A MACHINE LEARNING APPROACH FOR ACCURATE VALUATION OF IMPORTS IN UGANDA

PAUL SENTONGO

A DISSERTATION SUBMITTED TO THE FACULTY OF ENGINEERING, DESIGN AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF THE MASTER OF SCIENCE IN DATA SCIENCE AND ANALYTICS OF UGANDA CHRISTIAN UNIVERSITY

September, 2025

UGANDA CHRISTIAN UNIVERSITY
A Centre of Excellence in the Heart of Africa

Abstract

Accurate customs valuation is central to revenue mobilization, trade compliance, and economic stability in Uganda, where import duties contribute nearly one-third of domestic tax revenue. Yet persistent inefficiencies in conventional valuation methods such as reliance on importer-declared invoice values, outdated price databases, and manual adjudication have resulted in systemic undervaluation, misinvoicing, and annual revenue losses exceeding USD 200 million. This thesis investigates the potential of machine learning (ML) to transform customs valuation by developing and deploying predictive models trained on more than 70,000 import declaration records from Uganda Revenue Authority's ASYCUDA system (2020–2024).

Three supervised ML algorithms; Random Forest, Extreme Gradient Boosting (XGBoost), and Artificial Neural Networks (ANN) were implemented following a rigorous pipeline that included exploratory data analysis, feature engineering, and model optimization. All models demonstrated strong predictive performance ($R^2 > 0.93$), with Random Forest achieving near-perfect accuracy ($R^2 = 0.997$, MAE = UGX 560.35, RMSE = UGX 1,868.23). Compared to Uganda's current average based approach (MAE = UGX 124,797.76), this represents a 99.55% reduction in error, underscoring the transformative capacity of ML for valuation precision.

Beyond model benchmarking, the study contributes technically by operationalizing the Random Forest model into a Streamlit based prototype web application, offering real-time decision support for customs officers. Empirically, it provides the first quantified evidence of ML's potential to address valuation fraud and inefficiencies in Uganda. Practically, it establishes a replicable framework for low-resource settings, integrating ML with existing trade platforms such as ASYCUDA.

The findings have significant policy implications: adopting ML-driven valuation can curtail revenue leakages, enhance compliance with WTO Customs Valuation Agreements, and support Uganda's Vision 2040 and National Development Plan III goals for domestic revenue mobilization. Limitations such as reliance on secondary data, exclusion of informal trade, and simulation-based deployment highlight opportunities for future research. These include incorporating regional datasets, exploring explainable AI techniques (e.g., SHAP, LIME) to improve transparency, and piloting ML integration within operational customs systems.

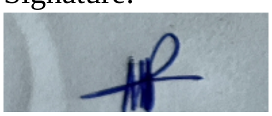
This thesis thus advances the discourse on AI in public sector modernization, demonstrating that machine learning is not merely a technical innovation but a strategic enabler for fiscal sustainability, trade integrity, and digital transformation in Uganda's customs administration..

Declaration

I, Paul Sentongo, confirm that the work presented in this thesis, titled *"A Machine Learning Approach for Accurate Valuation of Imports in Uganda"*, was independently carried out by myself as part of the requirements for the award of a master's degree.

To the best of my knowledge, this thesis has not been submitted, either in part or in full, for the fulfillment of any academic qualification at any other institution. All external sources and references that have informed this research have been properly cited and acknowledged throughout the document.

Paul Sentongo
Candidate
Signature:



Date: 22nd/09/2025

School of Computing, Faculty of Engineering Design and Technology
Uganda Christian University

Approval

This is to certify that this thesis, titled *"A Machine Learning Approach for Accurate Valuation of Imports in Uganda"*, has been done under my supervision and is now ready for submission.

Mr. Ian Raymond Osolo
Academic Supervisor

Sign:...

School of Computing, Faculty of Engineering Design and Technology
Uganda Christian University

Date: 22nd/09/2025

Dedication

This work is dedicated to my family, whose unwavering support, encouragement, and sacrifices have been the foundation of my academic journey. To my parents, who instilled in me the values of hard work, integrity, and resilience, I owe the strength to persevere through every challenge. To my siblings and loved ones, thank you for believing in me even in moments when the path seemed uncertain.

I also dedicate this thesis to the countless researchers, educators, and practitioners who inspire the pursuit of knowledge for the betterment of society. May this work stand as a small contribution to the broader vision of harnessing science and technology to transform lives in Uganda, Africa, and the world at large.

Above all, I dedicate this achievement to God, whose grace, wisdom, and providence have guided me through every stage of this journey.

Acknowledgments

I am deeply grateful to God for His guidance, grace, and strength throughout this research journey. My sincere appreciation goes to my supervisor, Mr Ian Raymond Osolo, for his invaluable mentorship, patience, and constructive guidance. Your insightful feedback and encouragement shaped not only the direction of this thesis but also my growth as a researcher.

I would like to acknowledge the support of the Uganda Christian University and the Department of Computing and Technology, whose academic environment and resources provided me with the opportunity to pursue this study. Special thanks also go to my lecturers, colleagues, and classmates for the stimulating discussions and constant encouragement that enriched this work.

The Uganda Revenue Authority (URA) deserves recognition for making data accessible for this study. Without this crucial resource, the development and evaluation of machine learning models for customs valuation would not have been possible.

To my family, thank you for your love, sacrifices, and unwavering belief in me. Your support has been my anchor throughout this academic pursuit.

Finally, I extend my gratitude to all my friends, peers, and the wider research community whose direct or indirect contributions inspired and sustained my efforts. This work is, in many ways, a collective achievement.

Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
API	Application Programming Interface
ASYCUDA	Automated System for Customs Data
BCTC	Behavioral Change Tariff Compliance
CIF	Cost, Insurance and Freight
DOI	Diffusion of Innovation
EDA	Exploratory Data Analysis
EACCMA	East African Community Customs Management Act
FOB	Free on Board
GATT	General Agreement on Tariffs and Trade
HS Code	Harmonized System Code
IQR	Interquartile Range
LIME	Local Interpretable Model-Agnostic Explanations
MAE	Mean Absolute Error
ML	Machine Learning
OECD	Organisation for Economic Co-operation and Development
PII	Personally Identifiable Information
RECTS	Regional Electronic Cargo Tracking System
RMS	Root Mean Square
RMSE	Root Mean Squared Error
SHAP	SHapley Additive exPlanations
TOE	Technology–Organization–Environment

URA	Uganda Revenue Authority
USD	United States Dollar
UGX	Ugandan Shilling
UNCTAD	United Nations Conference on Trade and Development
WCO	World Customs Organization
WTO	World Trade Organization

Chapter 1

Introduction

International trade plays a central role in driving Uganda's economic development, yet challenges persist in the way customs valuation is conducted. The existing approaches are largely manual, inconsistent, and susceptible to manipulation and fraud. These limitations contribute to revenue leakage, valuation disputes, and non-compliance with global trade norms.

This research explores the application of machine learning techniques to modernize and optimize the valuation of imported goods in Uganda. By introducing predictive models that are trained on actual import declaration data spanning from 2020 to 2024, this study aims to enhance the accuracy and objectivity of valuation practices.

The core objective is to evaluate the effectiveness of machine learning algorithms when applied to real-world customs data, and to benchmark their performance against traditional valuation systems. Additionally, the study seeks to operationalize these models by integrating them into a web-based application suitable for deployment within the Uganda Revenue Authority's operational framework. This integration is intended to support revenue collection, promote transparency, and facilitate compliance with international trade standards.

1.1 Background to the study

Customs valuation lies at the heart of international trade governance and revenue mobilization. It is the process by which authorities determine the value of imported goods for purposes of taxation, tariff application, and compliance monitoring. Across the globe, customs duties account for a substantial portion of public revenue, especially in developing economies where reliance on international trade taxes remains high. The World Trade Organization (WTO) Customs Valuation Agreement provides a common legal framework for valuation, emphasizing transparency, fairness, and alignment with international best practices. Yet, despite this harmonized legal foundation, implementation in many countries is undermined by systemic undervaluation, misclassification, and trade fraud, which collectively result in massive revenue leakages and distortions in market competitiveness.

