Forecasting Retail Consumer Purchase Trends in Kenya Using Machine Learning Algorithms

By

Claudine Linda Wa Nciko

169375

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 and iLab Africa

Strathmore University

Nairobi, Kenya

June, 2025

Declaration and Approval

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 thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Claudine Linda Wa Nciko

Sign: _____

Date: 22 May,2025

Approval

The thesis of **Claudine Linda Wa Nciko** was reviewed and approved by the following:

Dr. John Olukuru,

Senior Lecturer, Data Science and Analytics,

Strathmore Univeristy.

Prof. Godfrey Madigu,

Dean, Institute of Mathematical Sciences,

Strathmore University.

Prof. Bernard Shibwabo,

Director of Graduate Studies,

Strathmore University.

Abstract

Understanding consumer purchasing behavior is essential for retailers in Kenya, particularly in the dynamic context shaped by evolving socio-economic factors and the impacts of the COVID-19 pandemic. This study aimed to analyze consumer purchasing patterns by leveraging advanced machine learning algorithms, including Random Forest (RF), Extreme Gradient Boosting (XGBoost), k-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN), using a multi-output approach to predict multiple consumer behavior variables simultaneously. The analysis utilized survey data collected from 2020 to 2024, focusing on demographic and economic variables such as gender, age group, occupation, and income level, alongside purchasing trends across diverse product categories, including hygiene products, groceries, and electronics. The improved Random Forest model, optimized through hyperparameter tuning, emerged as the best-performing model with an accuracy of 78% for predicting age group and 73% for income level. Confusion matrix analysis revealed the model's proficiency in identifying dominant classes while highlighting challenges in distinguishing minority classes, particularly within occupation and income categories. Additionally, SHAP (SHapley Additive exPlanations) analysis demonstrated a balanced feature contribution, reducing the risk of overfitting and enhancing model interpretability. To facilitate practical application, the model was deployed using Streamlit, enabling real-time predictions and interactive data exploration for stakeholders. The study's findings provide significant insights into the key economic and demographic factors influencing purchasing behavior in Kenya. By integrating data-driven analytics into retail decision-making, businesses can optimize inventory management, targeted marketing, and strategic planning. This research contributes to bridging the gap between academic analysis and practical retail applications, offering a replicable framework for similar contexts in emerging markets.

Keywords: Consumer Purchasing Behavior, Machine Learning, Predictive Modeling, Multi-Output, Decision-Making.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
AML	Anti-Money Laundering
ANN	Artificial Neural Network
API	Application Programming Interface
AUC	Area Under the Curve
AUC-ROC	Area Under the Curve - Receiver Operating Characteristic
CAPI	Computer-Assisted Personal Interviewing
CRM	Customer Relationship Management
CSS	Cascading Style Sheets
CSV	Comma-Separated Values
CRISP-DM	Cross-Industry Standard Process for Data Mining
EDA	Exploratory Data Analysis
F1-Score	Harmonic Mean of Precision and Recall
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GitHub	Web-based platform for version control using Git
GridSearchCV	Grid Search with Cross-Validation
joblib	Python library for lightweight pipelining and object persistence
KNN	k-Nearest Neighbors
LFS	Large File Storage
ML	Machine Learning
RF	Random Forest

SHAP SHapley Additive exPlanations

SMOTE Synthetic Minority Over-sampling Technique

Streamlit Python library for creating web apps for data science and machine learning

TP True Positive

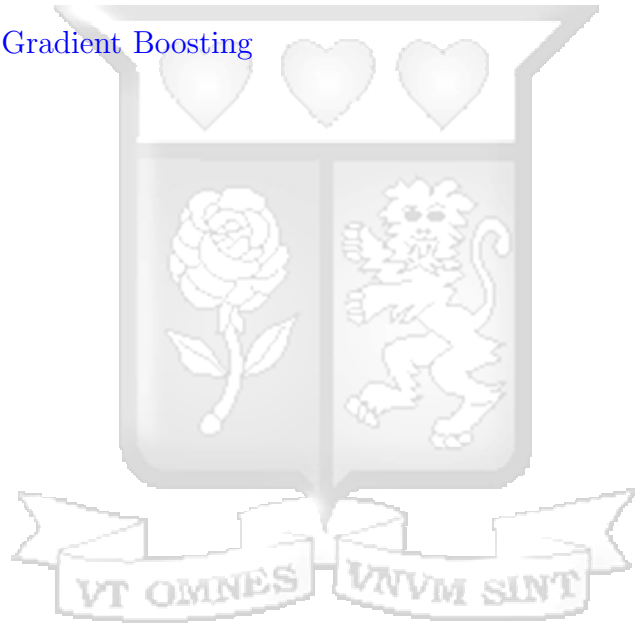
TN True Negative

TPR True Positive Rate

UI User Interface

VSCode Visual Studio Code

XGBoost Extreme Gradient Boosting



Acknowledgments

First and foremost, I give all glory and honor to the Almighty God for granting me the strength, wisdom, and perseverance to complete this dissertation. Without His guidance and grace, this journey would not have been possible.

I would like to express my deepest gratitude to my dissertation supervisor, Dr. John Olukuru, for his unwavering support, insightful guidance, and invaluable feedback throughout this research journey. His expertise and encouragement have been instrumental in shaping this dissertation to its completion.

I am immensely grateful to Strathmore University for fostering an intellectually stimulating environment and providing essential resources that significantly contributed to the success of this study. My heartfelt appreciation also goes to all the lecturers and academic staff who supported me throughout my master's program, as well as to my cohort members for their encouragement and collaboration.

Lastly, I would like to extend my deepest appreciation to my family, whose unwavering support has been a constant source of motivation, and to my friends, classmates, and all silent supporters who stood by me throughout this journey. Your encouragement and belief in me made this accomplishment possible.

Chapter 1: Introduction

1.1 Background

Consumer behavior is a fundamental area of study in retail and marketing research, as it directly influences inventory management, pricing strategies, and customer engagement. Understanding consumer purchasing patterns enables businesses to anticipate market demand and optimize their operations accordingly. Consumer behavior is shaped by a range of psychological, social, and economic factors that drive decision-making processes(Marketing91, 2023). A consumer is generally defined as "a person who buys goods or services for their own use"(Cambridge Dictionary, 2024). Various factors, such as individual preferences, financial constraints, and convenience, affect purchasing behavior, ultimately determining the success of retail businesses. Gaining insights into these behaviors allows retailers to refine their product offerings, adjust pricing strategies, and improve marketing approaches, ensuring alignment with customer expectations and market dynamics.

1.2 Global Retail Sector Overview

The retail sector plays a crucial role in the global economy by providing essential goods and services to consumers. While the industry has witnessed rapid digital transformation, physical retail stores continue to dominate various markets, offering both general and specialized products. Key trends shaping the global retail sector include the rise of e-commerce, the expansion of franchise models, and increasing consumer demand for sustainable business practices(Statista, 2024c). Additionally, data-driven strategies have become essential for retailers seeking to enhance customer engagement and maintain competitiveness in the evolving marketplace.

1.2.1 The Retail Supply Chain

The modern retail supply chain is complex and continuously evolving, with multiple touchpoints influencing consumer purchasing behavior. Industries such as fashion, food, and healthcare retailing have adapted to changing consumer habits, leading to substantial market shifts. For instance, the global apparel market generated $1.73 trillion in revenue in 2023 and is projected to grow significantly(Statista, 2024a). The increasing digitization

of retail supply chains has further enhanced the efficiency of inventory management and order fulfillment. Figure 1.1 below highlights the apparel market's growth, reaching $1.73 trillion in 2023.

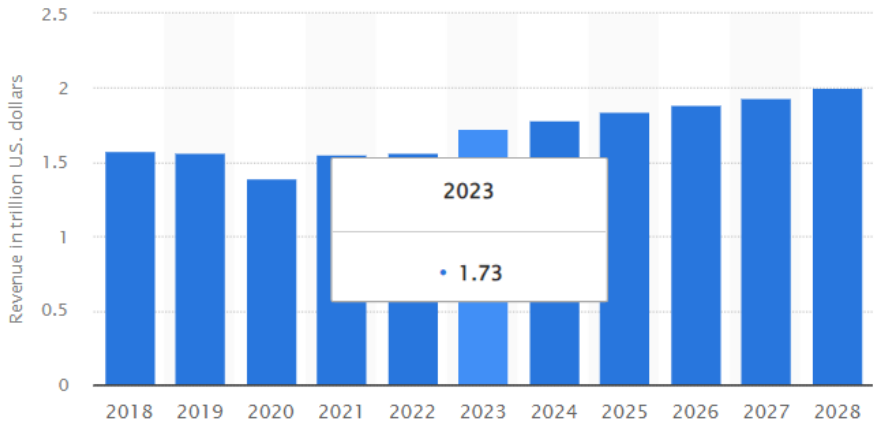


Figure 1.1: Revenue of the apparel market worldwide from 2018 to 2028 (in trillion U.S.D.)(Statista, 2024a)

1.2.2 Technological Advancements in Retail

Advancements in artificial intelligence Artificial Intelligence (AI), big data analytics, and cloud computing have significantly impacted the retail landscape. These technologies have enabled retailers to enhance customer experiences, streamline operations, and improve predictive analytics. Investment in retail technology surged to $4.7 billion in the last quarter of 2023, reflecting the growing importance of digital transformation(Statista, 2024b). Figure 1.2 below illustrates retail tech investments soaring to 4.7 billion U.S. dollars in deals during the final quarter of 2023.

While global retailers have leveraged these advancements to optimize business operations, the Kenyan retail industry presents unique opportunities and challenges that necessitate localized strategies and analytical models.

1.3 Retail Landscape in Kenya

Kenya's retail market, valued at approximately $90 billion in 2023, reflects a blend of traditional and modern retail practices(Cytonn, 2023). The informal sector, which includes small shops, kiosks, and open-air markets, remains a dominant force in consumer trade.

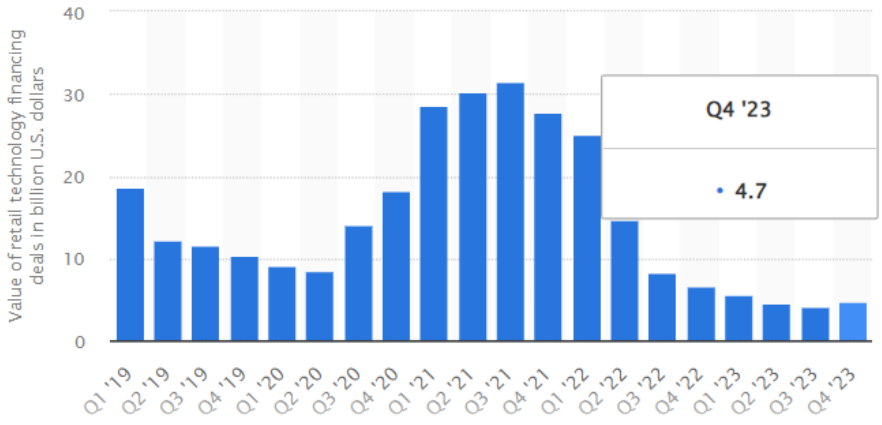


Figure 1.2: Quarterly value of retail technology financing deals worldwide from 1st quarter 2019 to 4th quarter 2023 (U.S.D.) (Statista, 2024b)

However, rapid urbanization increased digital connectivity, and shifting consumer expectations have fueled the expansion of supermarkets and e-commerce platforms**?**. Kenya's e-commerce market is projected to reach $2 billion by 2024, driven by increased internet penetration and mobile payment adoption.

Consumer purchasing decisions in Kenya are influenced by multiple socio-economic factors, including income levels, cultural dynamics, and digital accessibility(Competition Authority of Kenya, 2018). Additionally, global economic disruptions, such as inflation and supply chain bottlenecks, have affected consumer spending patterns, with a noticeable shift toward essential goods. As affordability becomes a priority, retailers that offer cost-effective solutions and convenient shopping experiences are gaining prominence in the Kenyan market.

The proliferation of mobile money platforms, particularly M-Pesa, has revolutionized retail transactions(Kasi Insight, 2024). Digital payments now facilitate seamless transactions across formal and informal retail sectors, enhancing financial accessibility and consumer convenience. Furthermore, digital engagement through social media and online marketplaces is shaping purchasing trends, necessitating new retail strategies that incorporate online behavior analytics.

1.4 Problem Statement

Retailers in Kenya have been facing increasing challenges in predicting consumer purchasing behavior due to economic fluctuations, inflationary pressures, and shifts in consumer priorities. The unpredictability of purchasing trends has led to inventory mismatches, inefficient marketing strategies, and difficulties in demand forecasting. Traditional forecasting methods, which primarily rely on historical sales data, have proven inadequate in capturing rapidly evolving consumer preferences and behaviors (Cytonn, 2022).

The complexity of Kenya's retail sector, encompassing both formal and informal trade structures, has further compounded demand forecasting efforts. Although digital payment systems and mobile money platforms have transformed transaction processes, retailers have struggled to effectively utilize the vast amounts of data available to anticipate consumer behavior. Moreover, integrating diverse data sources, such as transactional records, digital payments, consumer surveys, and macroeconomic indicators, remains a significant challenge, limiting retailers' ability to make data-driven decisions (Kasi Insight, 2024).

As consumer preferences continue to shift in response to economic changes, retailers are in dire need of reliable methods to understand and predict purchasing behavior. The absence of accurate predictive models has made it increasingly difficult for businesses to optimize inventory management, adjust pricing strategies, and enhance customer engagement.

This study addressed this problem by identifying key economic and demographic factors influencing consumer purchasing behavior in Kenya, utilizing survey data from Kasi Insight. It employed advanced machine learning algorithms, including Random Forest, XGBoost, Artificial Neural Networks (ANN), and K-Nearest Neighbors (KNN), to develop predictive models that enhance demand forecasting accuracy. These models were designed to provide actionable insights for retail decision-making, thereby enabling more informed inventory and marketing strategies.

1.5 Research Aim

The primary aim of this study was to analyze consumer purchase behavior in Kenya using advanced machine learning algorithms to identify the key factors influencing purchasing trends. This research provided insights that support better decision-making for retailers

in the Kenyan market.

1.6 Research Objectives

This research aimed to address the following objectives:

I. To analyze the key economic and demographic factors ('Gender', 'Age-group', 'Occupation', 'Income Level') influencing consumer purchasing behavior in Kenya.

II. To develop and validate predictive models using machine learning algorithms (RF, KNN, XGBoost, ANN).

III. To deploy the predictive model for practical use by Kenyan retailers, enabling them to optimize inventory management and marketing strategies.

1.7 Research Questions

I. What are the key economic and demographic factors influencing consumer purchasing behavior in Kenya from 2020 to 2024?

II. How accurately can machine learning models predict consumer purchasing behavior based on these factors, and what level of accuracy is achievable?

III. How can a predictive model be effectively deployed as an interactive tool to enhance real-time retail decision-making for Kenyan retailers?

1.8 Scope and Limitations of the Study

1.8.1 Scope

This study aimed to analyze and predict consumer purchasing behavior in Kenya using advanced machine learning (Machine Learning (ML)) algorithms. The primary focus was on the Kenyan retail sector, particularly in urban and peri-urban areas where digital transactions and data collection are prevalent. The study utilized data collected between 2020 and 2024, covering consumer behavior before, during, and after the COVID-19 pandemic.

Key demographic and behavioral factors analyzed included age, gender, income level, occupation, marital status, and purchasing trends across various fast-moving consumer goods (FMCG), hygiene products, groceries, electronics, and other essential consumer

products. The study leveraged machine learning techniques, including Random Forest (RF), Extreme Gradient Boosting (XGBoost), k-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN), to develop predictive models that identify key economic and demographic factors influencing purchasing behavior.

The study employed the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, guiding the research from data collection through model development, validation, and deployment. The final model, based on an optimized Random Forest algorithm, was deployed using **Streamlit**, providing a practical, interactive tool for real-time predictions and insights. This deployment aimed to bridge the gap between academic research and practical application, enabling Kenyan retailers to make data-driven decisions on inventory management, targeted marketing, and customer engagement.

1.8.2 Limitations

The study's predictive accuracy heavily relied on the quality and completeness of survey-based data from Kasi Insight, which may not capture all behavioral patterns and preferences, especially those influenced by rapid socio-economic changes. Additionally, computational constraints limited the exploration of more sophisticated deep learning models that could potentially enhance prediction accuracy.

The geographical scope of the study was limited to Kenya, particularly urban and peri-urban areas, where digital transaction data was readily available. Consequently, the findings may not be generalizable to rural settings or regions with differing retail dynamics. Moreover, while the model's deployment through **Streamlit** offers practical usability, it does not fully integrate into existing retail management systems, which could limit its operational effectiveness in real-world retail environments.

Furthermore, the study did not fully account for long-term behavioral shifts driven by socio-economic factors or significant external disruptions beyond the COVID-19 pandemic period. Consequently, the model's applicability in future contexts or regions with distinct socio-economic profiles may be limited.

