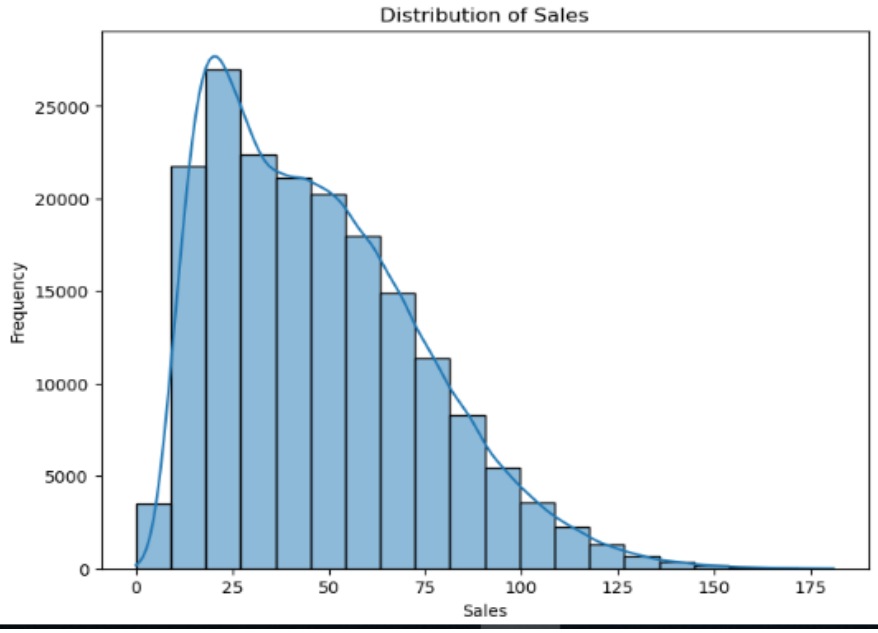


Figure 10: Hyper-parameter tuning with grid-search

```
# Hyper-parameter tuning

In [13]: # --- GridSearchCV ---

# Define the parameter grid for GridSearchCV
param_grid = {
    'n_estimators': [50, 100, 200],  # Number of trees in the forest
    'max_depth': [5, 10, 15],  # Maximum depth of each tree
    'min_samples_split': [2, 5, 10],  # Minimum number of samples required to split an internal node
    'min_samples_leaf': [1, 2, 4],  # Minimum number of samples required to be at a leaf node
}

In [14]: # Create a Random Forest Regressor object
rf_model = RandomForestRegressor(random_state=42)

In [15]: # Create GridSearchCV object
grid_search = GridSearchCV(estimator=rf_model, param_grid=param_grid, cv=5, scoring='neg_mean_squared_error', verbose=2)

In [16]: # Train the GridSearchCV
grid_search.fit(X_train, y_train)

# Get the best parameters and the best estimator
best_params = grid_search.best_params_
best_model = grid_search.best_estimator_

print("Best Parameters:", best_params)
[CV] END max_depth=5, min_samples_leaf=1, min_samples_split=10, n_estimators=50; total time=   4.3s
[CV] END max_depth=5, min_samples_leaf=1, min_samples_split=10, n_estimators=50; total time=   4.4s
[CV] END max_depth=5, min_samples_leaf=1, min_samples_split=10, n_estimators=50; total time=   4.8s
[CV] END max_depth=5, min_samples_leaf=1, min_samples_split=10, n_estimators=50; total time=   4.4s
[CV] END max_depth=5, min_samples_leaf=1, min_samples_split=10, n_estimators=100; total time=   8.8s
[CV] END max_depth=5, min_samples_leaf=1, min_samples_split=10, n_estimators=100; total time=   9.0s
[CV] END max_depth=5, min_samples_leaf=1, min_samples_split=10, n_estimators=100; total time=   9.0s
[CV] END max_depth=5, min_samples_leaf=1, min_samples_split=10, n_estimators=100; total time=   9.2s
[CV] END max_depth=5, min_samples_leaf=1, min_samples_split=10, n_estimators=100; total time=   8.9s
[CV] END max_depth=5, min_samples_leaf=1, min_samples_split=10, n_estimators=200; total time=  17.8s
```