In Uganda, customs duties represent nearly 30% of the total tax revenue base, making customs valuation a critical instrument for fiscal stability and financing of development priorities. However, prevailing approaches within the Uganda Revenue Authority (URA) remain largely manual and rule-based, depending on importer-submitted invoices, historical price averages, or static reference databases. These methods are vulnerable to manipulation by traders and clearing agents, as well as to the natural fluctuations of global market prices. Consequently, revenue losses due to undervaluation and fraudulent declarations are estimated to exceed USD 200 million annually, equivalent to a significant percentage of domestic revenue collections. These challenges undermine the government's ability to finance its National Development Plan III and broader Vision 2040 aspirations, which emphasize digital transformation and sustainable economic growth.

Globally, emerging technologies such as machine learning (ML) and artificial intelligence (AI) are redefining governance processes, offering solutions that move beyond human subjectivity to data-driven decision support. In customs, ML models can detect hidden patterns in trade data, forecast fair market values, and flag anomalies for further scrutiny, thereby enhancing both efficiency and compliance. Countries such as India, Brazil, and South Korea have already integrated ML into their customs risk management systems with demonstrable success. In Africa, pilot initiatives in Tanzania and Kenya show the promise of predictive models for revenue assurance and trade facilitation. Against this backdrop, this study situates Uganda's customs valuation reforms within the wider global wave of AI-driven governance, with the aim of exploring how ML can address local inefficiencies while aligning with international standards.

1.2 Problem Statement

Despite the strategic importance of customs to Uganda's revenue base, valuation processes remain plagued by inefficiency, subjectivity, and vulnerability to manipulation. The reliance on invoice-based methods, coupled with URA's use of static average values, creates an environment where fraudulent under-invoicing, misclassification, and transfer pricing thrive. Traders often exploit gaps between actual transaction values and reference prices, thereby depriving the government of critical revenue. The persistence of such practices has made valuation fraud one of the most entrenched forms of tax evasion in the country.

At the institutional level, the ASYCUDA World system, which serves as Uganda's core customs management platform, has limited capacity for predictive analytics. While ASYCUDA provides electronic submission, clearance, and record-keeping, its valuation module is still rule-driven and unable to adapt to the dynamic shifts in global trade. For instance, when commodity prices fluctuate due to global shocks such as the COVID-19 pandemic, supply chain disruptions, or regional currency instability the system lacks the intelligence to automatically recalibrate valuation benchmarks. This rigidity reduces its effectiveness and undermines URA's enforcement capacity.

The consequences of valuation inefficiencies are far-reaching. At a fiscal level, annual revenue losses of over USD 200 million constrain the government's ability to fund public services and infrastructure. At a governance level, subjective valuation erodes trust in URA's systems, fuels perceptions of corruption, and generates disputes between traders and customs officials. At a regional level, inconsistent valuation practices hinder the harmonization objectives of the East African Community (EAC), undermining the Customs Union and Common Market protocols. Thus, there is an urgent need to introduce more adaptive, objective, and evidence-based valuation approaches that can strengthen both revenue mobilization and trade facilitation. Machine learning, with its capacity to analyze large datasets, detect anomalies, and provide real-time predictions, presents a transformative opportunity to address these persistent challenges.

1.3 Objectives of the Study

The overall objective of this study is to design, implement, and evaluate machine learning models for improving customs valuation in Uganda, with an emphasis on bridging the gap between academic research and practical application in low-resource settings.

The specific objectives are:

- To develop machine learning models capable of predicting the value of imported goods using historical customs data.
- To evaluate the performance of these models in comparison with traditional valuation methods currently used by URA.
- To deploy the best-performing model in a prototype web application that demonstrates operational feasibility and potential integration with URA's ASYCUDA system.

These objectives collectively seek to establish not only the technical feasibility of ML for customs valuation but also its practical utility in enhancing compliance, revenue assurance, and institutional modernization.

1.4 Research Questions

This study is guided by the following key research questions:

- Can machine learning improve the accuracy and objectivity of customs valuation in Uganda?
- Which machine learning models offer the best predictive performance in this context?
- What institutional and technical considerations are necessary for the successful deployment of ML-based valuation systems within URA?

1.5 Justification of the study

The justification for this study is threefold: policy relevance, academic contribution, and practical significance. From a policy perspective, accurate valuation is central to Uganda's fiscal sustainability, given the heavy reliance on trade taxes for domestic resource mobilization. This research directly supports the URA's modernization agenda, the Ministry of Finance's commitment to reducing revenue leakages, and the government's broader Vision 2040, which envisions digital transformation as a driver of inclusive growth. It also aligns with international commitments such as the WTO Customs Valuation Agreement and the Trade Facilitation Agreement, both of which emphasize transparency and efficiency in customs operations.

From an academic perspective, there is a scarcity of empirical research on the application of machine learning to customs valuation in Sub-Saharan Africa. While studies in Asia and Latin America have demonstrated success, little is known about how these technologies perform in low-resource environments characterized by data quality challenges, infrastructural constraints, and institutional rigidities. This study therefore fills a critical gap in literature by providing the first quantified evidence from Uganda, thereby contributing to both the global discourse on AI in governance and the local body of knowledge on tax administration.

Practically, this study offers URA and other regional customs authorities a replicable model for adopting ML in operational contexts. The prototype application developed in this research illustrates how predictive models can be embedded into existing systems, providing decision support for valuation officers in real time. Such a tool has the potential to transform enforcement practices, improve compliance, and significantly reduce fraud and revenue leakage. By bridging the gap between theoretical modelling and practical deployment, the study advances the frontier of AI adoption in the public sector.

1.6 Significance of the study

This research is the first of its nature where machine learning techniques have been employed to handle import valuation in Uganda, leveraging real trade data from the ASYCUDA system. In addition to the development of models, the research offers a fully deployed prototype system, hence bridging the gap between theoretical modelling and practical implementation, especially in low-resource trade environments.

Successful implementation of machine learning mechanisms in customs valuation has notable economic, technological and policy implications for Uganda. Economically, improving the accuracy of import item valuations contributes to improved revenue collection, which aligns well with the country's National Development Plan III (NDP III)'s objective of strengthening domestic revenue mobilization. By minimizing undervaluation, the machine learning methods help to reduce revenue leakages, hence supporting the plan's initiatives. From the technological point of view, the research introduces the first machine learning-specific mechanism for Uganda's customs valuation framework. This improves efficiency and transparency in the framework, hence laying a foundation for further research in digital transformation in trade and tax administration.

Overall, the research contributes to the goal of promoting sustainable growth and development as outlined in Vision 2040 Uganda.

1.7 Hypotheses

The hypotheses were formulated based on the premise that introducing advanced technologies, particularly machine learning approaches, could significantly improve the accuracy of import valuations and streamline customs operations.

- **Null Hypothesis (H_0)**: Machine learning techniques yield no statistically significant reduction in Mean Absolute Error (MAE) compared to conventional valuation methods.
- **Alternative Hypothesis (H_1)**: Machine learning methods achieve a statistically significant reduction in MAE when determining the customs value of imported items.

1.8 Scope and Limitations

The research focuses on import valuation in Uganda's customs framework. It considers several categories of imported items from the Uganda Revenue Authority trade data that ranges from 2020 - 2024. The dataset records were collected from the ASYCUDA system with top records including the HS code, port of entry, country of origin and the unit price for each item, which we process in this study. We exclude exports, informal channels of trade and non-dutiable items. The limitations include the potential bias in the historical data since it is recorded in manual entries and this could affect generalisation of the models, intermittent network connectivity and limited computing resources.