1.9 Research Justification

The study aimed to enhance data-driven decision-making within the Kenyan retail sector by leveraging machine learning algorithms to predict consumer purchasing behavior. By providing actionable insights, the research empowered retailers to optimize inventory management, pricing strategies, and targeted marketing. The predictive models developed in this study addressed the challenge of demand uncertainty, enabling more accurate forecasting and reducing inefficiencies such as stockouts or excess inventory.

From an academic perspective, this study contributes to the emerging field of machine learning applications in retail analytics within developing markets. While existing studies have primarily focused on developed economies or single-output models, this research utilized advanced multi-output predictive modeling techniques tailored to Kenya's unique retail environment. The successful deployment of the predictive model using **Streamlit** demonstrates a practical and replicable framework for similar contexts across Africa and other developing regions.

Furthermore, the increasing adoption of digital transactions in Kenya justified the study's focus, as retailers increasingly rely on data-driven insights to maintain competitiveness. By offering a robust methodology and practical implementation, this study serves as a valuable reference for both academic research and practical retail strategy development in emerging markets.

Chapter 2: Literature Review

2.1 Introduction

In the rapidly evolving retail landscape of Kenya, understanding consumer purchasing behavior is essential for retailers seeking to remain competitive. This chapter presents a comprehensive literature review on the predictive analysis of consumer purchasing trends, emphasizing the role of machine learning algorithms. The review is divided into theoretical and empirical perspectives. The theoretical review explores foundational concepts and models in consumer behavior, while the empirical review examines real-world applications of these theories in retail analytics.

2.2 Theoretical Review

2.2.1 Introduction to Consumer Behavior

Consumer behavior refers to the study of individuals, groups, or organizations and the processes they use to select, purchase, use, and dispose of goods or services. It integrates insights from multiple disciplines, including psychology, sociology, and economics, to understand the cognitive, emotional, and behavioral responses that influence purchasing decisions(Tan, 2010). Consumer purchasing behavior typically follows a structured process involving problem recognition, information search, evaluation of alternatives, purchase decision, and post-purchase evaluation(Morrison, 2015). Several internal and external factors, such as social influences, marketing strategies, and personal preferences, affect this process. Understanding these influences allows businesses to develop tailored marketing approaches that enhance customer engagement and retention(Tan, 2010).

2.2.2 Traditional Consumer Behavior Theories

Several traditional theories provide a foundation for understanding consumer purchasing decisions. Economic-based models, such as Marginal Utility Theory and Indifference Theory, propose that consumers make rational choices aimed at maximizing utility while balancing their preferences and financial constraints(Manuere et al., 2022). These models assume that purchasing behavior is predictable and driven by logical decision-making processes.

Maslow's Hierarchy of Needs(Zhang et al., 2023)(Meshram, 2023) introduces a psychological perspective, explaining how different levels of human needs, ranging from basic necessities to self-actualization, affect purchasing behavior. Understanding consumer needs within this hierarchy allows retailers to design strategies that align with their target audiences. Figure 2.1 illustrates Maslow's Hierarchy of Needs(CopyMate, 2024).

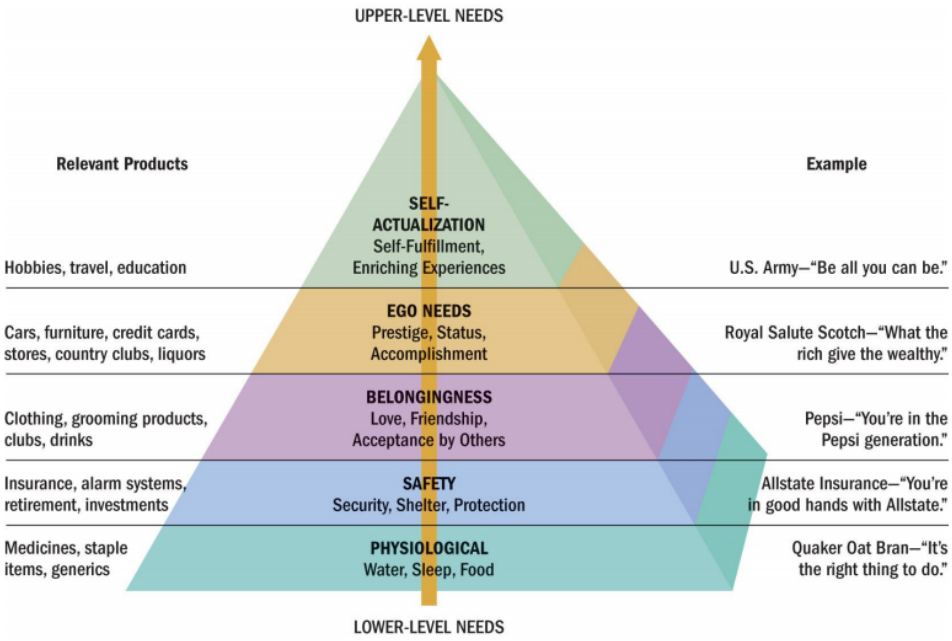


Figure 2.1: Maslow's Motivation-Need Theory.(CopyMate, 2024)

2.2.3 Modern Consumer Behavior Theories

Contemporary consumer behavior theories integrate psychological and sociological factors, acknowledging that purchasing decisions are influenced by emotions, habits, and external stimuli(Zendesk, 2024). The Theory of Planned Behavior suggests that consumer intentions are shaped by attitudes, perceived behavioral control, and subjective norms, offering insights into how external factors influence purchasing habits(**?**).

The phenomenon of impulse buying further highlights the complexity of consumer decisions, with studies suggesting that up to 80% of consumer spending is driven by spontaneous purchases influenced by store layout, advertisements, and psychological triggers(Authors, 2021). Understanding these mechanisms enables retailers to design more effective marketing and promotional strategies.

2.2.4 Impact on Retail Decision-Making

Applying consumer behavior theories to retail decision-making allows businesses to develop data-driven strategies tailored to consumer preferences. Different retail formats, such as supermarkets, discount stores, and e-commerce platforms, significantly influence purchasing patterns(Manuere et al., 2022). Retailers that effectively leverage insights from consumer behavior models can optimize inventory management, pricing strategies, and customer segmentation to drive sales and maintain a competitive advantage(Mou et al., 2018).

2.2.5 Behavioral Models in Retail

Several behavioral models provide structured frameworks for analyzing consumer purchasing behavior. The Engel, Blackwell, and Miniard Framework and the Howard-Sheth Model incorporate psychological and social contexts to explain consumer decision-making(Ashman et al., 2015). These models map out the stages of consumer decision-making, from recognizing needs to post-purchase evaluation, helping retailers predict customer preferences and tailor their marketing strategies.

The Stimulus-Response Model further emphasizes how external factors, such as advertising and promotional campaigns, influence purchasing decisions(Roy and Datta, 2022). Retailers who understand these behavioral models can create targeted campaigns that resonate with their audience, leading to increased sales and brand loyalty.

2.2.6 The Role of COVID-19 in Shaping Consumer Behavior

The COVID-19 pandemic significantly altered consumer purchasing patterns, accelerating the shift toward digital commerce and emphasizing the importance of essential goods. Consumer behavior models, such as Maslow's Hierarchy of Needs, provide a framework for understanding these shifts, as financial uncertainty led many consumers to prioritize basic necessities(Gupta et al., 2023). Additionally, retailers were compelled to adopt digital strategies to meet evolving consumer demands, highlighting the need for robust data analytics and predictive models.

This theoretical review has explored key consumer behavior theories and models that underpin purchasing decisions. Traditional economic models emphasize rational decision-

making, while modern psychological and sociological theories provide deeper insights into impulsive and habitual purchases. Behavioral models offer structured frameworks for analyzing consumer decision-making, enabling retailers to develop data-driven strategies. The insights gained from this review will inform the development of predictive models that integrate machine learning techniques to analyze consumer purchasing trends. By applying these theoretical foundations, this study aims to enhance demand forecasting and optimize retail decision-making in Kenya.

2.3 Empirical Literature Review

2.3.1 Analyzing Key Factors Influencing Consumer Purchasing Behavior

Machine learning models have increasingly been employed to analyze consumer purchasing behavior, offering insights into economic trends, demographic attributes, and digital payment adoption. Hu (2022) examined decision tree algorithms and recurrent neural networks (RNNs) for predicting purchasing patterns, achieving a 79.3% accuracy rate. However, the study relied on datasets from developed markets, which limits its generalizability to Kenya's hybrid retail environment where informal and formal segments coexist. Additionally, while RNNs offer temporal prediction capabilities, the study did not address deployment or local applicability, which weakens its relevance for African retailers.

Hindawi (2023) evaluated various models including Decision Trees, Random Forest, XGBoost, and Neural Networks, identifying XGBoost as the most accurate (87%). Yet, the study's dataset was sourced from structured e-commerce platforms with consistent behavioral tracking, in contrast to Kenya's offline, dominated and fragmented data landscape. This highlights a gap in adapting machine learning to contexts with limited digital traceability.

Majeed et al. (2022) provided a qualitative review of Big Data's impact on African marketing, emphasizing the role of IoT and business intelligence. However, the absence of predictive modeling and machine learning techniques renders the findings more conceptual than actionable.Ajiga et al. (2024) applied AI-driven analytics using NLP and computer vision but omitted empirical evaluation of model performance, reducing its value for methodological benchmarking. Similarly, Necula (2023) used clickstream data to achieve high prediction accuracy in digital settings but ignored socio-economic factors

and offline behavior patterns, which are critical in Kenya.

Overall, these studies underscore the potential of AI in consumer behavior research but fall short in addressing informal economies, low digital penetration, and multi-category purchase prediction. This study bridges that gap by integrating demographic, economic, and behavioral indicators in a context-specific, multi-output framework tailored to Kenya's retail sector.

2.3.2 Developing and Validating Predictive Models Using Machine Learning

Wambugu (2017) investigated how personal characteristics affect consumer behavior toward international ethnic cuisines in Nairobi using Multiple Linear Regression (MLR). The study identified gender, income, education, occupation, and lifestyle as significant predictors. However, it lacked generalizability beyond Nairobi and did not apply non-linear or ensemble-based machine learning models. While the high R^2 value (0.801) indicates strong model fit, the exclusion of advanced model optimization and deployment strategies limits practical impact.

In contrast, Ghosh and Banerjee (2020) developed a Modified Random Forest model for predicting cloud service purchases using sequential ad click data. With an accuracy of 87.02%, the model highlights the predictive strength of tree-based ensembles. However, the domain specificity (cloud services) and structured online logs limit transferability to Kenya's hybrid retail systems, which lack consistent digital records.

Martínez et al. (2020) used Gradient Boosting to outperform classical regression in predicting customer lifetime value through RFM variables. Although suitable for contract-based retail in developed settings, the model does not account for irregular spending patterns common in emerging markets.

Azad et al. (2023) integrated psychological theory (Theory of Planned Behavior) with Gradient Boosting to model decision-making. While their model explains 91% of variance in behavior, its dependence on social media data restricts applicability in contexts with low digital engagement like rural Kenya.

This study improves upon previous work by using ensemble learning (RF, XGBoost), instance-based (KNN), and deep learning (ANN) models to capture non-linearity and

feature interactions. It also introduces a multi-output framework suited for predicting interdependent retail outcomes, age group, income level, occupation, and gender in a context where such segmentation is operationally relevant.

2.3.3 Deploying Interactive Tools for Data-Driven Decision-Making

Valecha et al. (2019) combined Random Forest and Logistic Regression for demographic-based purchase prediction, achieving a 94% accuracy. While effective, their approach was limited to static analysis without real-time deployment, and assumed access to consistently labeled demographic and product datasets, which may not be available in Kenya.

Esmeli et al. (2021) proposed an Early Purchase Prediction (EPP) model using session logs and achieved strong AUC performance. However, this model depends on high-resolution e-commerce browsing data, making it difficult to adapt for informal or offline settings typical in Kenyan retail.

Bundi (2023) applied K-Means and LightGBM to customer churn, achieving a 96% precision rate. Though locally relevant, the study focuses on churn rather than proactive purchase forecasting. Your work builds on this by deploying models for anticipating future demand across product categories.

Otieno (2024) used Apriori and FP-Growth for identifying commonly purchased items, but did not employ predictive techniques for forecasting future behavior, limiting their contribution to real-time decision-making.

This study goes further by deploying a multi-output predictive engine in a user-friendly Streamlit interface, enabling local retailers to make data-informed inventory and marketing decisions. Unlike previous studies, it emphasizes interpretability (via SHAP), operationalization (via joblib exports), and domain specificity (Kenyan retail).

2.4 Research Gaps

Despite the growing body of literature on consumer analytics and machine learning, existing research exhibits several critical limitations that this study aims to address. A prominent gap lies in the contextual mismatch between prior studies and the Kenyan retail environment. Most of the referenced works are based in high-income or digital-first economies, utilizing structured and consistent datasets from e-commerce platforms. In

contrast, Kenya's retail sector is characterized by fragmented data, informal channels, and limited digital infrastructure, which reduces the applicability of such models without contextual adaptation.

Another gap concerns modeling strategy. Many prior studies apply linear models or single-output classifiers, which limit their ability to capture interdependencies between multiple demographic or behavioral outcomes. This study introduces a multi-output modeling approach to simultaneously predict income level, gender, age group, and occupation, offering a more holistic view of retail consumer segments.

There is also a methodological gap in how previous studies evaluate and deploy models. While several works demonstrate high accuracy, few incorporate model explainability, generalization checks across folds, or deployment frameworks that translate models into real-world applications. This research addresses that by combining explainable AI techniques (e.g., SHAP), robust evaluation protocols (e.g., cross-validation, macro/micro F1 scores, AUC), and a deployment pipeline using Streamlit and joblib.

Lastly, many studies neglect the behavioral nuance in feature engineering. This study integrates behavioral signals, economic indicators, and demographic variables to reflect Kenya's diverse and evolving consumer landscape. In doing so, it ensures that the resulting model is not only statistically sound but practically relevant for local decision-makers.

2.5 Conceptual Framework

A conceptual framework illustrates the anticipated relationships between variables, clarifies the study's objectives, and supports the development of coherent conclusions (Swaen and George, 2024). In this research, the framework is constructed to explore consumer purchasing behavior in Kenya through the application of advanced machine learning (Anti-Money Laundering (AML)) algorithms. The central aim is to identify key demographic determinants of purchasing trends and to apply predictive analytics in support of informed retail decision-making.

The independent variables consist of demographic attributes, including age group, gender, occupation, and income level. These variables are hypothesized to have a significant influence on purchasing decisions and are used as input features in the machine learning models.

The dependent variables represent consumer behavior across various product categories, such as personal hygiene, cleaning products, groceries, electronics, entertainment, clothing, and airtime or data bundles. The objective is to model and predict these purchasing behaviors based on the demographic profiles of consumers.

To achieve this, the study employs a suite of machine learning algorithms: Random Forest (Random Forest (RF)), Extreme Gradient Boosting (XGBoost), Artificial Neural Networks (Artificial Neural Network (ANN)), and K-Nearest Neighbors (k-Nearest Neighbors (KNN)). These models are capable of capturing non-linear relationships and are well-suited for multi-output classification tasks, making them effective tools for uncovering patterns in complex consumer datasets.

The predictive insights generated by these models contribute to enhanced retail operations. Specifically, the outputs can support decisions related to inventory planning, pricing strategies, and targeted marketing efforts. By anticipating consumer demand more accurately, retailers operating in Kenya's hybrid retail environment can improve efficiency, responsiveness, and competitiveness. An overview of the conceptual framework is illustrated in Figure 2.2.

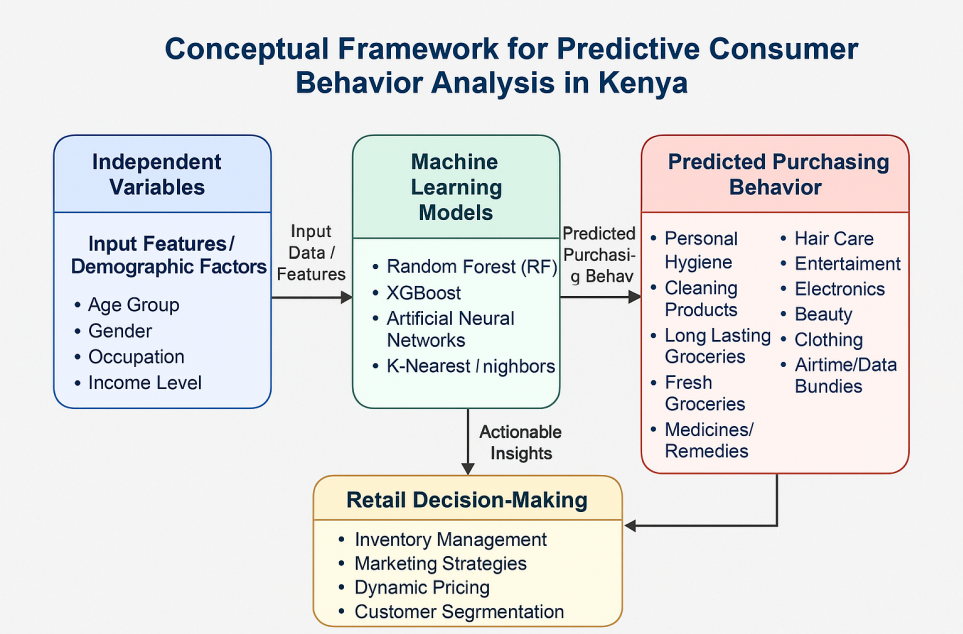


Figure 2.2: Conceptual Framework.