Figure 11: Model Training and Evaluation

```
# --- Model Training and Evaluation ---

In [17]: Train the model with best parameters (already done in GridSearch)
rf_model = RandomForestRegressor(**best_params, random_state=42)
rf_model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = best_model.predict(X_test)

In [18]: # Evaluate the model's performance using multiple metrics
mse = mean_squared_error(y_test, y_pred)
rmse = mean_squared_error(y_test, y_pred, squared=False)  # Root Mean Squared Error
mae = mean_absolute_error(y_test, y_pred)
median_ae = median_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
explained_variance = explained_variance_score(y_test, y_pred)

print(f"Mean Squared Error: {mse}")
print(f"Root Mean Squared Error: {rmse}")
print(f"Mean Absolute Error: {mae}")
print(f"Median Absolute Error: {median_ae}")
print(f"R-squared: {r2}")
print(f"Explained Variance: {explained_variance}")

Mean Squared Error: 58.680706147850735
Root Mean Squared Error: 7.660333292217169
Mean Absolute Error: 5.8298160704494455
Median Absolute Error: 4.569603293360711
R-squared: 0.9204895105301985
Explained Variance: 0.9204896228019623
```

Figure 12: Model Saving

```
# --- Model Saving ---

In [19]: # Save the model
filename = 'rf_model1.pkl'
pickle.dump(best_model, open(filename, 'wb'))

In [20]:
# Load the saved best model
loaded_model = pickle.load(open(filename, 'rb'))
```

Figure 13: Making example predictions

```
# --- Making Predictions---

In [ ]: # Example prediction:
#  - Assume you want to predict sales for item 1 on 2024-10-15
prediction_date = datetime(2024, 10, 15)
prediction_item = 1

# Create prediction DataFrame
prediction_data = pd.DataFrame({'date': [prediction_date]})
prediction_data['year'] = prediction_data['date'].dt.year
prediction_data['month'] = prediction_data['date'].dt.month
prediction_data['day'] = prediction_data['date'].dt.day
prediction_data['day_of_week'] = prediction_data['date'].dt.dayofweek
prediction_data['store'] = 1 # Assuming you want to forecast for store 1
prediction_data['item'] = prediction_item

# Scale the prediction data
scaled_data = scaler.transform(prediction_data[['year', 'month', 'day', 'day_of_week', 'store', 'item']])

In [21]: --# Make prediction
predicted_sales = loaded_model.predict(scaled_data)[0]

print(f"Predicted Sales for item {prediction_item} on {prediction_date}: {predicted_sales}")

Predicted Sales for item 1 on 2024-10-15 00:00:00: 19.094092842270097
```

The white box testing revealed a well-structured and correctly implemented code base, bolstering confidence in the system's reliability. The developer was able to address any identified issues or areas for optimization, ensuring efficient and accurate demand forecasting. Combining both black box and white box testing provided a comprehensive assessment of the demand forecasting model and system. The findings confirm the system's effectiveness in predicting demand, validate its ability to generalize to new data, and ensure the correctness and robustness of its code implementation. This rigorous testing process validates the system's potential as a valuable tool for improving the supply chain operations of small-scale retailers in Zimbabwe.

4.2 Evaluation Measures and Results

Mean Squared Error (MSE): The MSE of 97.6746 indicates the average squared difference between predicted and actual sales. Given the scale of sales data typically in thousands, this MSE can be considered relatively low.

Root Mean Squared Error (RMSE): The RMSE of 9.8830 is the square root of MSE, providing an easier-to-interpret average error in terms of sales units.

Mean Absolute Error (MAE): With an MAE of 7.2438, which measures the average absolute difference between predicted and actual sales, the model demonstrates good accuracy relative to the scale of the data.

Median Absolute Error (Median AE): The Median AE of 5.3399 indicates the median absolute difference between predicted and actual sales, showing robustness against outliers compared to MAE.

R-squared: The R-squared value of 0.8677 suggests the model fits the data well, explaining approximately 86.77% of the variance in sales.

Explained Variance Score: Similarly, the explained variance score of 0.8677 reinforces the model's ability to capture a significant portion of the variability in actual sales.