1.9 Theoretical and Conceptual Framework

The study is anchored in multiple theoretical perspectives to provide a holistic understanding of technology adoption in customs administration. The Principal–Agent Theory explains the persistent information asymmetry between importers (agents) and customs authorities (principals), where traders have incentives to misreport values for personal gain. Machine learning addresses this asymmetry by reducing dependence on subjective declarations and enabling independent verification of transaction values.

The Diffusion of Innovation (DOI) Theory and the Technology Organization Environment (TOE) Framework collectively explain how URA, as an institution, can adopt and internalize disruptive technologies such as machine learning. DOI emphasizes stages of adoption from awareness to decision, implementation, and confirmation while TOE highlights the interplay between technological feasibility, organizational readiness, and external environmental pressures. Together, they provide a useful lens to analyze URA's path toward AI-driven modernization.

Additionally, Institutional Theory offers insights into how customs administrations, as public agencies, adapt to external demands for transparency, accountability, and modernization. Optimal Tax Theory further justifies the relevance of efficient customs valuation by linking it to principles of equity, efficiency, and fiscal sustainability in taxation. By integrating these theoretical perspectives, the study establishes a comprehensive framework that situates machine learning not only as a technical intervention but also as an institutional and governance innovation.

1.10 Thesis Structure

This thesis is organized into five chapters. Chapter One introduces the background, problem statement, objectives, research questions, justification, and theoretical framework. Chapter Two reviews existing literature on customs valuation, tax fraud, and machine learning applications, situating Uganda's challenges within global experiences. Chapter Three details the methodology, including data sources, preprocessing, model development, and prototype deployment. Chapter Four presents results and discussions, linking findings to research objectives, theoretical underpinnings, and policy implications. Finally, Chapter Five concludes the study by summarizing contributions, offering policy and technical recommendations, highlighting limitations, and suggesting directions for future research.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

The literature on customs valuation and the use of advanced technologies such as machine learning (ML) intersects multiple domains, including international trade law, public finance, governance, and computational intelligence. Customs valuation has long been recognized as a cornerstone of international taxation and trade facilitation, directly influencing government revenue, trade competitiveness, and compliance with global agreements. However, persistent valuation fraud and inefficiencies continue to undermine revenue collection, particularly in developing countries such as Uganda where customs duties remain a dominant source of domestic tax revenue. Against this backdrop, the advent of machine learning and other artificial intelligence (AI) methods has created a new frontier in addressing longstanding governance challenges by leveraging data-driven predictive systems. This chapter critically reviews the existing body of knowledge, starting with conceptual and legal foundations of customs valuation, then exploring empirical evidence of valuation fraud, before turning to the global and regional applications of machine learning in tax administration and trade. The chapter concludes by identifying gaps in literature, particularly the scarcity of empirical work on ML for customs valuation in Sub-Saharan Africa, thereby establishing the relevance of this study.

2.1.1 Customs Valuation: Conceptual and Legal Foundations

Customs valuation is defined under the World Trade Organization (WTO) Agreement on Implementation of Article VII of the General Agreement on Tariffs and Trade 1994 (commonly known as the WTO Customs Valuation Agreement). The agreement prescribes the transaction value method—based on the price actually paid or payable for goods as the primary basis of valuation, with fallback methods applied in cases of fraud, misreporting, or insufficient documentation. These include comparative value, deductive value, and computed value methods, all designed to maintain transparency and fairness in international trade.

While the agreement sought to create a uniform and equitable system, implementation has proven uneven. Developed economies with advanced customs systems have largely integrated digital solutions to support valuation, including electronic databases, risk management tools, and automated verification systems. In contrast, many developing countries, including those in Sub-Saharan Africa, continue to rely on outdated methods that are vulnerable to abuse. Uganda's adoption of the Automated System for Customs Data (ASYCUDA World) marked an important step toward modernization, but as highlighted in Chapter 1, its valuation module remains rule-based, static, and ill-equipped to handle dynamic market fluctuations. This structural limitation underscores the need for new technologies capable of offering adaptive, data-driven valuation benchmarks.

2.1.2 Customs valuation fundamentals

Customs valuation is a pillar of international trade, ensuring revenue optimization, fair trade practices, and full compliance with global trade regulations such as the World Trade Organisation on customs valuation. This directly impacts economies, as in the case of developing countries such as Uganda, where customs duties contribute approximately 30% of the entire tax revenue. Notable discrepancies in declared import values, misclassification of items, and under and overvaluation of goods always undermine revenue collection, causing losses. The reliance on manual verification methods worsens these problems, and this forms the basis of this research, aiming at the development of machine learning-based methods that automate the processes, hence modernising the customs operations.
In this section, we identify gaps in the literature, especially for low resourceful environments like Uganda's and we provide a conceptual guideline to our machine learning methodology approach.

2.1.2.1 Limitations of the traditional valuation methods

The WTO provides 6 methods of valuation as explained in this study already. In theory, they represent a transparent and consistent way of valuing imports, whereas in practice, the customs officials often rely on outdated price databases, manual verifications and the importer's price declarations and as a result, Undervaluation and Misinvoicing: Systematic undervaluation of imports and misinvoicing practices are prevalent, leading to substantial revenue losses reported at approximately \$4.9 billion due to import undervaluation between 2006 and 2015 (Monitor, 2023). Although standardised, the existing traditional methods are often lacking in practice due to their subjectivity, complexity and potential for fraud (Keen, 2019). Misinvoicing and valuation fraud were the major challenges in customs valuation, leading to significant loss of revenue and trade imbalance.
The Uganda Revenue Authority in the annual capacity evaluation document (URA, 2022) noted a tremendous gap in its technological infrastructure. Similarly, Mukasa et al.'s (2023) study noted that only 45% of customs officials had access to real-time market price facts, whilst the other 60 % simply relied on outdated valuation databases. These findings matched with the regional research with the aid of the East African Community (EAC, 2023), which highlighted technological challenges as the principal barrier to the implementation of effective customs management.

2.1.3 Revenue Losses from Undervaluation and Fraud

Undervaluation, misclassification, and fraudulent invoice declarations are among the most persistent forms of tax evasion globally. Studies by the World Bank and International Monetary Fund estimate that developing countries collectively lose billions annually through trade misinvoicing, with Africa alone losing upwards of USD 50 billion each year. In Uganda, the Uganda Revenue Authority (URA) reports annual revenue leakages of over USD 200 million from customs fraud, representing a substantial share of potential tax receipts. These losses weaken fiscal capacity, constrain public investment, and exacerbate reliance on external borrowing.
Empirical research has documented the mechanisms by which undervaluation occurs. For example, traders frequently declare artificially low invoice values for high-value goods such as electronics, vehicles, and textiles. Others exploit tariff classification loopholes by misdescribing goods to fall

under lower-duty categories. In some cases, collusion between traders and customs officials facilitates deliberate under assessment of duties. These practices are sustained by information asymmetries, whereby traders possess more accurate knowledge of true values than the customs authority, reflecting the classic principal–agent problem described in institutional economics.
In Uganda, these problems are compounded by weak data infrastructure, limited enforcement capacity, and reliance on outdated reference price databases. As documented by URA's annual performance reports, the average-value method currently employed often fails to capture real-time price movements, making it easy for traders to exploit outdated benchmarks. Such inefficiencies call for adaptive systems that can draw on large datasets, identify anomalies, and provide objective benchmarks, thereby reducing opportunities for fraud.