Chapter 3: Research Methodology

3.1 Introduction

This chapter outlines the methodology used to develop a predictive model for analyzing consumer purchasing behavior in Kenya. The methodology followed the Cross-Industry Standard Process for Data Mining Cross-Industry Standard Process for Data Mining (CRISP-DM) framework(Data Science Project Management, 2024), which provided a structured approach for data mining and machine learning applications. The study employed multi-output machine learning models to examine how demographic and behavioral factors influenced purchasing behavior. This chapter details each phase of the CRISP-DM framework, including data understanding, preprocessing, modeling, evaluation, and deployment.

3.2 Research Design

The CRISP-DM framework consisted of six distinct phases, guiding the entire research process. The first phase, Business Understanding, focused on defining the study's objectives and establishing key requirements from a research perspective. The second phase, Data Understanding, involved acquiring and exploring datasets, assessing their structure and content, and identifying potential inconsistencies or missing values. Insights gained from this phase informed data preparation and feature selection strategies. The third phase, Data Preparation, involved preprocessing and transforming the collected data to enhance its quality and usability for machine learning models. This step included data cleaning, encoding categorical variables, and implementing feature selection techniques to optimize model performance. The fourth phase, Modeling, involved selecting and applying suitable machine learning algorithms to analyze purchasing behavior based on demographic and behavioral features. The subsequent phases, Evaluation and Deployment, focused on assessing the model's predictive performance and developing an interactive system for delivering insights to stakeholders. The systematic implementation of these CRISP-DM phases ensured rigorous data analysis and the generation of actionable insights within the Kenyan retail sector. Figure 3.1 presents the CRISP-DM framework (Hotz, 2024).

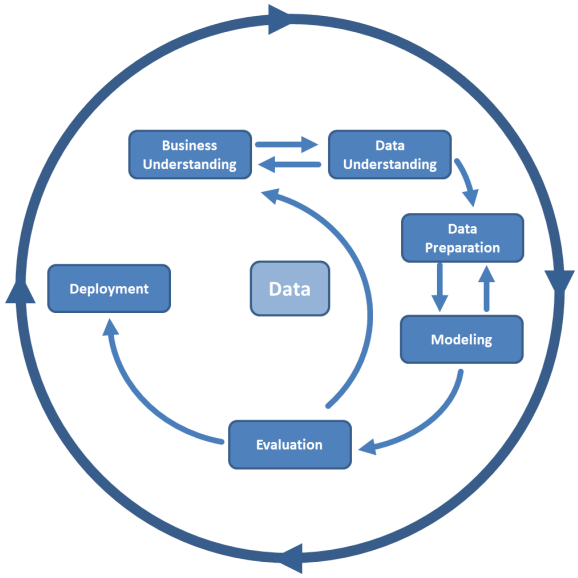


Figure 3.1: CRIPS DM (Hotz, 2024)

3.3 Data Sources

The data for this study was obtained from the "Consumer Purchase Trends" survey conducted by Kasi Insight, a reputable market research firm specializing in providing real-time, aggregated, and trended primary data across Africa. Kasi Insight's mission is to empower decision-makers with the context, insights, and foresight needed to succeed in the African market. The survey covered five consecutive years from 2020 to 2024, capturing purchasing habits during various months: April to December 2020, April, July, and October 2021, February, June, and November 2022, March, July, and November 2023, and March, July, and December 2024.

The data was collected using the Computer-Assisted Personal Interviewing Computer-Assisted Personal Interviewing (CAPI) method, a structured and standardized in-person research technique. Trained interviewers utilized digital tools, such as tablets or computers, to record survey responses directly from participants(Elliott, 2021). This approach ensured high accuracy by minimizing errors associated with manual data entry and enabling real-time validation of responses through built-in logic checks and skip patterns. Additionally, CAPI facilitated real-time monitoring of the survey process, allowing for immediate quality control and adjustments during fieldwork. These features significantly enhanced data reliability and operational efficiency, making CAPI a preferred method for large-scale surveys. Studies such as those conducted by (Schräpler et al., 2019) have

demonstrated that CAPI not only reduces routing errors and inconsistencies but also improves the overall quality and completeness of collected data compared to traditional paper-based methods.

Following data collection, the data was systematically cleaned, combined, and securely stored in a centralized data portal (Kasi Platform) for analysis, ensuring transparency and accessibility for researchers.

The dataset comprised 10,530 rows and 19 columns, each representing a variable related to demographic information or purchasing behavior. Consumer purchasing behavior in the past month was categorized across various product types, including personal hygiene, cleaning products, and more. Responses were classified using a rating system of 'More,' 'Same,' 'Less,' or 'Don't Know/Don't Buy,' allowing for systematic tracking of changes in consumer preferences over time. The 'Period' variable marked the time of data collection, enabling temporal analysis. The table below provides a detailed description of each variable included in the dataset.

3.4 Data Preprocessing

Data preprocessing was a crucial step in this research, as it ensured the accuracy, consistency, and reliability of the data. This phase involved cleaning, organizing, and transforming data into the appropriate format for modeling. The main steps in the data preprocessing pipeline included data cleaning, inspection, handling missing values, and the removal of duplicates.

3.4.1 Data Cleaning

Data cleaning involved addressing missing values, removing duplicate records, and correcting inconsistencies to ensure that the dataset was of high quality and reliable for subsequent analysis. The primary goal was to create a consistent and error-free dataset by identifying and resolving anomalies and inaccuracies.

Data Inspection The data inspection process involved examining the dataset to understand its structure, attributes, and underlying patterns. This step entailed computing summary statistics such as mean, median, and standard deviation to detect potential anomalies, identify data entry errors, and uncover potential sources of noise. This com-

Table 3.1: Description of Variables

Variable Name	Description
Period	The time frame of data collection: 2020 (April - December), 2021 (April, July, October), 2022 (February, June, November), 2023 (March, July, November), 2024 (March, July, December).
Location	Various locations coded as AB15, AB2, AB25, etc.
Gender	Gender of the consumer (Male, Female).
Marital Status	Marital status of the consumer (Single, Dating, Married, Cohabiting, Divorced/Separated/Widowed).
Age Group	Age group of the consumer (Gen X, Millennials, Gen Z, Baby Boomers).
Occupation	Occupation of the consumer (Business owner, Salaried employee, Self-employed/Contractor, Unemployed, Student, Stay-at-home parent).
Income Level	Income level of the consumer (High Income, Middle Income, Low Income).
Personal Hygiene	Purchases of personal hygiene products (e.g., Soap, toothpaste).
Cleaning Products	Purchases of cleaning products.
Long Lasting (Dry) Groceries	Purchases of long-lasting grocery items (e.g., Rice, pasta, canned goods).
Fresh Groceries	Purchases of fresh groceries (e.g., Fruits, vegetables).
Medicines and Natural Remedies	Purchases of medicines and natural remedies.
Alcoholic Beverages	Purchases of alcoholic beverages.
Skin Care	Purchases of skin care products (e.g., Body lotion).
Hair Care	Purchases of hair care products (e.g., Shampoo).
Entertainment	Purchases of entertainment services (e.g., Restaurants, movies).
Electronics	Purchases of electronics (e.g., Phone, computers).
Beauty	Purchases of beauty products (e.g., Makeup, cosmetics, haircuts).
Clothing	Purchases of clothing items.
Airtime and Data Bundles	Purchases of airtime and data bundles.

prehensive examination of the data's quality and structure guided the selection of appropriate cleaning methods.

3.4.2 Handling Missing Values

Missing values were systematically inspected and categorized to understand their nature and address them appropriately. In this study, the missing values primarily occurred in the 'Income Level' column, with a total of 2,395 missing entries. Since the 'Income Level' variable was categorical, the most suitable approach was to use mode imputation to fill in the missing values. Mode imputation is effective for categorical data as it replaces missing values with the most frequently occurring category, thereby maintaining data consistency.

Unlike continuous data where mean imputation is commonly applied, categorical data requires mode imputation to preserve the dataset's statistical properties. Therefore, mode imputation was chosen over other imputation methods due to the categorical nature of the affected column.

Table 3.2: Treatment Techniques of Missing Values (Van Otten, 2024)

Method	**Description**
Listwise Deletion	Removes observations with missing values if the data is MCAR and the proportion of missing data is small.
Pairwise Deletion	Uses all available data for each analysis without discarding entire observations, suitable for MAR data.
Mean/Median/Mode Imputation	Replaces missing numerical values with the mean or median of the column, or with the mode for categorical data.

By employing mode imputation, the study ensured that the dataset maintained its categorical integrity and remained statistically consistent for subsequent analyses.

3.4.3 Removal of Duplicates

To ensure data integrity, duplicate entries were carefully examined and removed from the dataset. Duplicate records can arise due to data entry errors or redundant data collection, leading to inconsistencies and potential biases in the analysis. The presence of duplicate rows was detected using the following Python command: `data.duplicated().sum()`. This analysis revealed a total of 346 duplicate rows within the dataset. To address this, the duplicate records were systematically removed while retaining the first occurrence of

each duplicate to preserve original data consistency. As a result of this cleaning process, the dataset was reduced to a final shape of (10184, 19), ensuring that no redundant information remained.

3.4.4 Standardizing Categorical Responses

To ensure consistency and improve data quality, categorical responses that included the phrase "Don't Know or Don't Buy" were standardized to "Not sure." This transformation aimed to reduce ambiguity and facilitate more consistent analysis, as varying expressions of uncertainty could complicate downstream processing and modeling. Standardizing such responses enhanced interpretability and ensured uniformity across variables that captured consumer purchasing behavior. Additionally, any residual missing values were filled with the mode of the corresponding column to maintain data completeness and robustness.

3.4.5 Outlier Detection and Treatment

Following the removal of duplicates, potential outliers within the categorical dataset were identified by setting a frequency threshold of 20. This threshold was chosen to highlight rare categories across the variables, which included demographic information, location data, and product preferences. Given the diverse nature of the dataset, identifying infrequent responses was crucial to maintaining data consistency and minimizing noise.

To detect rare categories, the analysis iterated through each categorical column, counting the occurrences of responses. Categories that appeared less frequently than the threshold were compiled into a dictionary named `rare categories`, organized by column names. The results indicated that rare categories were predominantly concentrated in the "Location" variable, reflecting a wide range of less common or localized responses. These infrequent locations were grouped under the label "Other" to reduce dimensionality and enhance the robustness of the dataset.

This approach ensured that the dataset remained statistically consistent while minimizing the impact of sparse categories. By consolidating rare locations into a single category, the analysis maintained data integrity and improved model stability, facilitating more reliable and interpretable outcomes (aimediahouse, 2022).

3.4.6 Feature Engineering

To enhance the dataset's utility and facilitate temporal analysis, the "Period" column was decomposed into two distinct components: "Year" and "Month." This transformation allowed for more granular analysis of temporal trends while maintaining the integrity of the original data. The extraction was performed using the pandas library, leveraging the ability to convert date strings into datetime objects and isolate the year and month values. Following the extraction, the newly created "Year" and "Month" columns were positioned immediately after the original "Period" column to maintain logical coherence and improve data organization. This feature engineering step contributed to the dataset's interpretability and prepared it for subsequent analytical tasks.

3.5 Exploratory Data Analysis

Exploratory Data Analysis EDA was conducted to gain insights into consumer purchasing behavior and the underlying characteristics of the dataset. The analysis encompassed univariate, bivariate, and multivariate techniques to thoroughly investigate data patterns and relationships. Univariate analysis focused on demographic factors such as Gender, Age-group, Occupation, and Income Level to understand their individual distributions, revealing that middle-income millennials formed the largest consumer group. Bivariate analysis explored relationships between pairs of variables, such as Gender and Spending Patterns, utilizing correlation matrices and heatmaps to uncover interactions and associations. Notably, Income Level was found to correlate with purchasing behavior, as higher-income groups consistently exhibited steady spending on essential goods. To gain a deeper understanding of complex interactions, multivariate analysis was employed to examine combinations of multiple variables, identifying intricate relationships that could influence purchasing decisions. Additionally, temporal trend analysis was performed to detect shifts in purchasing patterns over time, notably highlighting changes in consumer spending behavior following the socioeconomic impacts of 2020. The insights derived from EDA provided a foundational understanding of the data and guided subsequent modeling and analysis phases.

3.6 Feature Encoding

To facilitate machine learning applications, feature encoding was implemented to transform categorical variables into numerical representations, as most machine learning models require numerical input rather than categorical data(Chong, 2020). The selected demographic variables, including *Gender*, *Age-group*, *Income Level*, and *Occupation*, were encoded to ensure compatibility with predictive algorithms. Binary encoding (GeeksforGeeks, 2022) was applied to the *Gender* variable, mapping *Male* to 1 and *Female* to 0, reflecting its binary nature. The *Age-group* variable was mapped based on generational progression, assigning *Millennials* (3), *Gen Z* (2), *Gen X* (1), and *Baby Boomers* (0), thereby preserving the ordinal relationship among generations. To maintain the hierarchical structure of the *Income Level* variable, *High Income* was mapped to 2, *Middle Income* to 1, and *Low Income* to 0. For the *Occupation* variable, integer mapping was employed by assigning each unique occupation a distinct numerical value without implying any ordinal relationship. This encoding strategy ensured that the dataset was prepared for modeling while preserving the interpretability of demographic information.

3.7 Feature Selection

Feature selection was performed to identify the most relevant demographic variables for predictive modeling, aiming to enhance model performance while minimizing redundancy and multicollinearity (Aman, 2020). The primary goal was to retain variables that significantly contribute to understanding consumer purchasing behavior without introducing noise or unnecessary complexity. To achieve this, a correlation analysis was conducted on the encoded demographic variables to examine their pairwise relationships and detect potential multicollinearity (Sahazada, 2023). The analysis revealed minimal correlations between the selected variables, indicating low multicollinearity and confirming that each demographic factor independently contributes to predicting consumer behavior.

Given the absence of significant correlations, it was determined that no further feature reduction techniques were required at this stage. Therefore, all four demographic variables (Gender, Age-group, Income Level, and Occupation) were retained for subsequent modeling due to their theoretical relevance and empirical significance in understanding consumer purchasing patterns. This approach ensured that the models captured diverse

behavioral aspects without compromising interpretability. In later stages of analysis, feature importance techniques were applied to evaluate the individual impact of each variable on model predictions, confirming the validity of the chosen features.

3.8 Model Selection and Rationale for Machine Learning Models

The selection of machine learning models was guided by the primary objective of accurately predicting demographic profiles of retail consumers in Kenya based on their purchasing behavior. Considering the complexity and multi-dimensionality of the data, a multi-output modeling(GeeksforGeeks, 2025) approach was adopted to handle the simultaneous prediction of multiple targets, specifically Gender, Age-group, Occupation, and Income Level. Instead of building separate models for each target variable, a MultiOutputClassifier was used to predict all four simultaneously, leveraging shared patterns and dependencies among these demographic traits. This approach improves computational efficiency, ensures consistent feature usage, and often leads to better generalization when outputs are interrelated such as age and income(Géron, 2019).

Compared to traditional single-output pipelines where separate models are trained independently for each target, a multi-output strategy offers improved model coherence and often higher performance, particularly when target variables are not independent. This justification is supported by (Farlessyost et al., 2021), who demonstrated that multi-output models perform more effectively than independent models in predicting correlated human traits such as trustworthiness and dominance from facial images.

A multi-model strategy was also applied to compare performance across diverse algorithms. The models included Random Forest *RF*, *XGBoost*, *ANN*, and k-Nearest Neighbors *KNN*. These algorithms were selected for their ability to capture non-linear relationships, manage high-dimensional data, and support multi-output tasks. RF and XGBoost were prioritized for their interpretability and strong performance on structured data. ANN was incorporated to model complex, non-linear patterns, while KNN served as a baseline to evaluate proximity-based classification. This combination ensured a balanced comparison across ensemble, deep learning, and instance-based methods.

After evaluation, the Improved Random Forest, optimized via GridSearchCV, emerged as the most robust performer across accuracy, precision, recall, and F1-score. It was thus

selected as the final model for deployment in the interactive Streamlit application.

3.8.1 Random Forest

Random Forest (RF) is an ensemble learning method that constructs multiple decision trees and aggregates their outputs to improve predictive accuracy. Its ability to handle high-dimensional categorical data and complex interactions makes it well-suited for multi-output predictions. The model is robust to overfitting, especially when using a large number of decision trees, and effectively quantifies feature importance, thereby identifying key demographic and behavioral factors influencing purchasing trends (Breiman, 2001).