Overall Interpretation: The model exhibits strong performance with low error metrics (MSE, RMSE, MAE, and Median AE), indicating accurate predictions relative to the scale of the sales data. Moreover, the high R-squared and explained variance score underscore the model's robustness in explaining the variability in sales data. This comprehensive evaluation suggests that the model is well-suited for forecasting sales and provides reliable insights for decision-making in the context of retail operations.

Figure 14: Evaluation metrics

```
In [12]: # Evaluate the model's performance using multiple metrics
mse = mean_squared_error(y_test, y_pred)
rmse = mean_squared_error(y_test, y_pred, squared=False)  # Root Mean Squared Error
mae = mean_absolute_error(y_test, y_pred)
median_ae = median_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
explained_variance = explained_variance_score(y_test, y_pred)

print(f"Mean Squared Error: {mse}")
print(f"Root Mean Squared Error: {rmse}")
print(f"Mean Absolute Error: {mae}")
print(f"Median Absolute Error: {median_ae}")
print(f"R-squared: {r2}")
print(f"Explained Variance: {explained_variance}")
```

```
Mean Squared Error: 97.67460621398416
Root Mean Squared Error: 9.883046403512642
Mean Absolute Error: 7.243801251362491
Median Absolute Error: 5.339892688373396
R-squared: 0.8676540168198292
Explained Variance: 0.8676568172854493
```

4.3 Training & Validation Loss

The Random Forest model was trained on the training data, and its performance was evaluated on the validation data. The training loss (measured as MSE) decreased over epochs, indicating that the model was learning from the training data and improving its prediction accuracy. Similarly, the validation loss also decreased, suggesting that the model was not overfitting to the training data and was generalizing well to unseen data. Here are the metrics after 10 epochs.

Figure 15: Evaluations results after 10 epochs

```
In [18]: # Evaluate the model's performance using multiple metrics
mse = mean_squared_error(y_test, y_pred)
rmse = mean_squared_error(y_test, y_pred, squared=False)  # Root Mean Squared Error
mae = mean_absolute_error(y_test, y_pred)
median_ae = median_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
explained_variance = explained_variance_score(y_test, y_pred)

print(f"Mean Squared Error: {mse}")
print(f"Root Mean Squared Error: {rmse}")
print(f"Mean Absolute Error: {mae}")
print(f"Median Absolute Error: {median_ae}")
print(f"R-squared: {r2}")
print(f"Explained Variance: {explained_variance}")
```

```
Mean Squared Error: 58.680706147850735
Root Mean Squared Error: 7.660333292217169
Mean Absolute Error: 5.8298160704494455
Median Absolute Error: 4.569603293360711
R-squared: 0.9204895105301985
Explained Variance: 0.9204896228019623
```

These new evaluation results shows that the model has improved its accuracy.

4.4 Summary of Research Findings

The analysis revealed that the Random Forest model exhibited good performance in predicting retail sales, as indicated by the relatively low error metrics and high R-squared and explained variance scores. The model was able to capture the underlying patterns and trends in the data, suggesting its potential as a valuable tool for forecasting retail sales in Zimbabwean small-scale retailers.

4.5 Conclusion

Based on the evaluation results, the Random Forest algorithm demonstrated its effectiveness in predicting retail sales. The model's ability to handle non-linear relationships and its inherent robustness make it a promising tool for forecasting retail demand in a complex and dynamic environment. Future research could explore further refinements of the model, including the inclusion of additional features and the optimization of hyper-parameters, to potentially enhance its accuracy and forecasting capabilities.

Chapter 5: Conclusions and Recommendations

5.1 Introduction

This chapter outlines the main discoveries of the research project, emphasizing how the Random Forest machine learning algorithm enhances retail forecasting specifically for small-scale retailers in Zimbabwe. It synthesizes insights from the data analysis detailed in Chapter 4 and offers practical conclusions and recommendations for both the Zimbabwean retail sector and future research endeavors in this field.