2.1.4 Digital Transformation and AI in Tax Administration

Globally, tax administrations are increasingly turning to digital technologies to enhance efficiency, compliance, and transparency. Digital platforms such as e-filing, electronic invoicing, and integrated customs management systems have already transformed revenue authorities in countries like Brazil, India, and South Korea. The integration of artificial intelligence represents the next frontier, particularly in areas requiring pattern recognition, anomaly detection, and predictive modeling.
Machine learning, as a subset of AI, is particularly suited to customs valuation because of its ability to analyze large, high-dimensional datasets and uncover patterns not visible through traditional statistical methods. ML algorithms such as Random Forest, Gradient Boosting, and Neural Networks have been applied in diverse domains including fraud detection, credit risk assessment, and supply chain optimization. In customs, their application has ranged from predicting shipment risk profiles to detecting misclassification of goods and forecasting likely undervaluation. For example, Sharma et al. (2021) reported a Random Forest model in India achieving 89% accuracy in detecting undervalued imports, while studies in Brazil showed that gradient boosting techniques significantly reduced misclassification errors in textile imports.
In Africa, adoption remains nascent but promising. Tanzania piloted an XGBoost-based model to analyze vehicle import valuations, recording a 36% reduction in valuation errors compared to traditional methods. Similarly, Kenya's tax authority has integrated predictive analytics into its customs management, though with limited public documentation of outcomes. These experiences demonstrate both feasibility and potential for scaling machine learning in customs operations, while also underscoring the need for rigorous empirical studies to assess model accuracy, institutional readiness, and governance implications.

2.1.5 Literature on Customs Valuation in Uganda

Uganda's customs valuation system, administered by URA through ASYCUDA World, has been subject to critique in both academic and policy literature. Studies by the Economic Policy Research Centre (EPRC) and the African Tax Administration Forum (ATAF) highlight persistent undervaluation in sectors such as automobiles, petroleum products, and consumer electronics. These studies attribute inefficiencies to weak enforcement, limited technical capacity, and the absence of adaptive valuation tools.
Academic research in Uganda has tended to focus on broader tax administration challenges, such as compliance costs, tax morale, and the impact of electronic systems on efficiency. Few studies have specifically examined customs valuation from a machine learning perspective, creating a significant knowledge gap. The limited research that exists has concentrated on descriptive analysis

of revenue losses or policy discussions on compliance strategies, without deploying computational models to predict or detect undervaluation. This study therefore contributes a unique empirical perspective by applying ML directly to URA's ASYCUDA dataset, benchmarking models against existing methods, and demonstrating a deployable prototype.

2.2 Challenges in Traditional Valuation Systems

The literature reveals widespread limitations associated with traditional valuation practices. These systems rely predominantly on declared invoice values and human interpretation of pricing trends. However, global studies have shown that reliance on static price reference databases is inadequate in addressing the complexities of modern trade. Undervaluation, invoice fraud, and misclassification remain prevalent, undermining the integrity of customs operations. For example, a 2023 report by Uganda's Ministry of Finance indicated that over UGX 2.7 trillion was potentially lost in under-declared imports between 2018 and 2022. Several international case studies have similarly illustrated how discretionary customs assessments lead to inconsistent revenue collection and poor enforcement outcomes. Despite the existence of six WTO-sanctioned valuation methods including transaction value, deductive, computed, and fallback methods most developing countries primarily rely on a single approach, thereby limiting flexibility and adaptability.

2.2.1 Machine learning in customs trade

Machine learning methods have become a force in transforming trade operations worldwide, particularly in addressing undervaluation and detecting fraudulent schemes where these systems have been successfully installed. There has been notable success, especially in anomaly detection and improving accuracy in the valuation and classification of goods. For reference, in India, Sharma et al. (2021) developed and deployed a random forest model to identify undervalued shipments while comparing declarant values with the trade database values, and their model achieved a remarkable accuracy of 89% in identifying price discrepancies, enabling the tax authorities to recover an estimated $47 million in underpaid duties within 6 months of the model's implementation.

Similarly, Tanzania's implementation of the XGBoost model to analyse import data registered success, with valuation errors reduced by a notable 36% with the model identifying irregularities in declarations and misclassification of goods by their HS codes (Kiprop, 2023) which highlights the potential of machine learning to handle non-linear relationships and high-dimensional data that is inherent in customs trade records. However, supervised learning methods have a notable limitation in low-compliance environments such as Uganda, where labelled datasets of recorded fraud cases are scarce (Ferreira et al., 2020). They clearly warn against models trained on such incomplete or biased datasets, which were evidenced by Nigeria's customs union, where a neural network flagged 20% of legitimate shipments due to overfitting on outdated fraud patterns.

This emphasized the need to employ methods that incorporate both unsupervised and supervised learning methods, such as clustering, to reduce over-reliance on labeled data while adapting to ever-changing fraud tactics.

2.3 The Emergence of Machine Learning in Trade Administration

Machine learning has emerged as a transformative tool in public sector data analysis. In the customs domain, ML is increasingly being used to automate HS code classification, detect fraudulent declarations, and predict transaction values. The core strength of ML lies in its ability to identify complex patterns in large datasets without relying on predefined rules. Studies from India, China, and Brazil have shown promising results in using algorithms such as Random Forest, Gradient Boosting, and Artificial Neural Networks to automate valuation tasks (Sharma and Kumar., 2021) For instance, they developed a model using XGBoost that flagged undervalued imports with over 92% accuracy. Their approach relied on structured trade datasets enriched with metadata such as product weight, quantity, and origin. Similarly, (Szabo, 2017) applied logistic regression to customs data from the European Union, demonstrating how statistical learning could detect misclassified HS codes with high precision.

2.3.1 Predictive Analytics in Customs

Predictive analytics is the application of statistical techniques and machine learning (ML) models to analyse past data and forecast future trends and results. Predictive analytics is critical in trade facilitation because it improves decision-making, lowers risks, and optimises the efficiency of international trade operations. Predictive algorithms can forecast market trends, identify potential logistics bottlenecks, and improve customs risk assessments by using massive amounts of trade-related data.

Predictive analytics is critical in trade facilitation because it improves decision-making, lowers risks, and optimises the efficiency of international trade operations. Predictive algorithms can forecast market trends, identify potential logistics bottlenecks, and improve customs risk assessments by using massive amounts of trade-related data.

Time series analysis, regression models, and classification algorithms are the most common predictive models utilised in trading processes today. Time series forecasting, for example, is often used to forecast shipment arrivals and delays using historical patterns as well as external factors such as weather and geopolitical events. Regression models are used to assess the relationships between numerous trade parameters and their influence on logistics performance. Furthermore, categorization algorithms assist in identifying high-risk transactions by analyzing trends in trade documents, shipment data, and compliance histories.

Despite these advancements, the growing complexity and volume of international trade necessitate the development of more sophisticated predictive models capable of managing large and diverse datasets. This has sparked a rising interest in using sophisticated ML approaches to improve the effectiveness of predictive analytics in streamlining customs trade operations.

The application of machine learning techniques in customs gained traction in recent years, and a study by (Jentsch et al., 2019) noted the potential of machine learning in predicting accurate customs values based on historical customs records. Their study showed promising results in figuring out cases of undervaluation. Similarly, (Vetter et al., 2021) explored the application of ensemble learning algorithms for detecting customs fraud whilst combining more than one algorithm to enhance accuracy and robustness over the traditional rule-based techniques.

Applications of these ML algorithms have shown considerable gains in predicting trade patterns and risk assessment. However, integrating these models into existing trade systems presents several obstacles, which will be described in the following section.

2.3.2 HS Code Classification Using the Naive Bayes Algorithm

The Harmonized System (HS) code is a standardized six-digit numerical identifier used internationally to classify traded goods. While the base six digits are globally recognized, individual countries may extend the code to 8, 10, or even 12 digits to suit their domestic tariff and statistical requirements.

HS classification refers to the process of assigning the correct code to a product based on its attributes and intended use. Accurate classification ensures appropriate tariff rates are applied and facilitates trade data analysis. Errors in HS code assignment can result in revenue loss, trade disputes, or retroactive penalties.

Recent research has explored the potential of machine learning in improving the precision of HS code assignment. In particular, the Naive Bayes algorithm a probabilistic classifier known for its simplicity and computational efficiency has shown promise for multiclass classification tasks involving trade data. By leveraging the conditional probabilities of feature values, the model can predict the appropriate HS code for a given item.

For example, a study by (Muslim, 2022) demonstrated that applying Naive Bayes to customs datasets yielded a classification accuracy of 99.97%, with a minimal error rate of 0.03%. These results suggest that automated classification can not only improve operational efficiency but also help safeguard government revenue by minimizing tariff miscalculations and duty evasion.