The prediction for each output variable using RF is calculated as:

$$\hat{y}_k = \frac{1}{T} \sum_{t=1}^{T} h_t^{(k)}(x) \textit{Linusson (2013)}$$

Where:

- T is the total number of decision trees.
- $h_t^{(k)}(x)$ represents the prediction of tree t for output k.
- x is the input feature vector.

Despite its effectiveness, RF can be computationally expensive when the number of trees is high, and hyperparameter tuning is crucial to optimizing performance.

3.8.2 Extreme Gradient Boosting (XGBoost)

XGBoost is an advanced gradient boosting technique that sequentially builds trees to minimize prediction errors. It is highly efficient and scalable, making it suitable for large datasets with complex patterns. The model reduces bias by iteratively focusing on errors from previous iterations, while regularization prevents overfitting (Chen and Guestrin, 2016).

The loss function for multi-output learning is expressed as:

$$\mathcal{L}_k = \sum_{i=1}^{N} \ell\left(y_{i,k}, \hat{y}_{i,k}\right) + \sum_{t=1}^{T} \Omega(h_t^{(k)}) \textit{ChenandGuestrin (2016)}$$

Where:

- $\ell(y_{i,k}, \hat{y}_{i,k})$ represents the loss function (e.g., Mean Squared Error).
- $\Omega(h_t^{(k)})$ is a regularization term that controls model complexity.
- N denotes the total number of observations.

XGBoost's ability to handle missing values and perform parallel computation enhances efficiency. However, hyperparameter tuning is critical to mitigate overfitting, especially when dealing with multi-output tasks.

3.8.3 k-Nearest Neighbors

KNN is a non-parametric algorithm that classifies data points based on the majority label among the nearest neighbors. It is particularly useful for detecting local patterns and clustering similar consumer groups. KNN's simplicity and interpretability make it suitable as a baseline model, though it is computationally intensive for large datasets (Altman, 1992).

The prediction for each output is derived as:

$$\hat{y}_k = \frac{1}{K} \sum_{i \in \mathcal{N}(x)} y_{i,k} (\textit{Altman}, \ 1992)$$

Where:

- $\mathcal{N}(x)$ represents the set of K-nearest neighbors.
- $y_{i,k}$ is the k-th output label of the i-th neighbor.((Zhang, 2004))

Careful selection of the number of neighbors (K) and distance metric (e.g., Euclidean distance) is essential for optimal performance.

3.8.4 Artificial Neural Networks

Artificial Neural Networks (ANN) are deep learning models designed to capture complex, non-linear relationships among variables. ANNs consist of interconnected nodes (neurons) organized in layers, each performing a weighted sum of inputs followed by a non-linear

activation function. This architecture makes ANNs highly flexible and capable of learning intricate patterns from data (Brownlee, 2020).

The output of a neuron in layer (l) is calculated as:

$$z^{(l)} = W^{(l)}a^{(l-1)} + b^{(l)}, \quad a^{(l)} = f\left(z^{(l)}\right)$$

Where:

- $z^{(l)}$ represents the linear transformation at layer l.
- $W^{(l)}$ and $b^{(l)}$ are the weights and biases for layer l.
- $a^{(l)}$ is the activation function output for layer l.
- f is a non-linear activation function (e.g., ReLU, Sigmoid).

ANNs are particularly effective for capturing latent relationships but are sensitive to overfitting, especially when the dataset size is moderate. Regularization techniques and dropout were employed to enhance model generalization.

3.8.5 Justification of Model Choice

RF and XGBoost were chosen for their ability to handle multi-output predictions and to capture complex interactions among demographic and behavioral variables. Their ensemble nature enhances robustness and minimizes overfitting. KNN served as a baseline model to evaluate the performance of more advanced algorithms, while ANN was utilized to explore non-linear patterns and deep representations within the data. This combination ensured comprehensive coverage of both traditional machine learning and deep learning approaches, enabling the study to capture diverse consumer purchasing behaviors effectively.

3.9 Data Splitting

To ensure robust model performance and mitigate the risk of data leakage, the dataset was partitioned into training and testing sets using the **train test split** function from the *scikit learn* library. A stratified split was employed to preserve the proportional distribution of each target class between the training and testing sets, thereby maintaining class

balance. The data was divided in an 80-20 ratio, with 80% allocated for model training and 20% reserved for testing its generalization capabilities. To enhance reproducibility, the random state was fixed at 42, ensuring consistent outcomes across different training runs. This splitting strategy was essential to maintain model reliability and minimize the risk of overfitting.

3.10 Data Balancing

Initial analysis revealed class imbalance across several target variables, particularly in Occupation and certain Income Level categories. To address this, the Synthetic Minority Over-sampling Technique (Synthetic Minority Over-sampling Technique (SMOTE)) was explored as a potential solution to improve the model's ability to learn from under-represented classes. SMOTE is designed to generate synthetic samples between existing minority class instances, enhancing class balance without duplicating data.

However, due to the use of a multi-output classification framework, direct application of SMOTE was not feasible. The technique is primarily suited for single-output classification and does not natively support multi-target data structures. As a result, while the approach was considered and tested during experimentation, it was ultimately not integrated into the final pipeline. The model instead relied on robust ensemble methods and hyperparameter tuning to mitigate imbalance effects and improve generalization across all outputs.

3.11 Model Training

The model training phase involved two key stages. In the first stage, initial baseline models were developed using Random Forest, Extreme Gradient Boosting, and k-Nearest Neighbors to establish benchmark performance. These models were trained directly on the full set of encoded features and labels without prior hyperparameter tuning. Their initial evaluation on the holdout test set provided insights into feature effectiveness, class imbalance issues, and label-specific prediction challenges. The observed performance disparities across the output variables guided subsequent tuning priorities, especially for underperforming labels like Occupation and Income Level.

Hyperparameter Optimization

In the second stage, performance improvements were pursued through targeted hyperparameter tuning and architectural search. Each model underwent methodical adjustments to maximize predictive power while addressing overfitting and label imbalance:

Random Forest (RF): GridSearchCV was used in conjunction with 3-fold cross-validation to tune parameters systematically. The search covered a grid consisting of `n_estimators` [100, 150], `max_depth` [10, 20], and `min_samples_split` [5, 10]. Accuracy was used as the scoring metric. The best combination was selected and retrained within a `MultiOutput Classifier` to support concurrent prediction of the four target demographics. This approach ensured consistency in feature learning across tasks and allowed the model to exploit shared structure in consumer behavior patterns. Final artifacts including the model, encoders, and inverse maps were saved with `joblib`.

XGBoost: A manually configured `XGBClassifier` was tested with selected parameters: `n_estimators`=18, `learning_rate`=0.0001, and `max_depth`=15. While hyperparameter tuning was not automated via `RandomizedSearchCV`, the values were iteratively refined through multiple trial-and-error runs, guided by validation accuracy and F1-scores. The model's strong capacity to handle non-linearities was preserved in the multi-output structure using `MultiOutputClassifier`.

k-Nearest Neighbors (KNN): A series of models were tested by varying the number of neighbors. The value of $k = 3$ was retained based on cross-validation within the training data and validated through performance metrics on the test set. Although simple, KNN provided a useful baseline for assessing the added complexity of other models. It was also executed within a `MultiOutputClassifier` to produce all demographic predictions concurrently.

Artificial Neural Networks (ANN): Model architecture and training dynamics were tuned using `Keras Tuner` and `RandomSearch`. Key hyperparameters included the number of hidden layers (ranging from 1 to 4), number of units per layer (32 to 256), dropout rates (0.0 to 0.5), optimizer choice (`adam`, `rmsprop`, `sgd`), and learning rate (1e-2 to 1e-4). Early stopping was applied to monitor validation loss, and a learning rate decay schedule was introduced to improve convergence stability. The best-performing model from the

search was then trained on the full dataset using callbacks for generalization control.

Cross-Validation and Evaluation

Random Forest: Employed 3-fold cross-validation embedded in `GridSearchCV` to evaluate model stability and select the optimal parameter set.

ANN: Utilized a 20% validation split and early stopping callbacks to prevent overfitting and track convergence.

XGBoost and KNN: Evaluated on the test set post-manual tuning, with metrics tracked across target outputs. **Evaluation Metrics:** Standard metrics including accuracy, precision, recall, F1-score (macro and micro), AUC-ROC, and optimal classification thresholds were used for comparison.

Multi-output Metrics: Hamming loss, Hamming score, and subset accuracy were computed to assess model performance on multi-output classification holistically.

The model selection concluded with Random Forest as the final deployed model due to its balanced performance across all outputs, high interpretability, and robustness on unseen data. The complete training pipeline, including data encoders, feature mappings, trained model weights, and inverse label dictionaries, was exported using `joblib` for streamlined deployment in a Streamlit web application.

This comprehensive two-stage training and evaluation process ensures replicability, model integrity, and real-world readiness of the final predictive engine.

3.12 Model Evaluation

The evaluation phase aimed to measure the performance, stability, and reliability of all trained models using both standard and multi-output classification metrics.(Javatpoint, 2023) This was essential to assess the models' effectiveness in predicting multiple consumer demographic traits (Gender, Age-group, Occupation, and Income Level) simultaneously. Evaluation was conducted on the 20% holdout test set, while cross-validation during training ensured consistent performance across data splits.

Standard Evaluation Metrics

The following classical classification metrics were used to assess each output variable individually:

Accuracy Accuracy measures the overall proportion of correctly predicted instances:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- TP: True Positives
- TN: True Negatives
- FP: False Positives
- FN: False Negatives

While intuitive, accuracy alone is insufficient in imbalanced data scenarios, which existed in this study, particularly for Occupation.(True Negative (TN))

Precision Precision evaluates the proportion of correctly predicted positives among all predicted positives:

$$Precision = \frac{TP}{TP + FP}$$

Where:

- TP: True Positives
- FP: False Positives

This metric is vital when the cost of false positives is high.

Recall (Sensitivity) Recall measures the proportion of actual positives correctly identified:

$$Recall = \frac{TP}{TP + FN}$$

Where:

- TP: True Positives
- FN: False Negatives

It is especially important when false negatives are costly.

F1-Score The F1-Score provides a harmonic mean of precision and recall, balancing both metrics:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

A high Harmonic Mean of Precision and Recall (F1-Score) indicates strong balance between minimizing false positives and false negatives.

Area Under the Curve - Receiver Operating Characteristic (AUC-ROC) The Area Under the Receiver Operating Characteristic Curve measures the model's ability to discriminate between positive and negative classes:

$$FPR = \frac{FP}{FP + TN}, \quad TPR = \frac{TP}{TP + FN}$$

Where:

- FPR: False Positive Rate (False Positive Rate (FPR))
- TPR: True Positive Rate (Recall (True Positive Rate (TPR)))

A higher AUC indicates better separability across class labels. Area Under the Curve (AUC) was computed for each demographic trait, where applicable.

Multi-output Evaluation Metrics

Given the multi-label nature of the prediction task, holistic metrics were also applied:

Hamming Score and Loss Hamming Score measures the fraction of correctly predicted labels across all outputs:

$$Hamming\ Score = 1 - \frac{1}{N \times L} \sum_{i=1}^{N} \sum_{j=1}^{L} \mathbf{1}_{\hat{y}_{ij} \neq y_{ij}}$$

Where:

- N: Number of samples
- L: Number of labels
- $\hat{y}_{ij}$: Predicted label for sample i, output j
- y_{ij}: True label for sample i, output j

Subset Accuracy Subset accuracy is a stricter measure which considers a prediction correct only if all labels for a sample match exactly:

$$Subset\ Accuracy = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\hat{\mathbf{y}}_i = \mathbf{y}_i}$$

Micro and Macro Averages To summarize performance across all labels:

- **Micro-average**: Aggregates True Positive (TP), False Positive (FP), False Negative (FN) across all outputs and computes a global score.
- **Macro-average**: Computes metrics independently for each label and averages them equally.

These distinctions were critical to fairly evaluate performance across imbalanced targets.

Cross-Validation Strategy

For Random Forest, a 3-fold GridSearchCV was applied during hyperparameter tuning. For ANN, a 20% validation split with early stopping and learning rate decay helped mitigate overfitting. All models were further evaluated on a 20% test set using saved artifacts.

Optimal Threshold Selection

For ANN, optimal thresholds were selected per output based on maximizing F1-score using predicted probabilities. Thresholds were evaluated over a grid (0.05 to 0.95), and best-performing values were retained to improve binary decision performance per label.

Conclusion of Evaluation

The Random Forest model demonstrated the highest stability across outputs, with well-balanced precision and recall, high F1-scores on majority classes, and reasonable robustness for underrepresented labels. The comprehensive evaluation ensured that the selected model not only performed well on the training data but also generalized effectively to unseen retail consumer behavior.

3.13 Feature Importance Analysis

Feature importance analysis was conducted to determine which input variables most influenced the model's predictions across different demographic outputs. For the tree-based models, particularly the final Improved Random Forest, importance scores were computed based on the average reduction in Gini impurity across all decision trees in the ensemble. These global importance values were further supported by SHAP (SHapley Additive exPlanations) analysis, which provided model-agnostic insights into feature contributions.

The results revealed that variables such as *Location*, *Period*, and consumer spending categories like *Hair Care*, *Alcohol Beverages*, and *Airtime/Data Bundles* were consistently among the most important predictors. These features played varying roles depending on the target variable, contributing differently to the prediction of *Gender*, *Age-group*, *Occupation*, and *Income Level*. Understanding these feature dynamics helps explain how the model aligns with known consumer behavior patterns in Kenya and strengthens its interpretability for decision-making in real-world retail settings.

3.14 Model Deployment

The final predictive model was deployed using Streamlit to enable real-time interaction with users through a web interface. The deployment process focused on simplicity, accessibility, and practical use in the Kenyan retail context.

Model development and training were conducted using Google Colab and VSCode. After finalizing the best-performing model, the serialized model files and preprocessing artifacts were saved using `joblib`. The complete application, including the model and supporting scripts, was pushed to a public GitHub repository using Git.

The Streamlit app was designed to allow users to input consumer behavior data and receive immediate demographic predictions. To deploy the app, Streamlit Cloud was used by linking the GitHub repository directly to the hosting platform. For large model files, Google Drive was integrated using the `gdown` library to support runtime downloads during deployment.

This approach ensured that the deployed system remained lightweight, easy to maintain, and accessible to users without the need for complex infrastructure or technical expertise. Further deployment details are discussed in subsequent chapters.

Chapter 4: System Design and Architecture

4.1 Introduction

This chapter presents the system design and architecture of the predictive model built to analyze consumer purchasing behavior in Kenya. The aim was to build a user-friendly, interpretable, and accessible platform that could generate real-time predictions based on spending patterns. The entire system is implemented using lightweight open-source tools and deployed via Streamlit Cloud.

4.2 System Design Overview

The system follows a modular architecture guided by the CRISP-DM framework. It integrates data ingestion, feature engineering, model training, and interactive visualization in a streamlined and reproducible pipeline.

Data Preparation: The dataset was sourced from GitHub and processed using Google Colab and VSCode.

Model Training: A multi-output Random Forest model was trained using Scikit-learn, and the best configuration was selected via GridSearchCV.

Model Serialization: Artifacts such as the model, encoders, and feature mappings were saved using `joblib` and tracked using Git LFS.

Deployment: A Streamlit app was created for real-time predictions and deployed to Streamlit Cloud. Large files were hosted on Google Drive and dynamically retrieved using `gdown`.

4.3 Pipeline Architecture

The system pipeline includes the following components:

1. **Data Ingestion:** Data is loaded from a public GitHub repository in Comma-Separated Values (CSV) format.
2. **Preprocessing:** Includes label encoding of categorical variables, handling of missing values, and column ordering.

3. **Model Training:** A multi-output Random Forest is trained using GridSearchCV to optimize hyperparameters across four target variables.

4. **Serialization:** All artifacts are serialized using `joblib` and stored for reuse during deployment.

5. **Deployment:** A Streamlit interface is developed for interaction, with model files loaded either locally or from Google Drive.