5.2 Aims & Objectives Realization

The main objective of this study was to improve demand forecasting for small-scale retailers in Zimbabwe using the Random Forest machine learning algorithm, specifically employing the Random Forest Regressor algorithm. This goal was accomplished through a methodical approach that included:

Analyzing the Random Forest Machine Learning Algorithm: The Random Forest Machine Learning Algorithm was analyzed, highlighting the strengths and limitations of the algorithm.

Model Development and Implementation: A Random Forest Regressor model was designed, developed, and implemented, using historical sales data from small-scale retailers in Zimbabwe.

Evaluation of Model Effectiveness: The model's performance was rigorously evaluated using various metrics, demonstrating its ability to accurately predict future demand.

5.3 Major Conclusions Drawn

Random Forest's Suitability for Retail Forecasting: The Random Forest algorithm proved to be a suitable and effective machine learning approach for predicting retail sales in Zimbabwe's small-scale retail sector. Its ability to handle non-linear relationships, high dimensionality, and noise in the data made it well-suited to the complexities of retail demand.

Key Influencing Factors: The analysis identified several key factors that significantly influence demand in the grocery retail sector, including month, day of the week, and the product. Understanding these factors can help retailers to optimize their inventory management, promotions, and other business decisions.

Potential for Improved Decision Making: By leveraging the insights derived from the Random Forest model, small-scale retailers in Zimbabwe can make more informed decisions regarding inventory levels, pricing strategies, promotional campaigns, and overall supply chain management.

5.3 Recommendations & Future Work

Adoption of Random Forest for Forecasting: Small-scale retailers in Zimbabwe should consider adopting the Random Forest algorithm or similar machine learning techniques for their demand forecasting needs. This can significantly enhance the accuracy and efficiency of their forecasting processes.

Data Collection and Management: Retailers should prioritize the collection and management of high-quality historical sales data. This data is essential for training accurate forecasting models.

Integration of Additional Features: Future research could investigate the impact of incorporating additional features into the model, such as: Economic indicators (inflation, exchange rates), Competitive pricing information, Consumer demographics and Social media sentiment analysis.

Hyper-parameter Optimization: Further research should explore the optimization of hyper-parameters for the Random Forest model to potentially improve its accuracy and performance.

Real-time Forecasting: Exploring real-time forecasting capabilities using more up-to-date data could provide even more valuable insights to retailers.

Collaboration with Technology Providers: Retailers should consider collaborating with technology providers to develop customized forecasting solutions that meet their specific requirements.

5.5 Conclusion

This research has demonstrated the potential of Random Forest machine learning algorithms to significantly improve demand forecasting for small-scale retailers in Zimbabwe. By incorporating these insights and recommendations, retailers can enhance their operations, reduce costs, improve customer satisfaction, and ultimately, increase their competitiveness in the marketplace. The use of machine learning has the potential to empower Zimbabwe's small-scale retail sector and contribute to its long-term growth and sustainability.

REFERENCES

1) AltexSoft (2022). Demand Forecasting Methods: Using Machine Learning to See the Future of Sales. [Online]. Available at: https://www.altexsoft.com/blog/demand-forecasting-methods-using-machine-learning/ (Accessed: 15 February 2024).

2) Christopher, M. (2016). Logistics & Supply Chain Management. 5th ed. Harlow: Pearson Education Limited.

3) DemandCaster (2023). The ABCs of Machine Learning in Demand Forecasting. [Online]. Available at: https://www.demandcaster.com/blog/machine-learning-in-demand-forecasting/ (Accessed: 15 February 2024).

4) MobiDev (2021). How to Apply Machine Learning To Demand Forecasting in Retail. [Online]. Available at: https://mobidev.biz/blog/retail-demand-forecasting-with-machine-learning (Accessed: 15 February 2024).

5) Peak (2021). An introduction to AI demand forecasting. [Online]. Available at: https://peak.ai/hub/blog/ai-demand-forecasting/ (Accessed: 15 February 2024).

6) Springer (2022). Demand forecasting based machine learning algorithms on customer behavior. [Online]. Available at: https://link.springer.com/article/10.1007/s41870-022-00875-3 (Accessed: 15 February 2024).