The broader application of such techniques aims to reduce the likelihood of misclassification, avoid fines and penalties associated with incorrect declarations, and provide decision support for customs officers. Integrating data-driven models into classification workflows can enhance the transparency and consistency of trade valuation processes.

2.3.3 ML-Powered Customs Operations case studies

(CHEN, Z., 2024) highlighted the revolutionary impact of machine learning when it was applied to a customs dataset, discovering trends and anomalies. He goes on to say that by automating the valuation process, machine learning methods were able to improve risk assessment and detect discrepancies in declared values, allowing customs authorities to make more informed decisions, reduce errors, and ultimately lead to a more accurate and efficient valuation of items.

2.3.4 Brazil

Since 1997, the country's import declarations have been logged in Siscomex, an integrated commerce system. If errors are discovered during inspection by a customs officer, a corrected copy of the declarations is saved, and both copies are retained indefinitely. The AI system utilised is SISAM, which learns from both versions of the dataset to enhance mistake detection. To handle the dataset's many properties, Bayesian approaches with smoothing hierarchies are used. It uses both supervised and unsupervised learning approaches to adjust to legislation changes without the need for retraining, allowing the system to maintain high accuracy in its classifications and predictions.

If more than 75% of errors are identified in an import declaration, SISAM advises a physical examination by customs officials. The system's handling of complex errors currently outperforms random selection as noted by chen (2024).

2.3.5 China's Adoption of Artificial Intelligence in Customs Operations

In response to growing regulatory demands and limited inspection resources, China Customs has increasingly embraced advanced technologies to enhance efficiency and accuracy in border management. One key area of innovation is the integration of artificial intelligence (AI) into non-intrusive inspection (NII) systems.

An AI-powered image recognition framework has been developed to assist in the automatic analysis of X-ray and computed tomography (CT) scans taken during container inspections. By training on extensive historical datasets from devices such as the H986 scanner, the system learns to identify goods and risk profiles, developing its own detection algorithms over time. The purpose is to reduce reliance on manual interpretation and eventually enable automated decision-making for inspection clearance.

The deployment of this system has already yielded measurable results. Following the introduction of autonomous algorithm selection, local IT infrastructure saw improvements in processing speeds. Additionally, automated image recognition accuracy on NII devices improved by approximately 5%, while the false-positive rate dropped by 8%. Similar gains were observed with CT scanners, with accuracy improving by 6% and error rates falling by nearly 5% (WCO News, 2024).

Another innovation involves the Intelligent Passenger Face Recognition System, which leverages biometric identification for real-time risk assessment. This system is linked with thermal detection for quarantine checks and operates across three control zones: alert, processing, and re-examination areas. High-risk passengers, including individuals on watchlists or those flagged for repeated border crossings, are automatically identified and flagged for inspection using handheld devices equipped with facial recognition technology.

The accuracy of this facial recognition system now exceeds 99%, significantly enhancing the ability to intercept smugglers and enforce border controls. It has contributed to the apprehension of several trafficking networks. Simultaneously, because it is non-intrusive, it facilitates quicker and smoother processing for low-risk travelers.

China Customs continues to explore further applications of AI in areas such as risk profiling, behavioral analysis, and automated compliance evaluation. These innovations not only strengthen enforcement but also contribute to more seamless and intelligent border management experiences.

2.3.5.1 Belgium: Analyzing Behavioral Shifts in Response to Tariff Adjustments (BCTC)

This case study focuses on the behavioral impact of European Union (EU) customs tariff changes on the import activities of economic operators. The main objective is to identify fraudulent practices that may emerge following the implementation or escalation of tariff barriers. Such policy interventions are generally aimed at protecting the EU's internal market, preserving local manufacturing sectors, and mitigating the effects of global competition.

However, in response to these measures, certain actors may engage in deliberate misreporting to avoid paying the applicable duties. These tactics often include the misrepresentation of the product's country of origin, incorrect product classification codes, or both. The resultant revenue loss and distortion of fair competition undermine the intended objectives of these trade policies.

The study leverages historical import data to detect abnormal shifts in import behaviors. It seeks to identify deviations in an importer's profile that significantly diverge from previously observed patterns prior to the imposition of new tariffs. This analysis is centered on recognizing unusual activity that may signal an attempt to bypass the newly introduced duty measures.

By capturing such behavioral changes, the initiative aims to support EU customs authorities in fraud detection, enabling early interventions and more targeted compliance audits.

2.3.5.2 Uganda: The Regional Electronic Cargo Tracking System (RECTS)

In 2017, a trilateral summit involving the heads of state from Uganda, Kenya, and Rwanda led to a joint resolution to implement a shared digital cargo tracking framework. This initiative, known as the Regional Electronic Cargo Tracking System (RECTS), was established to enhance the security and transparency of goods in transit across borders within the East African region.

Administered collaboratively by the revenue authorities of the three nations, RECTS ensures secure, real-time end-to-end cargo monitoring from point of departure to final destination. It is designed to reduce cargo-related fraud, improve operational efficiency, and reinforce trust in regional trade logistics. The system is maintained exclusively by Bsmart Technologies, which provides the technological infrastructure and system integration (Julius and Christabel, n.d.).

RECTS is composed of four key components. These include: (1) electronic seals for dry and fuel cargo that transmit GPS signals; (2) customs officers stationed at release terminals to initiate cargo tracking; (3) a Centralized Monitoring Centre that oversees all movements; and (4) twelve Rapid Response Units strategically deployed along the transit corridors to address emergencies or deviations from authorized routes.

Additionally, a dedicated reconciliation team audits the transit documentation and verifies compliance with regional transportation regulations. This team is responsible for identifying discrepancies, resolving malpractice issues, and maintaining the integrity of transit flows across partner states.

The implementation of RECTS marks a significant step toward modernizing East Africa's trade monitoring infrastructure by integrating real-time technology into customs enforcement.

FIGURE 2.1 A customs officer inspecting a cargo container equipped with the RECTS tracking system, ensuring secure and transparent transit monitoring. Picture Source: URA Vol. 1, Issue 1 FY 2015/16

Benefits realised after the implementation of the project include:

2.3.5.3 Benefits of Implementing the RECTS Project

Following the deployment of the RECTS initiative, several notable improvements have been observed across the region's customs operations and logistics systems:

- **Shortened Transit Time:** The average duration of cargo movement decreased to between three and four days, improving delivery efficiency.
- **Revenue Enhancement:** The deployment of Rapid Response Units (RRUs) has led to an increase in revenue collection by intercepting non-compliant cargo.
- **Improved Data Integrity:** Strengthened control measures have enhanced the reliability and security of data associated with cargo tracking.
- **Enhanced Regional Collaboration:** The formation and coordination of joint technical working groups have fostered greater cooperation among member states.
- **Real-time Monitoring:** Live tracking capabilities have reduced incident response times to within 60 minutes, promoting swift intervention.
- **Reduction in Diversion Cases:** The incidence of cargo diversion has decreased significantly due to closer route monitoring.
- **Lower Business Costs:** Greater operational efficiency and minimized delays have contributed to a reduction in costs incurred by traders and logistics providers.

2.3.5.4 Ongoing Enhancement Initiatives

The technical teams from the participating revenue authorities continue to assess and refine the RECTS system to ensure it remains aligned with evolving trade and security demands. These joint working groups regularly evaluate the platform's functionality and suggest operational improvements.

One of the major enhancement plans involves expanding the network of Rapid Response Units across all transit corridors to provide broader coverage and faster enforcement. To achieve this, there is an ongoing effort to ensure that all RECTS components including personnel, technology, and infrastructure are adequately resourced and supported through continuous training and system upgrades. These developments aim to reinforce cargo oversight and boost the system's overall resilience and performance.