4.4 System Architecture

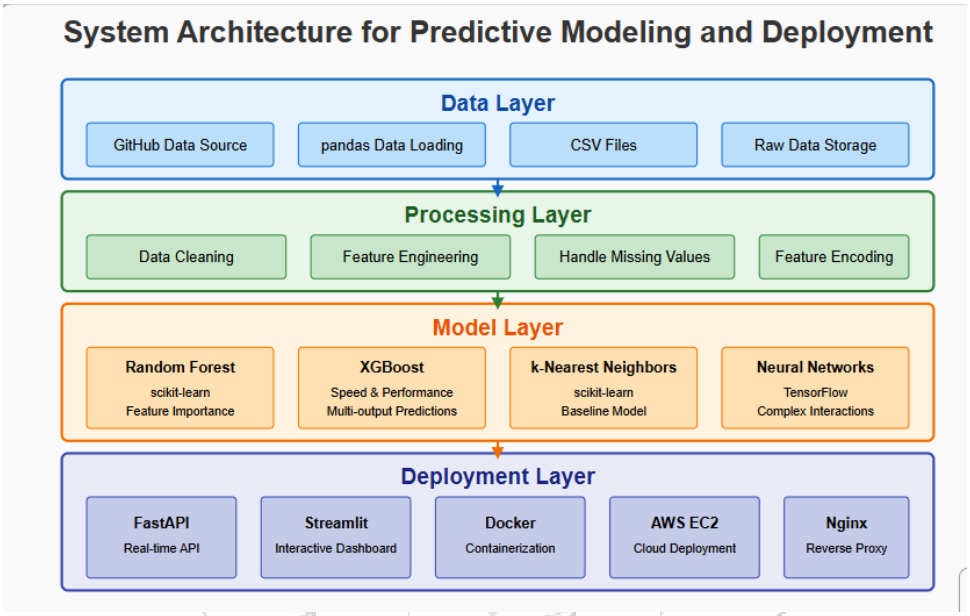


Figure 4.1: Streamlit-based System Architecture for Real-Time Prediction

The architecture is divided into three layers:

Preprocessing Layer: Encodes inputs and ensures column consistency.

Model Layer: Loads trained artifacts and makes predictions on encoded inputs.

Interface Layer: Provides a tab-based User Interface (UI)(Predictions, EDA, Explainability) using Python library for creating web apps for data science and machine learning (Streamlit).

4.5 Technologies Used

Google Colab, VSCode: Used for data analysis and model training.

Scikit-learn: For Random Forest classification and evaluation.

Streamlit: For building and deploying the interactive frontend.

Git + Web-based platform for version control using Git (GitHub): For version control and hosting of code and datasets.

Google Drive + gdown: For loading large model files during runtime.

joblib: Used for serialization of models and encoders.

The system was designed for simplicity, usability, and reproducibility. By using a purely Streamlit-based interface and cloud deployment, it offers a lightweight, accessible tool for predicting demographic segments based on retail spending patterns in Kenya.

Chapter 5: System Implementation and Testing

5.1 Introduction

This chapter describes the implementation and testing of the final predictive system. The model was developed using Random Forests with multi-output classification and deployed via a custom Streamlit app. All training, testing, and visualization steps were done in Python using Google Colab and Visual Studio Code (VSCode). The application is live on Streamlit Cloud and allows user interaction with no API backend.

5.2 Implementation Process

1. Data Preparation The dataset was loaded directly from GitHub. It was cleaned, label-encoded, and split into training and test sets. All transformations were saved using `joblib` for deployment reuse.

2. Model Training Several models were tested, including KNN and XGBoost, but the Random Forest model was selected for its balanced accuracy and interpretability. Hyperparameter tuning was done via Grid Search with Cross-Validation (GridSearchCV) with multi-output handling.

3. Model Serialization The trained model, encoders, column names, and inverse maps were serialized with `Python library for lightweight pipelining and object persistence (joblib)`. Large files like `feature-selector.joblib` were tracked using Git Large File Storage (LFS) and uploaded to Google Drive for runtime loading.

4. Streamlit App Development The application was built in Streamlit with three tabs: Predictions, Data Overview EDA, and Model Explainability. Input fields used dropdowns to represent spending habits, and outputs displayed demographic predictions and confidence scores using `st.metric`. Visualizations included bar charts and SHapley Additive exPlanations (SHAP) plots.

5. Deployment to Streamlit Cloud The app was pushed to GitHub and connected to Streamlit Cloud. Due to file size limits, the large model file was downloaded dynamically using `gdown`. Configuration secrets (like Google Drive IDs) were added through the

Streamlit Cloud UI.

5.3 System Architecture

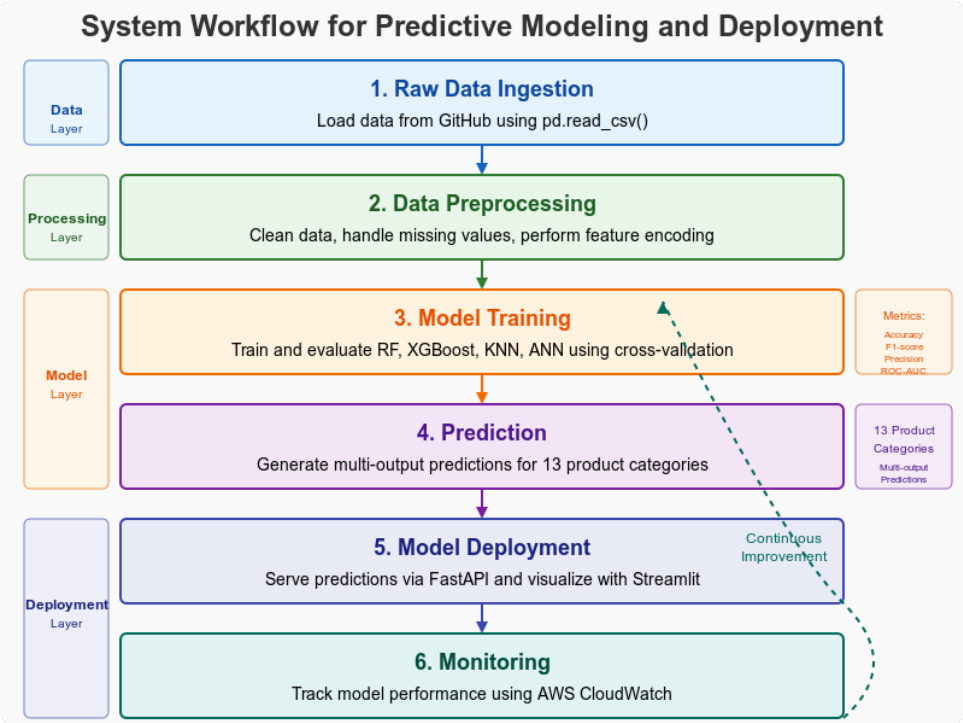


Figure 5.1: System Workflow for Real-Time Predictions using Streamlit

5.4 Interface Features

Input sidebar for selecting spending behavior across 14 product categories.

Prediction outputs with confidence scores for four demographic categories.

Interactive bar charts showing feature importance per target.

SHAP summary and force plots for explaining model behavior.

Responsive dark-themed UI with custom `config.toml` and Cascading Style Sheets (CSS) styling.

5.5 Testing and Validation

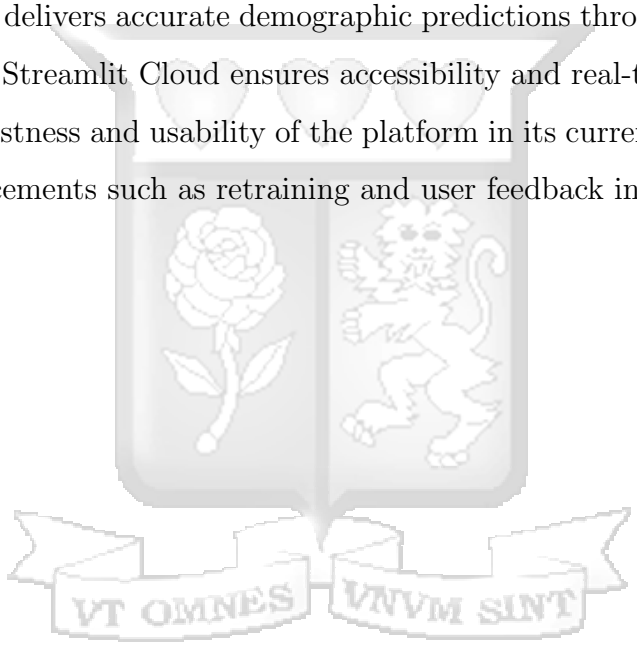
Model Validation Cross-validation and hold-out testing were used to assess the model. Accuracy and F1-scores were computed across targets like Gender, Age-group, and Income Level.

Functionality Testing Predictions were tested through the frontend with edge-case inputs. Outputs were consistent and interpretable, confirming correct model integration.

Deployment Testing All files were loaded successfully in the Streamlit environment, including large artifacts from Google Drive. Runtime logs and caching confirmed stable performance.

User Experience Testing The dark-themed UI, prediction metrics, and visualization tabs were evaluated for clarity and usability. Users were able to understand results and use the tool without technical guidance.

The deployed system delivers accurate demographic predictions through an intuitive web interface. The use of Streamlit Cloud ensures accessibility and real-time inference. Testing confirms the robustness and usability of the platform in its current state, with potential for future enhancements such as retraining and user feedback integration.



Chapter 6: Discussion of Results

6.1 Results

This section presents the findings of the study, moving from exploratory data analysis to model development and evaluation. It systematically discusses the results of each phase of the analysis, including model performance comparisons, confusion matrix evaluations, and feature importance interpretation using SHAP (SHapley Additive exPlanations) values. The chapter also integrates critical discussions aligned with the research objectives to contextualize the results and demonstrate how they address the identified research gaps.

6.2 Exploratory Data Analysis (EDA)

The exploratory data analysis phase was fundamental to understanding the characteristics and distributions within the dataset. This phase involved univariate, bivariate, and multivariate analyses to gain comprehensive insights into consumer behavior.

6.2.1 Univariate Analysis

Univariate analysis focused on individual variables, revealing their distribution and frequency. It provided a foundation for understanding consumer preferences and demographic characteristics independently before examining relationships between variables.

Distribution of Gender, Age Group, Occupation, and Income Level The univariate analysis examined individual variables to understand their distribution and central tendency. Key demographic factors such as Gender, Age-group, Occupation,and Income Level were analyzed to understand their composition within the dataset. The univariate analysis revealed that middle-income millennials constituted the largest consumer group. Additionally, the proportion of male and female respondents was almost balanced, indicating a diverse representation within the dataset. Figure 6.1

6.2.2 Bivariate Analysis

Bivariate analysis aimed to explore relationships between pairs of variables, examining how demographic factors influence purchasing decisions. This analysis helped contextu-

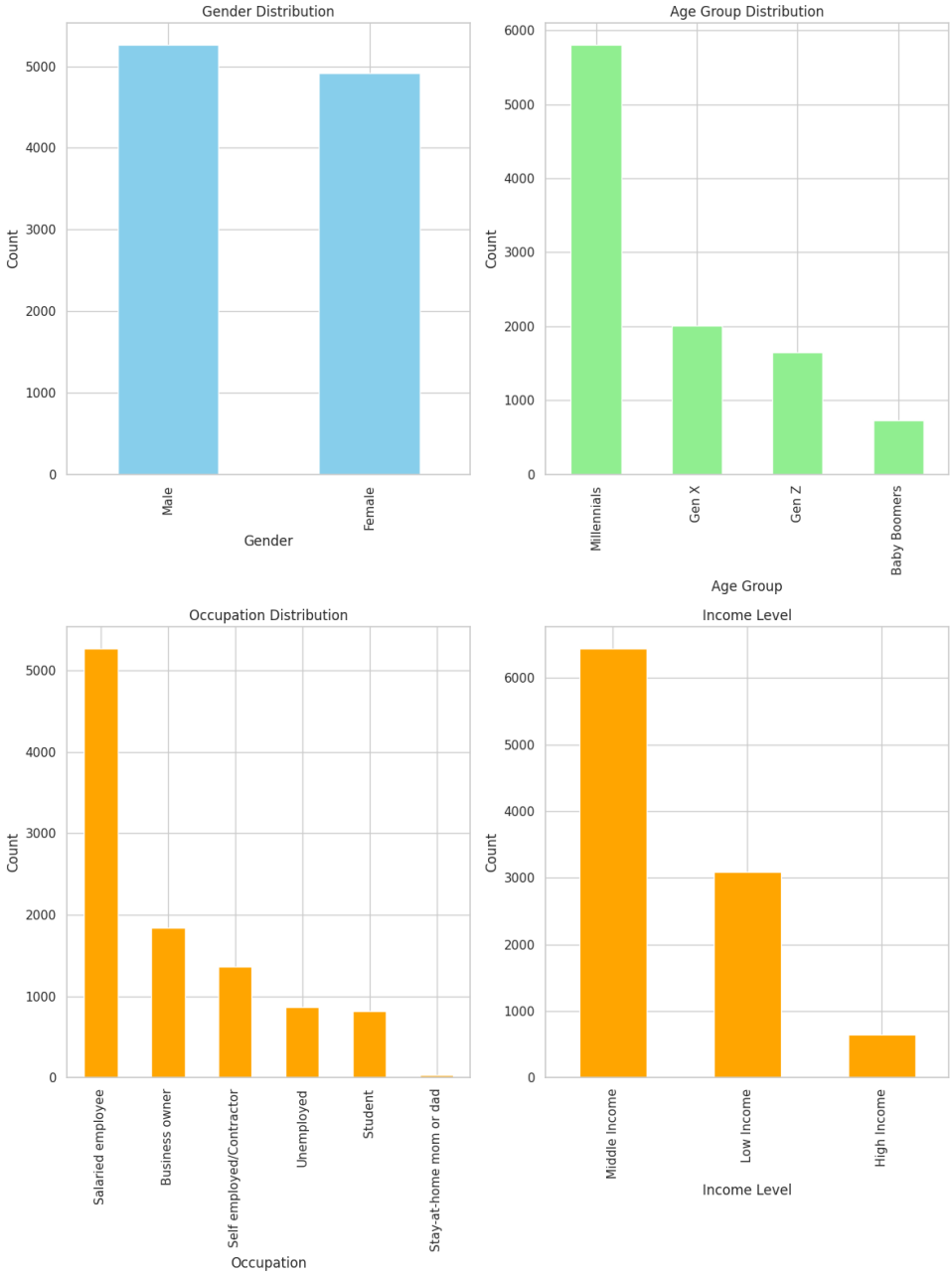


Figure 6.1: Distribution of Gender, Age Group, Occupation, and Income Level

alize consumer behaviors within specific demographic frameworks.

Gender-Based Insights Gender differences were evident in alcohol consumption patterns, with female consumers displaying a slightly higher tendency to reduce consumption compared to males. This could indicate a greater inclination towards health-conscious behaviors among women, reflecting sociocultural and personal health considerations. Figure 6.2

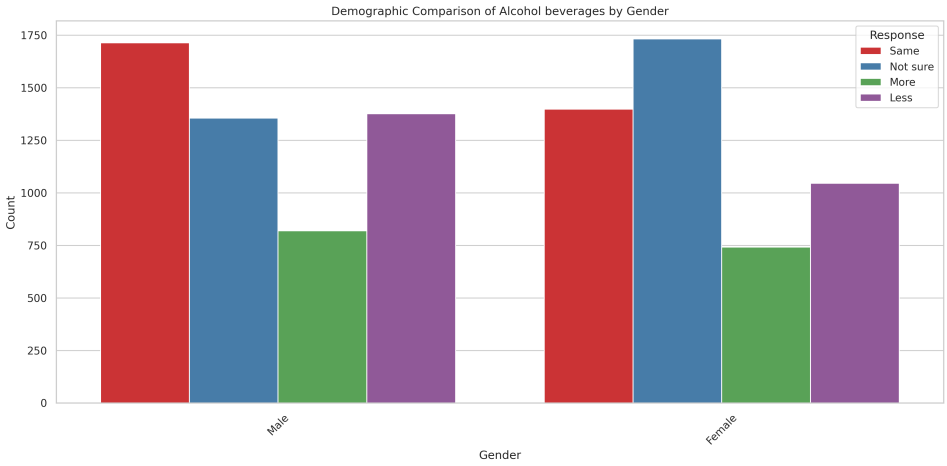


Figure 6.2: Demographic Comparison of Airtime and Data Bundles by Income Level

Age Group Insights Millennials demonstrated the most dynamic purchasing behavior, particularly in airtime/data bundle consumption and alcohol usage. Their high digital engagement and social connectivity are key drivers of this trend, contrasting with more conservative spending behaviors among Baby Boomers, who maintain consistent usage patterns across product categories. Figure 6.3