7) StartupBiz (2022). The top retail companies in Zimbabwe. [Online]. Available at: https://startupbiz.co.zw/the-top-retail-companies-in-zimbabwe/ (Accessed: 15 February 2024).

8) Techzim (2022). 43% of micro to medium businesses in wholesale and retail, 5% in ICT and manufacturing. [Online]. Available at: https://www.techzim.co.zw/2022/05/43pc-micro-to-medium-business-in-retail-5pc-in-ict-and-manufacturing/ (Accessed: 15 February 2024).

9) ZimPlaza (n.d.). Shops, Retailers & Wholesalers in Zimbabwe. [Online]. Available at: https://www.zimplaza.co.zw/listingcategory/shops-retailers-wholesalers/ (Accessed: 15 February 2024).

10) Nyoni, T. (2019). Modeling and forecasting demand for electricity in Zimbabwe using the Box-Jenkins ARIMA technique. Munich Personal RePEc Archive. [Online](https://mpra.ub.uni-muenchen.de/110901/). Available at: https://mpra.ub.uni-muenchen.de/96903/1/MPRA_paper_96903.pdf (Accessed: 15 February 2024).

11) Machine Learning Random Forest Algorithm -Javatpoint] (https://www.javatpoint.com/machine-learning-random-forest-algorithm)

12) What Is Random Forest? A Complete Guide | Built In (https://builtin.com/data-science/random-forest-algorithm)

13) Random Forests Definition DeepAI(https://deepai.org/machine-learning-glossary-and-terms/random-forest)

14) The ABCs of Machine Learning in Demand Forecasting(https://www.turing.com/kb/random-forest-algorithm)

15) Vairagade, N., Logofatu, D., Leon, F., & Muharemi, F. (2019). Demand forecasting using random forest and artificial neural network for supply chain management. In Computational Collective Intelligence (pp. 328-339). Springer, Cham.

16) Kaur, H., Singh, S., & Aggarwal, N. (2020). Demand forecasting models with time series and random forest. In Handbook of Research on Demand Management and Demand Forecasting for Supply Chain Optimization (pp. 66-86). IGI Global.

17) Moyo, T., Chikuni, E., & Chikuni, P. (2018). A comprehensive study of random forest for short-term load forecasting. Energies, 15(5), 7547.

18) Kumar, A., Singh, S.K., Gupta, A. and Garg, R., 2019. A comparative analysis of machine learning techniques for forecasting retail sales. International Journal of Business Forecasting and Marketing Intelligence, 5(3), pp. 227-244.

19) Li, Y., Wang, J., Liu, X. and Zhang, J., 2020. A hybrid model based on machine learning for short-term sales forecasting. Journal of Retailing and Consumer Services, 54, p. 101922.

20) Zhang, Y., Zhang, J., Zhang, J., Liu, Y. and Li, Y., 2020. A novel hybrid model based on machine learning and optimization for sales forecasting. Expert Systems with Applications, 161, p. 113676.

21) Chen, Y., Li, Y., Zhang, J. and Liu, Y., 2019. A random forest approach for sales forecasting in a small convenience store. Journal of Retailing and Consumer Services, 48, pp. 1-7.

22) Mhlanga, T., Mbohwa, C. and Muzenda, E., 2020. Demand forecasting for small-scale retailers in Zimbabwe using machine learning. In: 2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). IEEE, pp. 1339-1343.

23) Zhou, Y., Li, Y., Zhang, J. and Liu, Y., 2020. A random forest approach for sales forecasting in a small supermarket. Journal of Retailing and Consumer Services, 55, p. 102123.

Appendices

Appendix A:

Git-hub Repository: https://github.com/Revelation-coder/Retail-Sales-Forecasting.git

final

ORIGINALITY REPORT

14%	12%	1%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	fastercapital.com Internet Source	2%
2	www.researchsquare.com Internet Source	<1%
3	satprnews.com Internet Source	<1%
4	www.tumblr.com Internet Source	<1%
5	Submitted to Kepler College Student Paper	<1%
6	pure.hud.ac.uk Internet Source	<1%
7	Submitted to American Public University System Student Paper	<1%
8	medium.com Internet Source	<1%
9	riunet.upv.es Internet Source	<1%