2.3.5.5 The DATE model

Chen and Liu (2022) developed this, which resulted in a significant improvement in customs data processing. "DATE (Dual Attentive Tree Embeddings) demonstrated that tree-structured attention models can capture relationships in customs declarations, improving accuracy by 28 percent compared to traditional machine learning methods."

The model was introduced by the World Customs Organization's BACUDA project, and it represents a significant shift in customs fraud detection, with notable findings such as superior performance over the XG Boost model, particularly in classification tasks. This model was also successful in Nigeria (FSI, 2023). In addition, a more straightforward web-based user interface was implemented to prevent item misclassification effectively.

FSI (2023) conducted a study on the use of artificial intelligence in developing Nigeria's customs system, and the DATE model outperformed existing traditional methods. To avoid the misclassification of imported items, relevant users are provided with a web link where they can enter the item name and another unique identifier, and the model returns the appropriate class to which those items belong. This study closes a gap in the existing literature by developing a machine-learning model to address valuation issues within Uganda's customs regime.

2.4 Relevance to the Ugandan Context

In Uganda, the potential for machine learning in customs has not yet been fully realized. While the URA has experimented with digital declarations and data analytics, there is limited evidence of predictive modeling being used in valuation decision-making. The country faces unique challenges, including limited technical infrastructure, inconsistently structured trade data, and minimal budget allocations for AI-driven innovation. However, these challenges also present an opportunity to design context-sensitive solutions that address local capacity gaps. Applying ML models to Uganda's own historical import data can offer a more accurate, objective, and scalable approach to customs valuation. Moreover, by integrating predictive tools into existing platforms like ASYCUDA, valuation officers can benefit from real-time decision support, especially for high-risk or frequently misdeclared items.

2.4.1 Significance of machine learning in customs trade.

Similar innovations have been reported and explored mainly in the sub-Saharan African countries, such as the Kenya Revenue Authority's AI study for customs fraud detection (Mutua et al., 2021) and Ghana's e-customs portal that integrates machine learning methods for HS codes classification (Mensah & Owusu, 2022). However, there hasn't been any documented research about the prediction of unit prices of imports for Uganda's context, and this has formed the basis for this study.
Machine learning (ML) has emerged as a strong technology that has the potential to transform trade facilitation. ML identifies patterns, trends, and anomalies in large datasets by employing data-driven algorithms. This functionality is especially useful in trade environments that create massive amounts of data daily.
Automated risk assessment, real-time customs clearance time prediction, and supply chain operation optimization are all examples of machine learning applications in trade facilitation. Predictive models, for example, can analyse past customs data to identify shipments that are likely to have delays or violations, allowing customs officials to better deploy resources. Similarly, machine learning-based risk assessment models can aid in the identification of high-risk consignments, boosting inspection accuracy and efficiency. These examples highlight machine learning's transformational potential in improving trade processes and reducing inefficiencies.
Enhanced efficiency:The implementation of machine learning systems in customs trade streamlines operations by reducing costs and improving efficiency. While analyzing large trade datasets, these systems reduce human error, automate customs clearance, and reduce delays in document processing, hence enhancing efficiency.
Anomaly detection: Machine learning systems often support classifying trade transactions through outlier detection mechanisms.

2.4.2 Problems associated with deployment of ML solutions in low resource economies

Data quality and its availability: The use of predictive analytics and machine learning models in trade facilitation faces several challenges, the most significant of which are data quality concerns and system integration. Machine learning systems frequently rely on high-quality data to function properly. Biased data produces incorrect results and conclusions, undermining the ML system's reliability. Customs trade employs a variety of data sources, including customs records, shipping manifests, and regulatory documents. These datasets' discrepancies, missing values, and inconsistencies can all have a significant impact on the predictive model's accuracy.

Ethical and Regulatory Considerations: The use of machine learning models can inadvertently perpetuate biases that are frequently present in training data, resulting in unfair outcomes. Furthermore, the sensitivity of trade data raises privacy and security concerns, which limit the availability of large datasets for model training and analysis.

Technological Constraints: Special skills and knowledge are required for the successful implementation of machine learning systems. The scarcity of skilled professionals in artificial intelligence impedes the adoption and effective implementation of these technologies. This necessitates significant investment in infrastructure, particularly in high-performance computing resources and data storage systems.

Public Perception: Building public trust in these artificial intelligence systems is critical. There is a constant fear of job loss as AI automates many tasks. As a result, addressing the societal impact of AI systems necessitates the creation of programmes that up-skill the workforce and provide opportunities for employees to learn new skills.

AI Regulation: The absence of a structured legal framework for data collection and use of electronic data. The lack of standardization in data formats and communication protocols complicates ML model integration. In addition, traditional trade systems frequently lack the technical infrastructure required to support real-time machine learning predictions.

These issues highlight the importance of better data management techniques and robust integration frameworks for properly deploying ML in trade facilitation.

2.4.3 Theoretical Perspectives

Several theoretical frameworks underpin the literature on customs valuation and technology adoption. The Principal–Agent Theory is frequently invoked to explain valuation fraud, emphasizing the asymmetry of information between traders (agents) and customs authorities (principals). Traders have incentives to misreport values, while customs authorities face monitoring challenges. Machine learning reduces this asymmetry by leveraging data to independently validate transaction values.

The Diffusion of Innovation (DOI) Theory provides insights into how institutions adopt and internalize new technologies. For URA, ML adoption requires moving through stages of awareness, persuasion, decision, implementation, and confirmation. Similarly, the Technology–Organization–Environment (TOE) Framework highlights the interplay of technological readiness, organizational capacity, and environmental pressures. Uganda's push for digital transformation under Vision 2040, combined with international obligations under the WTO Valuation Agreement, creates both internal and external pressures conducive to adoption.

Finally, Institutional Theory enriches the discussion by examining how public agencies adapt to demands for transparency and modernization. In Uganda, international donor support, regional integration through the East African Community, and domestic pressure for revenue mobilization all constitute institutional drivers for customs reform. Integrating these theoretical perspectives allows this study to situate ML not only as a technical innovation but also as part of a broader institutional and governance transformation.

2.5 Gaps in the Existing Literature

Despite the growing global evidence on the use of machine learning in customs administration, critical gaps remain in the African and Ugandan context. First, few empirical studies have applied predictive ML models to real customs datasets in Sub-Saharan Africa, leaving uncertainty about their accuracy and feasibility in low-resource environments. Second, little is known about how such models can be operationalized into existing systems like ASYCUDA, especially given infrastructural constraints. Third, there is scant research on the institutional, organizational, and behavioral factors that influence adoption of AI in public revenue authorities. These gaps highlight the originality and relevance of this study, which not only benchmarks model performance but also deploys a prototype, providing both technical evidence and practical insights.

2.6 Conclusion

The literature establishes customs valuation as a critical but vulnerable component of trade governance, subject to manipulation through undervaluation and misclassification. While international frameworks such as the WTO Customs Valuation Agreement set legal standards, implementation challenges persist, particularly in developing economies. Machine learning has emerged as a promising tool for enhancing valuation accuracy, with successful applications documented in Asia and Latin America and pilot initiatives underway in Africa. However, empirical evidence from Uganda is limited, and no study to date has operationalized ML models within URA's customs framework. This gap underscores the need for this study, which seeks to demonstrate the feasibility, performance, and practical implications of ML adoption in Uganda's customs valuation system.

Chapter 3

METHODOLOGY

3.1 Introduction

This chapter presents the methodological framework employed in the study, from data collection and preparation through model development, evaluation, and prototype deployment. The approach adheres to Guo's (2018) seven-step machine-learning development paradigm, which ensures a logical progression from data collecting to model deployment.The objective was to design and rigorously test machine learning (ML) models for customs valuation, benchmark them against existing practices at the Uganda Revenue Authority (URA), and demonstrate operational feasibility through a functional prototype. Methodological decisions were informed by both theoretical considerations such as the Principal Agent problem and the Technology Organization–Environment (TOE) framework discussed in Chapter 1 and practical constraints associated with data availability, computational resources, and institutional readiness. The chapter is structured as follows: data sources and description, data preprocessing and feature engineering, model selection and development, evaluation procedures, deployment of the prototype application, and ethical considerations.