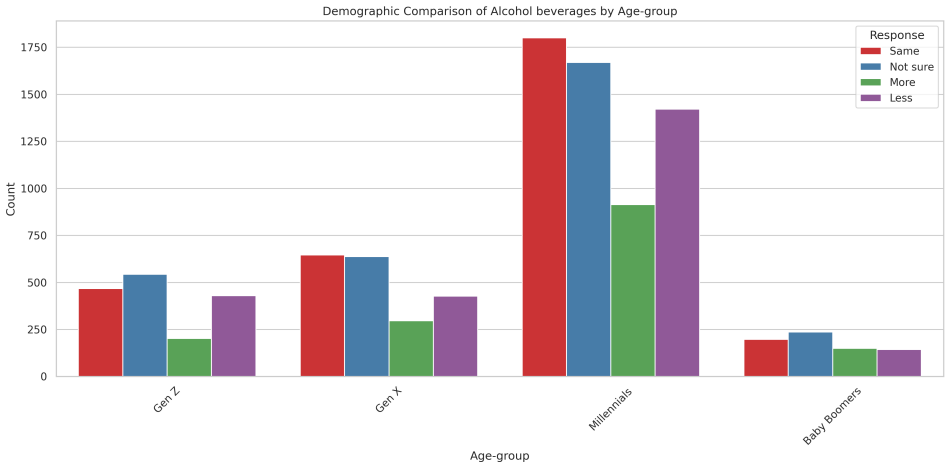


Figure 6.3: Demographic Comparison of Alcohol Beverages by Age-group

Income Level Insights Income level significantly impacted product consumption, particularly in airtime and data bundles. Middle-income consumers exhibited increased consumption, likely reflecting a balance between affordability and connectivity needs, while low-income groups displayed stable or declining usage patterns. Alcohol consumption, however, showed a nuanced trend, with middle-income groups maintaining stable consumption, while low-income groups showed increased usage, possibly reflecting stress-

induced consumption patterns. Figure 6.4

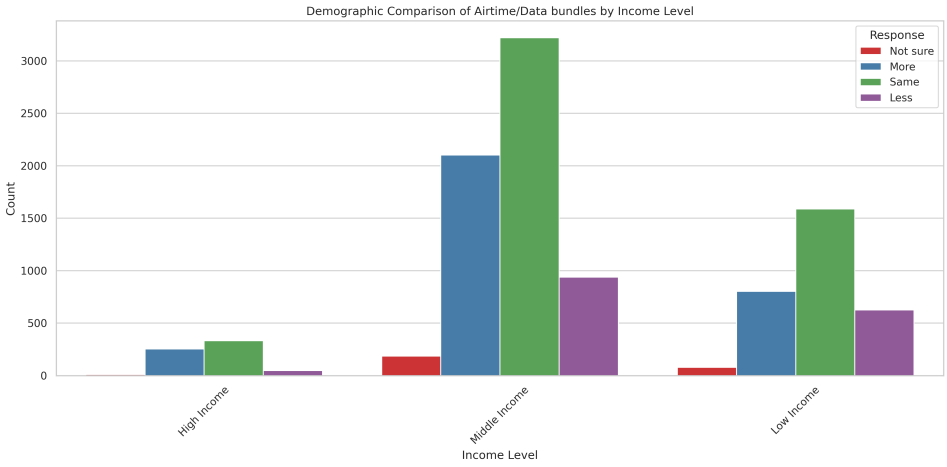


Figure 6.4: Demographic Comparison of Airtime and Data Bundles by Income Level

Occupation-Based Insights Occupation played a critical role in purchasing patterns, particularly for electronics and personal hygiene products. Salaried employees maintained stable purchasing behavior due to predictable incomes, while students exhibited lower consumption levels, possibly due to budget constraints. Business owners and self-employed individuals showed variability, reflecting diverse financial stability and business demands. Figure 6.5

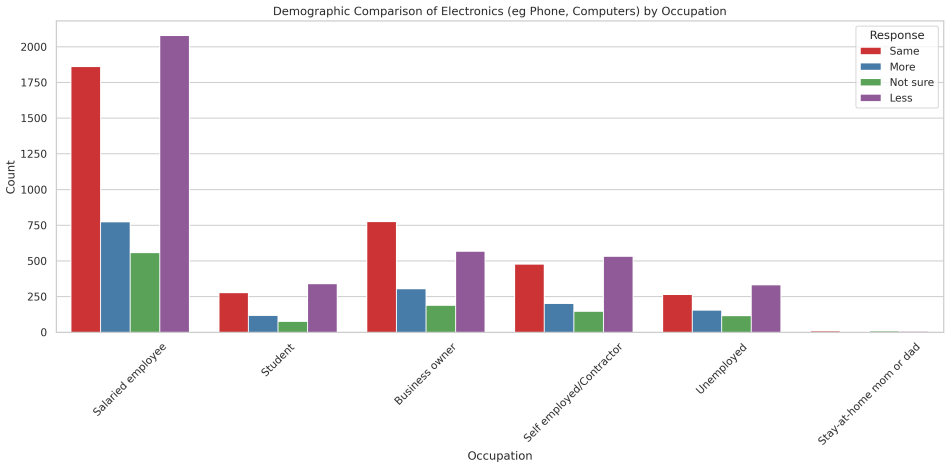


Figure 6.5: Demographic Comparison of Electronics (e.g., Phone, Computers) by Occupation

6.2.3 Multivariate Analysis

The multivariate analysis focused on examining the complex interactions between multiple variables. This approach was essential for understanding how demographic and economic factors jointly influence purchasing decisions.

Correlation Analysis Correlation analysis revealed that no single demographic factor overwhelmingly dominated consumer behavior patterns. The moderate negative correlation between age group and income level (-0.32) indicated that as consumers aged, their income levels tended to stabilize, reflecting established careers and financial stability. Figure 6.6

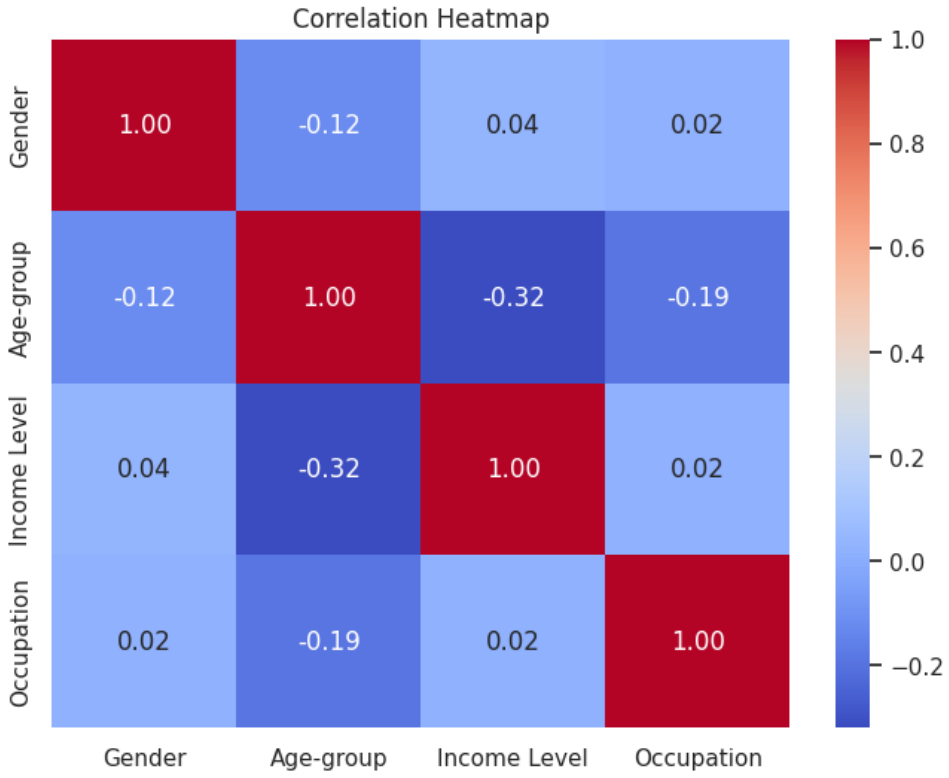


Figure 6.6: Correlation Analysis of Demographic and Economic Variables

6.2.4 Temproral Analysis

Temporal analysis examined the yearly shifts in purchasing behavior, especially considering the socio-economic impacts of the COVID-19 pandemic.

Yearly Distribution of Personal Hygiene Products The distribution of personal hygiene product usage reveals that most consumers consistently reported the same level

of usage over the years, with a peak observed in 2020. The proportion of respondents reporting increased usage has declined since then, while those reporting reduced usage remain minimal. This consistent pattern indicates a sustained demand for personal hygiene products. Figure 6.7

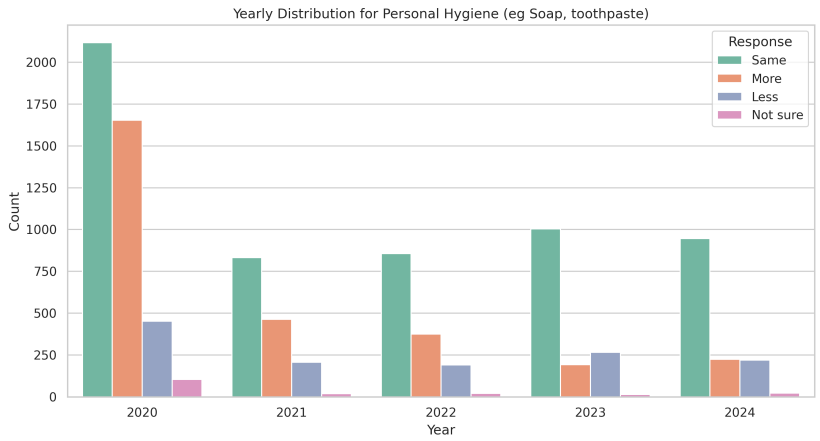


Figure 6.7: Yearly Distribution of Personal Hygiene Products (e.g., Soap, Toothpaste)

Heatmap of Yearly Distribution for Airtime and Data Bundles The heatmap highlights the dynamic nature of purchasing trends over time, revealing an increase in digital product consumption during and after the pandemic. It provides a visual representation of how user perceptions and behaviors related to airtime and data bundle usage have evolved over the years. The trends indicate a significant shift in 2020, followed by a period of adjustment and eventual stabilization. Figure 6.8

6.3 Model Performance and Comparison

To accurately predict consumer purchasing behavior, four machine learning models were employed and compared: Random Forest, K-Nearest Neighbors, XGBoost, and Artificial Neural Networks. The comparison was based on evaluation metrics such as accuracy, precision, recall, and F1-score, providing a multi-dimensional view of model performance. These models were evaluated both in their baseline form and after targeted tuning strategies, as described in Section 3.11.

6.3.1 Model Performance Comparison

Model evaluation results before and after optimization are summarized below:

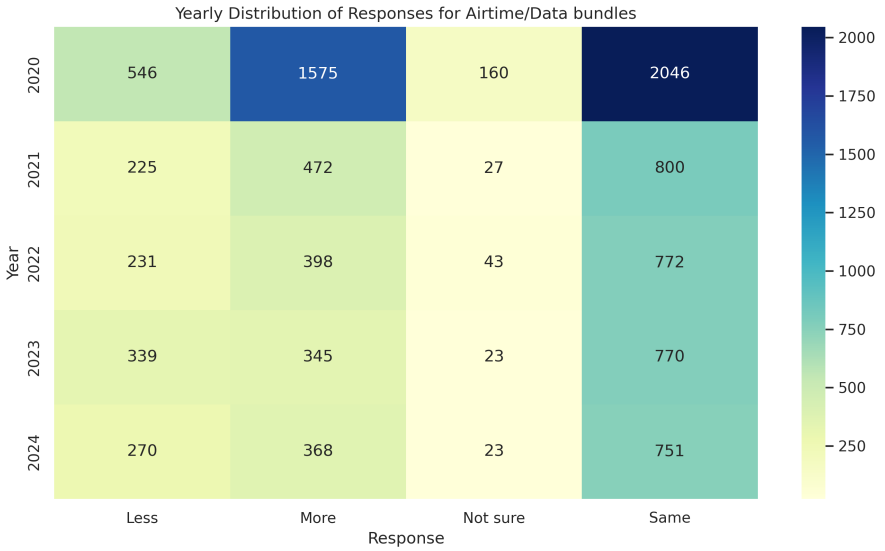


Figure 6.8: Heatmap of Yearly Distribution for Airtime and Data Bundles

Random Forest (Baseline) The baseline Random Forest model delivered the highest initial accuracy for Age-group (76%) and Income Level (69%), though it exhibited moderate precision and recall values across outputs. This indicated partial sensitivity to demographic signals but emphasized the need for tuning, especially for underrepresented classes such as Occupation. As shown in Table 6.1, the model's performance varies across categories.

Table 6.1: RF Classification Report Before Model Improvement

Category	Accuracy	Precision	Recall	F1-Score
Gender	0.59	0.59	0.59	0.59
Age-group	0.76	0.58	0.38	0.38
Occupation	0.50	0.26	0.20	0.20
Income Level	0.69	0.52	0.45	0.45

K-Nearest Neighbors The KNN model yielded consistent yet suboptimal results, with Age-group (70%) and Income Level (66%) performing better than Gender (51%) and Occupation (47%). The model's fixed distance-based mechanism struggled to generalize under the complex feature interactions present in consumer purchasing data. As shown in Table 6.2, the KNN model performs best in predicting income level.

XGBoost marginally outperformed KNN in Age-group (75%) and Income Level (71%), demonstrating its capacity to model non-linear relationships. However, its performance

Table 6.2: KNN Classification Report

Category	Accuracy	Precision	Recall	F1-Score
Gender	0.51	0.51	0.51	0.51
Age-group	0.70	0.44	0.42	0.42
Occupation	0.47	0.25	0.21	0.21
Income Level	0.66	0.51	0.46	0.47

remained limited on Gender (51%) and Occupation (46%), possibly due to under-tuned parameters and insufficient data variety. As detailed in Table 6.3, the XGBoost model yields strong performance on age-group and income level.

Table 6.3: XGBoost Classification Report

Category	Accuracy	Precision	Recall	F1-Score
Gender	0.51	0.25	0.50	0.34
Age-group	0.75	0.55	0.44	0.46
Occupation	0.46	0.24	0.23	0.23
Income Level	0.71	0.54	0.50	0.51

Artificial Neural Networks The ANN model was fine-tuned using Keras Tuner. It was evaluated specifically on Age-group and Occupation, showing divergent behaviors: high precision but low recall for Age-group (92%, 20%), and high recall but low precision for Occupation (77%, 13%). These findings suggest overfitting and class imbalance challenges. Table 6.4 highlights the post-improvement performance of the ANN model across age group and occupation prediction.

Table 6.4: ANN Classification Report (After Improvement)

Category	Accuracy	Precision	Recall	F1-Score
Age-group	0.25	0.92	0.20	0.33
Occupation	0.27	0.13	0.77	0.22
Macro Avg	-	0.52	0.48	0.27

Improved Random Forest (After Hyperparameter Tuning) With GridSearchCV optimization, the final Random Forest model displayed the best balance across metrics and targets. Age-group and Income Level achieved the highest accuracies (78% and 73% respectively), supported by well-rounded F1-scores. The model's stability and improved

recall highlighted the success of tuning efforts. As shown in Table 6.5, the RF model showed improvement particularly in predicting income level.

Table 6.5: RF Classification Report After Model Improvement

Category	Accuracy	Precision	Recall	F1-Score
Gender	0.43	0.43	0.43	0.42
Age-group	0.78	0.70	0.41	0.43
Occupation	0.53	0.35	0.21	0.19
Income Level	0.73	0.59	0.54	0.55

6.3.2 Justification for Choosing Improved Random Forest

The Improved Random Forest model was selected for deployment based on a combination of quantitative performance metrics, interpretability, and operational robustness within a multi-output learning framework. Among all evaluated models—K-Nearest Neighbors, XGBoost, Artificial Neural Networks, and RF—the improved RF exhibited the most consistent performance across all four demographic prediction tasks: Gender, Age-group, Occupation, and Income Level.

From a metric-based perspective, the RF model achieved the highest test set accuracy for Age-group (78%) and Income Level (73%), outperforming both KNN (70%, 66%) and XGBoost (75%, 71%). While its performance on Occupation (53%) and Gender (43%) remained moderate, it surpassed ANN in terms of overall F1-scores and showed less volatility across labels. Unlike the ANN model, which demonstrated highly imbalanced precision-recall trade-offs (e.g., 92% precision and 20% recall for Age-group), the RF model maintained balanced scores, with Age-group reaching an F1-score of 0.43 and Income Level at 0.55.

In addition, the RF model achieved a lower Hamming loss and higher subset accuracy relative to the other models in the multi-output setting. These multi-label metrics indicate that RF preserved inter-label dependencies better than KNN or XGBoost, which struggled to generalize across all demographic outputs. Moreover, its macro-averaged F1-score showed less skew, suggesting resilience to class imbalance—an issue that significantly impacted the Occupation and Gender targets.

Another justification lies in interpretability and deployment feasibility. RF's tree-based structure allowed for clear extraction of feature importance, further enhanced by SHAP

analysis, which provided model-agnostic explanations for decision-making. This is particularly valuable for non-technical retail stakeholders. In contrast, models such as ANN lacked transparency, limiting their practical utility despite competitive performance on isolated targets.

While the relatively strong performance of RF aligns with established literature on ensemble methods in classification tasks, this study extends those findings by confirming its effectiveness in a multi-output context using Kenyan retail data. The ability to model shared structure across interrelated outputs (e.g., Age-group and Income Level) made RF particularly well-suited for joint prediction, reducing the need to train and maintain separate models for each demographic dimension.

Future work could strengthen this justification through comparative benchmarking against single-output pipelines to quantify efficiency and generalization trade-offs. Nonetheless, the evidence presented, rooted in evaluation metrics, interpretability, and deployment readiness, supports the selection of the Improved Random Forest as the optimal model for the predictive engine.