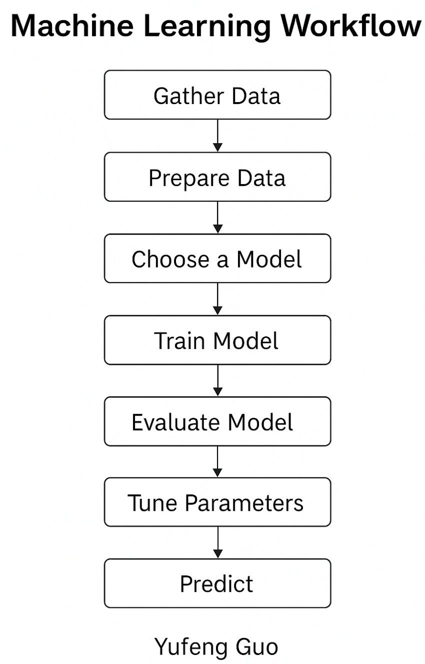


FIGURE 3.1 A flowchart of the steps of machine learning as per Dr Yufeng Guo.

3.1.1 Data Sources

The primary dataset for this study was obtained from the Uganda Revenue Authority's Automated System for Customs Data (ASYCUDA World), covering the period 2020 to 2024. ASYCUDA is a standardized customs management system developed by UNCTAD and deployed globally to facilitate electronic processing of customs declarations. The dataset comprised more than 70,000 import declaration records, representing a wide range of commodities including automobiles, electronics, textiles, and agricultural products. Access was granted under URA's research collaboration framework, with strict adherence to confidentiality agreements and ethical data use protocols.

3.1.2 Data Structure

Each record in the dataset included variables relevant to customs valuation: Harmonized System (HS) codes, declared transaction values, quantities, country of origin, freight and insurance costs, duty-paid values, and exchange rates at the time of clearance. Categorical variables such as HS codes captured commodity classifications, while continuous variables included unit prices, freight charges, and duty amounts. Metadata such as declaration dates and import channels (air, sea, or land) provided temporal and logistical context.

3.1.3 Data limitations

As with most administrative data, the ASYCUDA dataset contained missing values, inconsistencies, and potential misclassifications. Moreover, it excluded informal cross-border trade flows, which account for a substantial share of Uganda's total trade volume. These limitations, discussed further in subsequent sections, were addressed through data cleaning, imputation, and careful feature engineering.

3.1.4 Descriptive statistics

These statistics enabled us to gain preliminary insights about the dataset. This allowed for better understanding of the distribution of variables, quick analysis of values to detect extreme values and abnormal distribution. Understanding the central tendencies of the variables allowed for an organised model development process.
Observation Data Distribution: Most numerical features showed a right-skewed distribution, as evidenced by their means being higher than the medians (such as: Unit_Price_UGX, Value_per_-kg, FOB_per_kg). This hinted at the presence of high-value outliers, which would influence model behavior and initial recommendation required robust scaling.
Target Variable(Unit_Price_UGX): The unit price in UGX values ranged from 62.20 to 729,221.44, with a high standard deviation of 163,745.99, which indicated a substantial variance. Its median value (86,718.70) was significantly lower than the mean (151,323.88), confirming skewness.
Freight and Insurance Costs: Both Freight_USD and Insurance_usd exhibited wide variability, with maximum values exceeding 1,000 USD and 135 USD respectively. However, Freight_per_kg and Insurance_per_kg show that some shipments had minimal to no costs per kilogram, suggesting variable valuation or subsidization.
Mass and Quantity: The Gross_Mass_kg and Net_Mass_kg were highly correlated, with average values of 786 kg and 726 kg respectively. The mass-based price features (FOB_per_kg, Value_-per_kg) were also spread out significantly, which may affect model sensitivity to weight-based features.
Time Span: The data ranged from 2020 to 2024, which allowed for time series analysis using the Year and Month columns. This temporal information would be used to detect policy changes, macroeconomic shifts, or demand cycles.
Taxes and Valuation: The Tax_Rate had a relatively small standard deviation and ranged between 5% and 25%, which indicated tiered tax policies.

Distributional Characteristics of the target variable: (Unit Price)
To assess the shape of the distribution of key valuation variables, we computed the skewness and kurtosis of the dataset. The calculated values were as follows:

- **Skewness**: 1.519
- **Kurtosis**: 1.654

To further explore the distribution of the target variable, statistical tests were ran on the target variable to determine the skewness and Kurtosis.
A skewness of 1.519 indicates a moderately right-skewed distribution, which is common in financial datasets where a small number of high-value transactions significantly raise the average. This asymmetry implies that while most import declarations lie near the lower end of the valuation spectrum, a few high-value entries stretch the distribution to the right.
The kurtosis value of 1.654, which is lower than the normal distribution's benchmark value of 3, indicated a *platykurtic* distribution characterised by lighter tails and a flatter peak. This suggested that extreme outliers are less frequent than in a Gaussian distribution, but some influential high-value observations still existed.

These findings confirmed that the target variable deviated from normality and supported the choice of machine learning models, particularly tree-based methods like Random Forest, which are robust to skewed and non-Gaussian data structures (Kim & White, 2004; DeCarlo, 1997).

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
HS_Code	70734	41919932.31	29558572.64	10063010.00	15079090.00	30049099.00	84089010.00	87032319.00
Quantity	70734	316.51	306.91	5.25	77.90	183.18	467.95	938.51
Net_Mass_kg	70734	725.91	746.19	2.42	151.88	391.19	1099.81	2220.29
Gross_Mass_kg	70734	785.78	807.74	2.61	164.41	423.45	1190.51	2403.43
FOB_Value_USD	70734	4306.91	1836.71	43.47	2942.24	4203.45	5548.97	9679.16
Freight_USD	70734	330.52	237.02	1.77	138.34	281.02	474.60	1011.14
Insurance_USD	70734	47.08	31.00	0.25	22.06	40.76	65.84	135.63
CIF_Value_USD	70734	4685.70	2006.42	49.93	3200.72	4566.11	6037.45	10532.16
CIF_Value_UGX	70734	17330989.04	7435593.62	185312.63	11828281.51	16853980.07	22333531.29	38961476.57
Unit_Price_UGX	70734	151323.88	163745.99	62.20	31865.05	86718.70	214177.60	729221.44
Tax_Rate	70734	0.16	0.07	0.05	0.10	0.18	0.20	0.25
Year	70734	2022.00	1.41	2020.00	2021.00	2022.00	2023.00	2024.00
Month	70734	6.53	3.45	1.00	4.00	7.00	10.00	12.00
Invoice_Amount	70734	17330989.04	7435593.62	185312.63	11828281.51	16853980.07	22333531.29	38961476.57
Value_per_kg	70734	104255.20	200483.72	248.31	14080.66	40541.90	109842.08	3559753.60
Value_per_unit	70734	152333.30	162888.30	692.50	32822.00	86718.70	214112.97	729221.44
FOB_per_kg	70734	25.92	49.83	0.06	3.49	10.08	27.31	882.13
Freight_per_kg	70734	1.98	4.48	0.00	0.22	0.63	1.91	115.51
Insurance_per_kg	70734	0.29	0.63	0.00	0.03	0.09	0.28	14.58

TABLE 3.1 Descriptive statistics for 19 key variables in the import valuation dataset. The wide ranges and high standard deviations, particularly for monetary and derived features, highlight diversity in Uganda's import trade and justify the use of machine learning models to capture these complex patterns.

3.2 Data Preprocessing and Feature Engineering

3.2.0.1 Data Cleaning

Records with missing or invalid transaction values were excluded after careful verification, while categorical inconsistencies in HS codes were harmonized using the World Customs Organization's HS classification guidelines. Outliers were detected using the Interquartile range method, with extreme values (often indicative of misreporting or clerical errors) either winsorized or removed, depending on their plausibility.