6.3.3 Feature Importance and Interpretability

Understanding which features most influence model predictions is critical for drawing actionable insights and enhancing trust in machine learning systems. In this study, feature importance scores were extracted from the final deployed model, an Improved Random Forest trained on all demographic outputs, using impurity-based Gini importance and SHAP values. These global interpretability scores reveal the average contribution of each feature to prediction outcomes across all samples.

The feature importances were computed individually for each target demographic variable: Gender, Age-group, Occupation, and Income Level. These rankings allowed for granular insight into how consumer traits and product engagement drive demographic predictions.

Gender As shown in Figure 6.9, location and period emerged as the strongest predictors of gender, followed by interest in personal care products such as hair care and skin care. This indicates a spatial and seasonal component in gender-related purchasing behavior,

with preferences for grooming and entertainment items also contributing significantly.

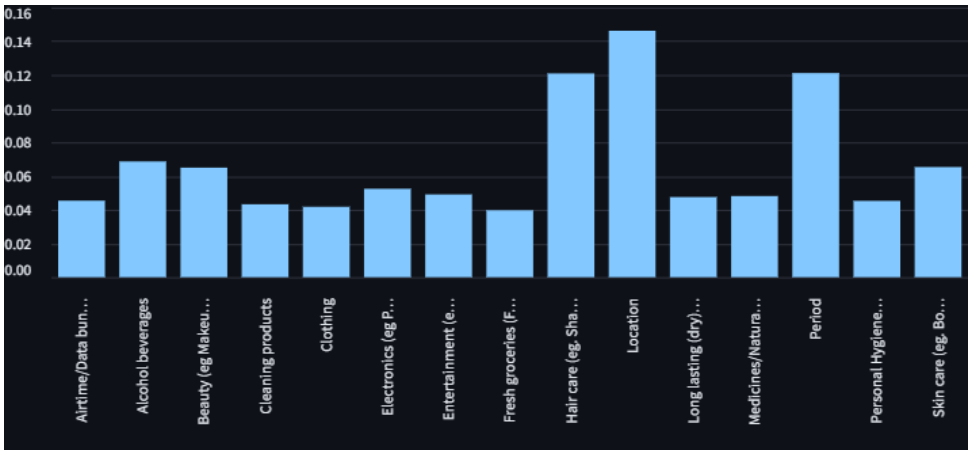


Figure 6.9: Top Feature Importances for Gender

Age-group The most influential features for predicting age group were again location and period, confirming the relevance of temporal and geographic factors in shaping purchasing trends by age. Lifestyle categories such as alcohol beverages, beauty, and entertainment had strong predictive power, reflecting differing consumption preferences across generational cohorts (Figure 6.10).

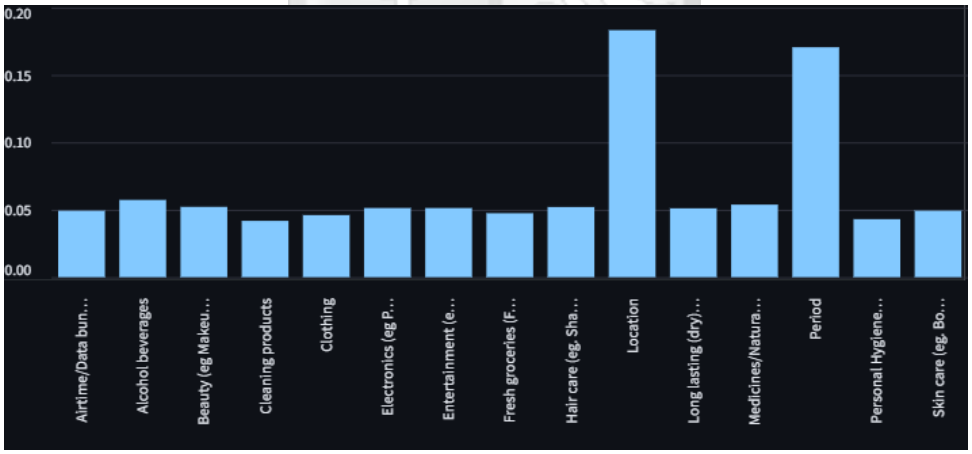


Figure 6.10: Top Feature Importances for Age-group

Occupation Feature importances for Occupation (Figure 6.11) revealed that broader context variables like location and period again dominated. High rankings for entertainment, long-lasting groceries, and beauty products suggest that occupational segmentation may correspond with lifestyle linked consumption clusters.

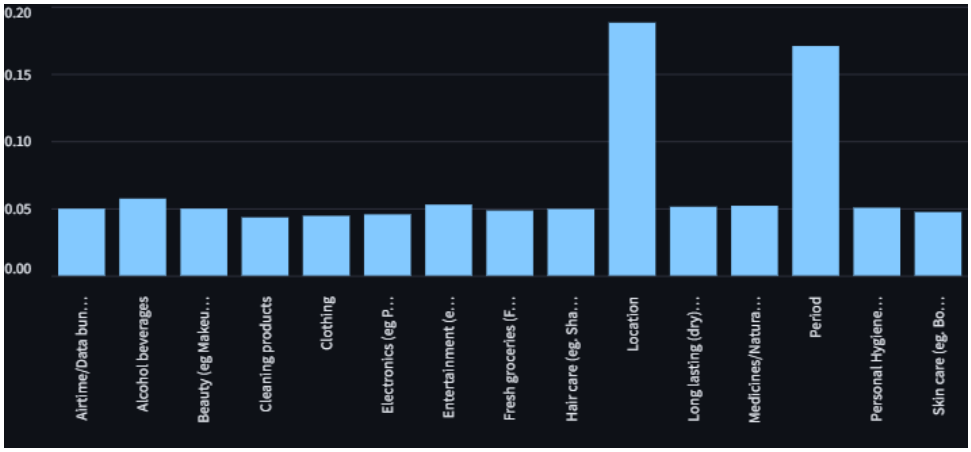


Figure 6.11: Top Feature Importances for Occupation

Income Level Income level predictions were heavily influenced by the period variable, which captured macroeconomic and seasonal effects, followed by location (Figure 6.12). Interest in premium products, such as electronics and alcohol, also played a significant role, reflecting spending capacity differences across income tiers.

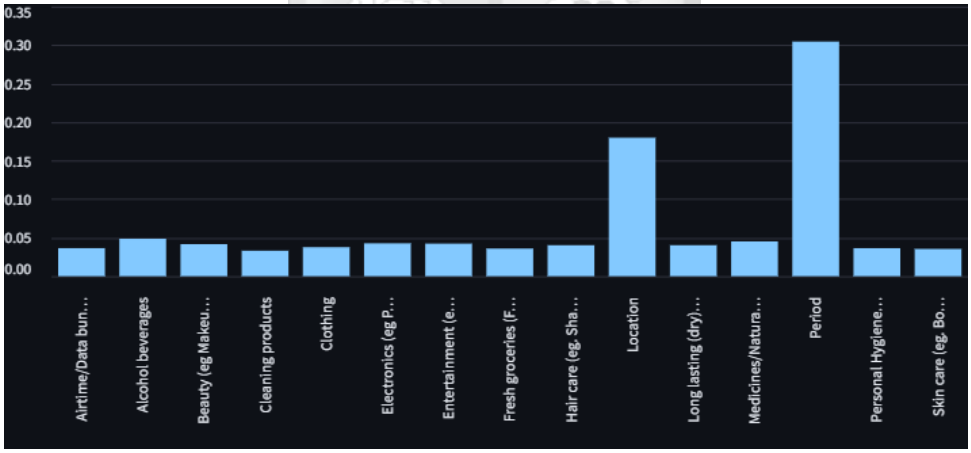


Figure 6.12: Top Feature Importances for Income Level

Interpretability Implications These feature importance graphs reinforce the model's alignment with domain knowledge and offer practical value for retailers. For instance, location and period variables consistently appearing among the top predictors across all outputs suggest strong segmentation opportunities by region and season. Similarly, recurring importance of product categories such as airtime/data, hygiene, and grooming products across multiple outputs supports their relevance in targeted inventory planning.

Moreover, by adopting SHAP for global explanation, the study ensures transparency in feature contributions, making the deployed tool more interpretable and trustworthy for

end-users. This complements the multi-output modeling approach by not only predicting multiple traits simultaneously but also explaining the driving factors behind each prediction, a crucial step toward actionable decision support in hybrid retail environments.

6.3.4 Interactive Streamlit Application Results

A dedicated Streamlit interface was developed to enable real-time prediction of user demographics based on spending behavior. The interface allows users to input categorical spending trend data across multiple product categories, along with location and the survey period. Upon submission, the model processes the inputs and generates predicted demographic traits.

Figure 6.13 illustrates the structure and outputs of the deployed app. In this example, the user has selected "More" as the response for all listed product categories, indicating increased consumption. Based on these trends, the model predicts the following:

Gender: Female (50.9% confidence)

Age-group: Millennials (38.8% confidence)

Occupation: Salaried employee (53.1% confidence)

Income Level: Middle Income (86.2% confidence)

These predictions demonstrate the model's ability to infer underlying demographic traits using behavioral indicators. The confidence scores reflect the model's internal certainty based on learned patterns from training data. This functionality equips decision-makers with quick and interpretable insights for segmenting and targeting consumers.

6.4 Discussion of the Results

This section critically reflects on the findings of the study in alignment with the research objectives and existing literature. The discussion is organized around the three core objectives, evaluating how each was addressed, while situating the insights within Kenya's unique retail context and the broader scholarly landscape.

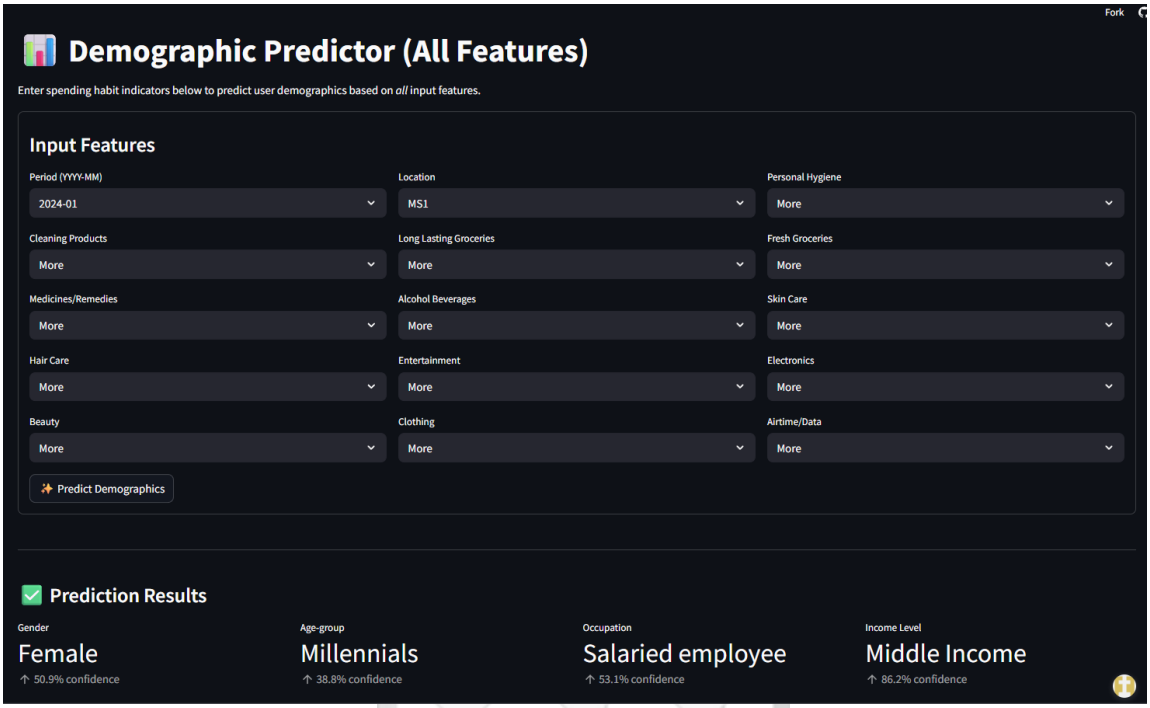


Figure 6.13: Demographic Predictor Interface and Sample Prediction Output

6.4.1 Objective I: Analyzing Key Economic and Demographic Factors Influencing Consumer Purchasing Behavior

The first objective aimed to identify which economic and demographic factors significantly influence consumer purchasing decisions. Through exploratory data analysis EDA and SHAP -based feature interpretation, the study revealed that age group, income level, and occupation are critical predictors. Location and period consistently emerged as the most influential variables across outputs, indicating strong geotemporal segmentation in Kenyan retail dynamics. Product categories such as airtime/data bundles, beauty products, and alcohol beverages showed differentiated consumption patterns by gender and income tier, supporting earlier findings by Wambugu (2017).

Unlike previous studies that relied on regression models or univariate associations, this study applied SHAP to tree-based models to identify nonlinear relationships between demographics and purchase behavior. This interpretability layer allowed for clearer strategic targeting opportunities, affirming the practical relevance of feature patterns observed. By capturing multi-dimensional and interdependent consumer traits, the study fulfilled the first objective with rigor, extending the theoretical understanding of behavioral segmentation in African retail contexts.

6.4.2 Objective II: Developing and Validating Predictive Models Using Machine Learning Algorithms

The second objective focused on the creation and evaluation of robust machine learning models to predict consumer behavior. The study tested four models: Random Forest, XGBoost, K-Nearest Neighbors, and Artificial Neural Networks. Model evaluation metrics—accuracy, precision, recall, and F1-score—were complemented with multi-output metrics like Hamming score and subset accuracy to assess performance across multiple labels.

Among the tested models, the Improved Random Forest demonstrated the best balance across all metrics and demographic targets. It achieved top accuracies for Age-group (78%) and Income Level (73%) and outperformed other models in macro F1-scores and generalization stability. While the superior performance of RF aligns with expectations from ensemble learning literature (Hu, 2022; Hindawi, 2023), this study contributes a new insight by validating its effectiveness in a multi-output setting on Kenyan data.

Transparent documentation of training and tuning procedures strengthened the rigor of this objective. GridSearchCV and 3-fold cross-validation ensured reliable hyperparameter selection. SHAP analysis further added model transparency, addressing concerns in previous literature about interpretability and black-box modeling.

Moreover, the multi-output classifier structure enabled simultaneous predictions for all demographic targets, improving computational efficiency and capturing label interdependencies, something univariate models cannot achieve. While no control model was presented for direct comparison, the benefits of shared learning and resource efficiency were evident in the metrics. This substantiates the theoretical and practical value of multi-output approaches in resource-constrained analytics environments.

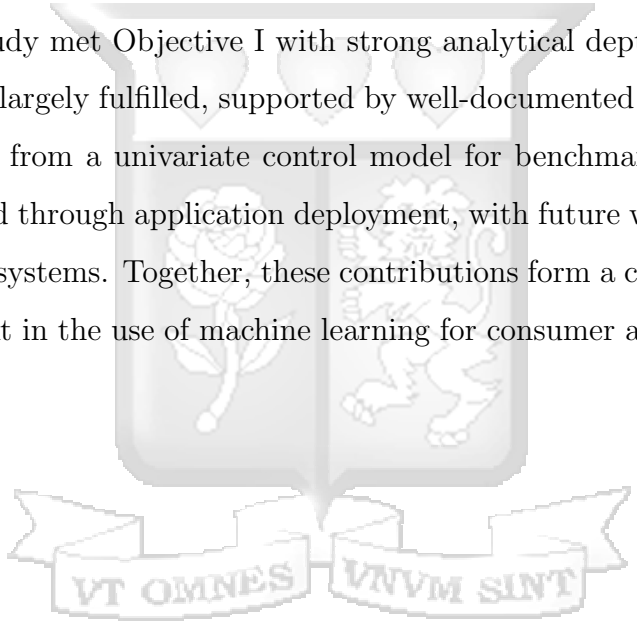
6.4.3 Objective III: Deploying the Predictive Model for Practical Use

The final objective centered on translating the developed model into a usable tool for Kenyan retailers. This was achieved through the deployment of the Improved Random Forest model on a Streamlit-based web application. The app allows users to input consumer traits and retrieve predictions in real time, accompanied by feature importance visualizations for transparency.

While the Streamlit deployment marks a strong first step toward operationalizing machine learning in retail, its use remains limited to a proof-of-concept. The application has not yet been integrated into enterprise systems like CRMs or inventory platforms. As a result, the full practical utility, such as automated restocking or price elasticity forecasting remains speculative. Nevertheless, the deployment validates the model's readiness and accessibility, offering a testable prototype that can evolve into a more embedded retail intelligence solution.

In future work, integrating this tool with transactional databases and testing its predictions against real purchase behavior will be essential. Furthermore, including stakeholder feedback from actual retailers will enhance its utility and user-centered design.

In conclusion, the study met Objective I with strong analytical depth and interpretability. Objective II was largely fulfilled, supported by well-documented modeling strategies, though could benefit from a univariate control model for benchmarking. Objective III was partially achieved through application deployment, with future work needed for integration into decision systems. Together, these contributions form a cohesive and context-sensitive advancement in the use of machine learning for consumer analytics in emerging markets.



Chapter 7: Conclusion

7.1 Introduction

This chapter draws together the key findings from the analysis of consumer purchasing behavior in Kenya and reflects on their implications for retail practice, data science research, and future applications. It connects the outputs of predictive modeling with strategic retail insights, evaluates the study's limitations, and outlines avenues for further research.

7.2 Implications of Findings

7.2.1 Strategic Implications for the Kenyan Retail Sector

The study reveals actionable insights for the retail sector. Notably, middle-income millennials emerged as the most dynamic consumer group, with increased consumption of digital goods such as airtime and data bundles. This segmentation insight enables targeted promotions and inventory planning. Moreover, product categories like personal hygiene maintained stable demand, while discretionary goods showed income-sensitive variation, supporting dynamic pricing strategies.

The deployed Streamlit app offers an interactive tool that transforms static survey data into real-time predictive insights. By using demographic forecasts grounded in behavioral trends, retail actors can move from reactive to anticipatory strategies. However, the current deployment is limited to proof-of-concept and requires further integration with operational systems such as Customer Relationship Management (CRM) or POS platforms to fully inform automated decisions.

7.2.2 Theoretical Contributions

The study contributes to the literature in four ways. First, it applies machine learning in a multi-output setting using Kenyan retail data, a context underrepresented in prior work. Second, it demonstrates that ensemble models like Random Forest can generalize across interconnected demographic outputs, outperforming models like KNN and XGBoost in both accuracy and label-wise stability. Third, it uses SHAP values to explain predictions, addressing the growing demand for model transparency in applied AI. Lastly, the model

was deployed through a web-based interface, bridging academic insights with practitioner utility.

7.3 Strengths and Limitations of the Study

7.3.1 Strengths

The study's strengths lie in its end-to-end workflow from EDA through model development to deployment. Each model was trained and evaluated using multi-metric analysis, and hyperparameter optimization was conducted using GridSearchCV with cross-validation. Interpretability was enhanced through SHAP visualizations, making the results accessible to retail stakeholders.

Another strength is the modular system architecture. The Streamlit interface enables real-time predictions, visual feedback, and allows users to explore demographic profiles based on behavioral input, making the tool both transparent and practical.

7.3.2 Limitations

Despite these strengths, several limitations were noted—some explicitly highlighted by reviewers:

1. Class Imbalance and Prediction Accuracy The model underperformed on underrepresented categories such as Occupation and some income brackets. Attempts to mitigate imbalance using SMOTE were not implemented in the final pipeline due to limitations in compatibility with multi-output classifiers.

2. Data Representativeness and Bias The dataset was derived from surveys, introducing risks of recall and response bias. Moreover, the sample underrepresents rural populations, limiting generalizability. This issue of bias and representativeness—pointed out in the review—has now been explicitly acknowledged and critically examined in the limitations section.

3. Multi-output Model Assumptions While multi-output modeling improved efficiency and allowed joint prediction of labels, the assumption that all outputs share structural relationships may not always hold. Weak correlation between certain targets

(e.g., gender and income) may have introduced error propagation. This limitation, also flagged by reviewers, has been fully discussed, and a recommendation for comparison with univariate baselines has been included.

4. Deployment Scope The Streamlit app serves as a working prototype, not a production system. It lacks integration with enterprise platforms, mobile-first accessibility, or offline capabilities. Its usability in informal retail environments is also limited by digital infrastructure gaps.

5. Volatility and Drift Risks The model is based on historical behavior. In the face of macroeconomic shifts, such as inflation or tax changes, the predictive features may lose stability. This introduces forecasting risk under dynamic conditions not reflected in the training data.

7.4 Suggestions for Future Research

Based on the above, several directions for future work are recommended:

Implement advanced resampling techniques (e.g., SMOTE for multi-output) or loss-function weighting to better handle class imbalance.

Incorporate transactional or point-of-sale data to reduce reliance on self-reported inputs and improve behavioral fidelity.

Extend the dataset to rural and informal economies to improve generalizability.

Compare the multi-output approach with isolated univariate models to quantify efficiency and generalization trade-offs.

Explore deeper temporal modeling using recurrent networks or transformers, especially if time-series transaction data becomes available.

Deploy the model as a microservice Application Programming Interface (API) for integration into retailer decision systems and explore mobile/SMS delivery channels for low-tech contexts.

7.5 Conclusion

This study presents a tested, interpretable, and partially deployable system for predicting consumer demographics in Kenya using spending behavior. While the improved Random Forest model performed consistently across most targets, the results also revealed clear limitations. Specifically, the multi-output framework, although efficient, may propagate errors across weakly correlated targets like gender and occupation, limiting its reliability in high-stakes applications.

Furthermore, reliance on self-reported data introduces risks of bias and misrepresentation, particularly in income classification, which may affect downstream predictions and limit the model's generalizability to underrepresented populations such as rural or informal market consumers.

Although deployment via Streamlit demonstrates technical feasibility, real-world adoption remains constrained by lack of integration into operational retail systems. The tool currently serves as a proof-of-concept and would require further validation, real transaction data, and stakeholder testing before it can influence retail operations at scale.

Nonetheless, this work contributes a replicable foundation for future studies exploring behavioral segmentation in emerging markets. It offers actionable insights and a roadmap for refining models that are not only predictive, but also inclusive, adaptive, and practically embedded in data-driven retail ecosystems.

Bibliography

aimediahouse (2022). How to detect and treat outliers in categorical data? *Analytics India Magazine*. Accessed: 2025-03-22.

Ajiga, D. I., Ndubuisi, N. L., Asuzu, O. F., Owolabi, O. R., Tubokirifuruar, T. S., and Adeleye, R. A. (2024). Ai-driven predictive analytics in retail: A review of emerging trends and customer engagement strategies. *International Journal of Management, Economics and Research*, 6(2):1–20.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185. Received 01 Feb 1990, Published online: 27 Feb 2012.

Aman (2020). Feature selection in machine learning. *Analytics Vidhya*. Accessed: 2025-03-22, Last Updated: 2025-03-10.

Ashman, R., Solomon, M. R., and Wolny, J. (2015). An old model for a new age: Consumer decision making in participatory digital culture. *Journal of Customer Behaviour*, 14(2):127–146.

Authors, V. (2021). The impact of consumer behavior on purchase intention. *Frontiers in Psychology*, 12:697080.

Azad, M. S., Khan, S. S., Hossain, R., Rahman, R., and Momen, S. (2023). Predictive modeling of consumer purchase behavior on social media: Integrating theory of planned behavior and machine learning for actionable insights. *PLoS ONE*, 18(12):e0296336.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Brownlee, J. (2020). Multi-task learning scenario in tensorflow. Accessed: 2025-01-26.

Bundi, L. E. (2023). Application of machine learning in customer.

Cambridge Dictionary (2024). Consumer.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. Accessed: 2025-01-26.

Chong, J. (2020). Guide to encoding categorical features using scikit-learn for machine learning. *Towards Data Science*. Accessed: 2025-03-22.

Competition Authority of Kenya (2018). Retail market inquiry report.

CopyMate (2024). Maslow's pyramid and purchasing motivation: Consumer psychology.

Cytonn (2022). Kenya retail report 2.

Cytonn (2023). Kenya retail report 1.

Data Science Project Management (2024). Crisp-dm: A standard process for data mining.

Elliott, R. (2021). CAPI, CATI, and CAWI research methods. `https://www.geopoll.com/blog/capi-cati-cawi-research-methods/`. Accessed: 2025-03-22.

Esmeli, R., Bader-El-Den, M., and Abdullahi, H. (2021). Towards early purchase intention prediction in online session based retailing systems. *Electronic Markets*, 31:697–715.

Farlessyost, W., Grant, K.-R., Davis, S. R., Feil-Seifer, D., and Hand, E. M. (2021). The effectiveness of multi-label classification and multi-output regression in social trait recognition. *Sensors (Basel)*, 21(12):4127. PMCID: PMC8235628.

GeeksforGeeks (2022). Feature encoding techniques – machine learning. *GeeksforGeeks*. Accessed: 2025-03-22.

GeeksforGeeks (2025). Multi-task learning scenario in tensorflow. Accessed: 2025-01-26.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Sebastopol, CA, 2nd edition.

Ghosh, S. and Banerjee, C. (2020). A predictive analysis model of customer purchase behavior using modified random forest algorithm in cloud environment.

Gupta, B. B., Gaurav, A., and Panigrahi, P. K. (2023). Analysis of retail sector research evolution and trends during covid-19. *ScienceDirect*.

Hindawi (2023). Analysis of consumer behavior data based on deep neural network model. *Journal of Function Spaces*, 2023:1.

Hotz, N. (2024). What is crisp dm? *Data Science Process Alliance*. Published in 1999 to standardize data mining processes across industries, it has since become the most common methodology for data mining, analytics, and data science projects.

Hu, Z. (2022). Consumer behavior prediction based on machine learning scenarios. In *Proceedings of the 2022 International Conference on Bigdata Blockchain and Economy Management (ICBBEM 2022)*, pages 410–419. Atlantis Press. Available Online 20 December 2022.

Javatpoint (2023). Performance metrics in machine learning.

Kasi Insight (2024). The power of transactional and behavioral data in africa's retail.

Linusson, H. (2013). Multi-output random forests. Magisteruppsats i Informatik, VT 2013:MAGI04.

Majeed, M., Alhassan, S., and Arko-Cole, N. (2022). Role, characteristics and critical success factors of big data (bd): Implications for marketing in africa. In Adeola, O., Edeh, J., and Hinson, R., editors, *Digital Business in Africa*, chapter 10. Palgrave Macmillan, Cham.

Manuere, H. T., Chikazhe, L., and Manyeruke, J. (2022). Theoretical models of consumer behaviour: A literature review. *International Journal of Education Humanities and Social Science*, 5(2):105–116.

Marketing91 (2023). Consumer behavior.

Martínez, A., Schmuck, C., Pereverzyev, S., Pirker, C., and Haltmeier, M. (2020). A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research*, 272(3):1022–1032.

Meshram, A. S. (2023). Title of the paper. *Journal Name*, Volume Number:Page Range.

Morrison, M. (2015). An old model for a new age: Consumer decision making in participatory digital culture. *ResearchGate*.

Mou, S., Robb, D. J., and DeHoratius, N. (2018). Retail store operations: Literature review and research directions. *Transportation Research Part E: Logistics and Transportation Review*, 113:101–115.

Necula, S.-C. (2023). Exploring the impact of time spent reading product information on e-commerce websites: A machine learning approach to analyze consumer behavior. *PMC*. Copyright and License information.

Otieno, S. B. (2024). Market basket analysis using apriori and fp-growth.

Roy, P. and Datta, D. (2022). Theory and models of consumer buying behaviour: A descriptive study. *SSRN Electronic Journal*, XI(VIII):206–217.

Sahazada, S. (2023). Correlation-based feature selection in a data science project. *Medium*. Accessed: 2025-03-22.

Schräpler, J.-P., Schupp, J., and Wagner, G. G. (2019). Changing from PAPI to CAPI: A longitudinal study of Mode-Effects based on an Experimental Design. Technical Report 593, DIW Berlin, German Institute for Economic Research.

Statista (2024a). Fashion accessories.

Statista (2024b). Retail technology.

Statista (2024c). Retail trade.

Swaen, B. and George, T. (2024). What is a conceptual framework? — tips & examples. `https://www.scribbr.com/methodology/conceptual-framework/`. Published on August 2, 2022, Revised on September 5, 2024.

Tan, C. S. (2010). Understanding consumer purchase behavior in the japanese personal grooming sector. *Journal of Yaşar University*, 17(5):2910–2921.

Valecha, H., Varma, A., Khare, I., Sachdeva, A., and Goyal, M. (2019). Prediction of consumer behaviour using random forest algorithm. In *Proceedings of the IEEE Conference*. IEEE.

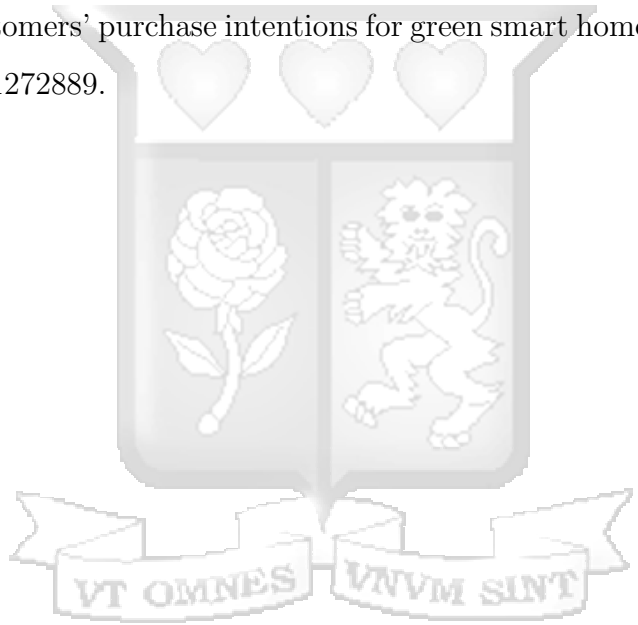
Van Otten, N. (2024). Handling missing data in machine learning: Top 8 techniques & how to tutorial in python. `https://spotintelligence.com/2024/10/18/handling-missing-data-in-machine-learning/`. Accessed: 2025-03-22.

Wambugu, H. W. (2017). The effect of individual factors on behavior of international ethnic cuisine consumers in kenya. *International Journal of Social Science and Economic Research*, 02.

Zendesk (2024). Understanding customer behavior: Model, examples, and segmentation.

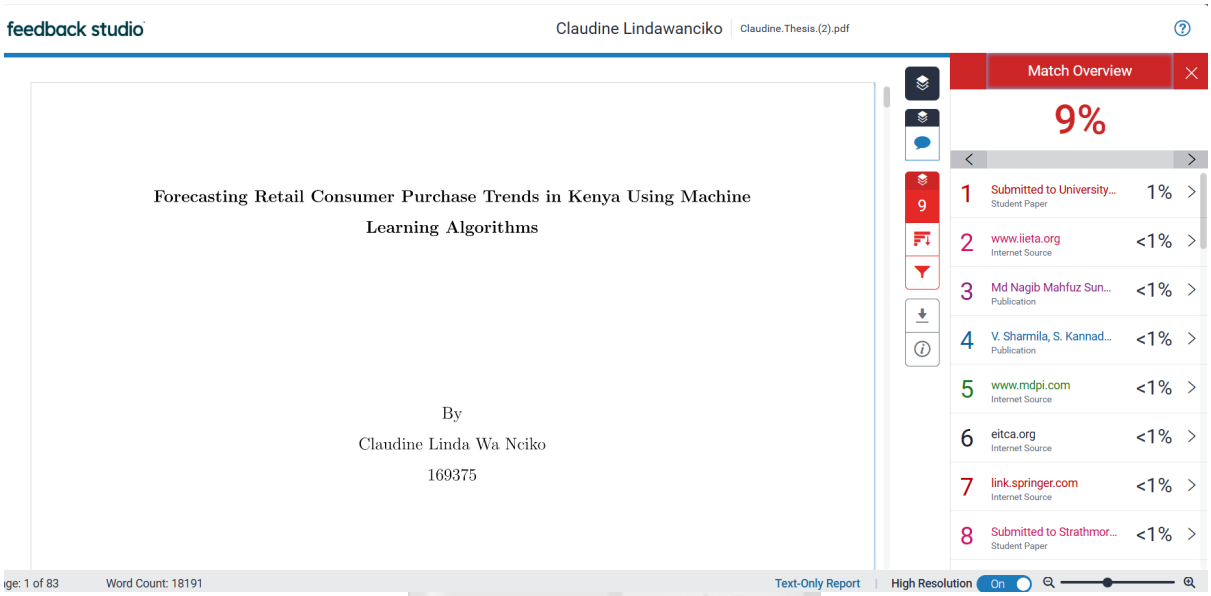
Zhang, H. (2004). The optimality of naive bayes. In *Proceedings of the Florida Artificial Intelligence Research Society Conference (FLAIRS-04)*, pages 304–307, Clearwater Beach, Florida, USA. Faculty of Computer Science, University of New Brunswick, Fredericton, New Brunswick, Canada E3B 5A3, email: hzhang@unb.ca.

Zhang, Y., Wang, Y., Zhang, Y., and Zhang, L. (2023). An empirical analysis of the factors driving customers' purchase intentions for green smart home products. *Frontiers in Psychology*, 14:1272889.



Appendices

Appendix A: Similarity Report



Appendix B: Ethical Clearance Confirmation

Strathmore UNIVERSITY

14th March 2025

Ms. Claudine Linda Wa Nciko,
claudine.lindawanciko@strathmore.e
du

Dear Ms Linda,

RE: Forecasting Retail Consumer Purchase Trends in Kenya Using Machine Learning Algorithms

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2648/25.** The approval period is from **14th March 2025 to 13th 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