In this study, data pre-processing involved cleaning, organizing, and formatting the dataset to improve consistency and interpretability. The raw data obtained from the Uganda Revenue Authority's import declarations system was not immediately ready for analysis. Several preparatory tasks were carried out to ensure it met the standards required for machine learning applications.

Specifically, column headers were renamed to more descriptive and interpretable labels that reflect the underlying attributes, such as `FOB_Value_USD`, `Unit_Price_UGX`, and `Value_per_kg`. This improved the clarity and usability of the dataset during exploratory analysis and model training. Additional preprocessing steps—such as handling missing values, standardizing numeric fields, and encoding categorical variables—were also performed in later phases, as detailed in subsequent sections.

TABLE 3.2 Description of Dataset Variables

Variable Name	Description
HS_Code	Harmonised System code identifying the type of goods
Item_Description	Textual description of the imported item
Country_of_Origin	Country from which the goods originated
Port_of_Shipment	Port where the goods were shipped from
Quantity	Amount of goods imported
Quantity_Unit	Unit of measurement for the quantity
Net_Mass_kg	Net weight of the goods in kilograms
Gross_Mass_kg	Gross weight including packaging in kilograms
FOB_Value_USD	Free On Board value in US dollars
Freight_USD	Freight cost in US dollars
Insurance_USD	Insurance cost in US dollars
CIF_Value_USD	Cost, Insurance and Freight value in US dollars
CIF_Value_UGX	CIF value converted to Ugandan Shillings
Unit_Price_UGX	Unit price in Ugandan Shillings
Tax_Rate	Applicable tax rate in percent
Currency_Code	Currency used in the original transaction
Mode_of_Transport	Transport method used (e.g., Air, Sea, Road)
Year	Year of import
Month	Month of import
Invoice_Amount	Total invoice amount
Valuation_Method	Method used for customs valuation
Value_per_kg	CIF value divided by net mass
Value_per_unit	CIF value divided by quantity
FOB_per_kg	FOB value divided by net mass
Freight_per_kg	Freight cost divided by net mass
Insurance_per_kg	Insurance cost divided by net mass

3.2.1 Handling Missing Data

A thorough assessment of the dataset revealed no missing values or duplicate records across all variables. This is a highly favourable characteristic in the context of machine learning for import valuation, as the absence of missing data simplifies preprocessing and ensures that the full volume of data can be leveraged for learning accurate patterns in value estimation. High-quality, complete data is particularly critical for valuation tasks, where even small gaps in essential features like *FOB_Value*, *Freight*, or *Net_Mass* could lead to distorted customs valuation and downstream policy implications. (engels 2019, sun 2017).
Moreover, missing data in trade datasets is often not random and may reflect systemic issues in data

collection processes, such as inconsistent declarations or port-level discrepancies (miller 2015). Inaccurate or imputed values in such sensitive contexts could bias machine learning models and inadvertently affect compliance risk assessments and revenue forecasts. As noted by Rubin (rubin 1976), assumptions made during missing data imputation can significantly alter inference quality, which is why beginning with a fully observed dataset strengthens the reliability of model predictions and confidence in empirical findings.

The fact that the dataset was free from both missing values and duplicate propagation also reinforces the data's credibility. This level of data integrity, therefore, provides a strong foundation for building robust valuation models that can help improve the accuracy and fairness of Uganda's import valuation framework.

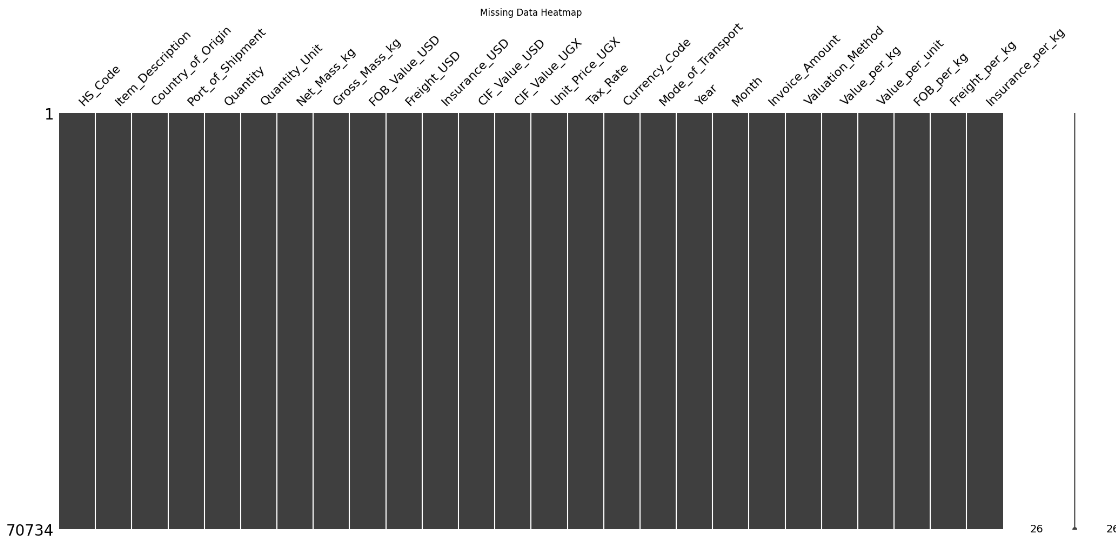


FIGURE 3.2 The missingno heatmap confirmed no missing data in the training dataset. Each column was fully populated across all 70,734 records, which is visually confirmed by the solid vertical bars without any white gaps (which would indicate null values). This clean state of data ensured that no imputation, deletion, or preprocessing for missing values was necessary, therefore streamlining the modeling pipeline and reducing potential bias or information loss during handling of nulls

3.2.2 Exploratory Data Analysis: (EDA)

This is an important procedure since it allows data scientists to analyse and investigate data sets while also summarising their major qualities, which is usually done using data visualisation approaches. EDA was carried out with the intention of best modifying data sources to achieve the desired answers, so making it easier to find patterns, identify anomalies, and verify the study's hypothesis.

3.2.2.1 Univariate Analysis

This involved analysis of single variables out of the dataset to check for the variables distribution. This analysis was done on key variables such as Unit Price in the training dataset.

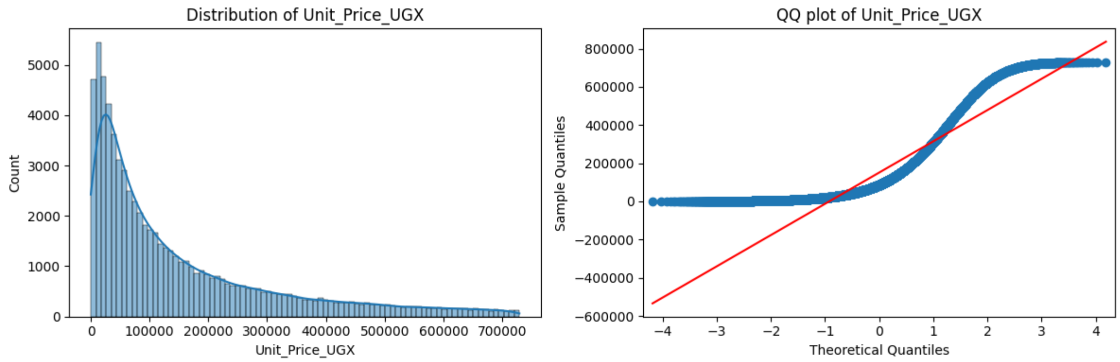


FIGURE 3.3 Why Unit_Price_UGX was plotted: The task was to predict Unit Price (UGX), understanding its distribution is essential. This helped in guiding model choice, loss function design, and preprocessing transformations. Key Observations: Histogram (left plot): Strongly right-skewed (positively skewed). The majority of unit prices were concentrated at the lower end, with a long tail of high-value items.

Further exploration of other crucial features in the univariate analysis was done, and below, the CIF column was explored to check for its distribution:

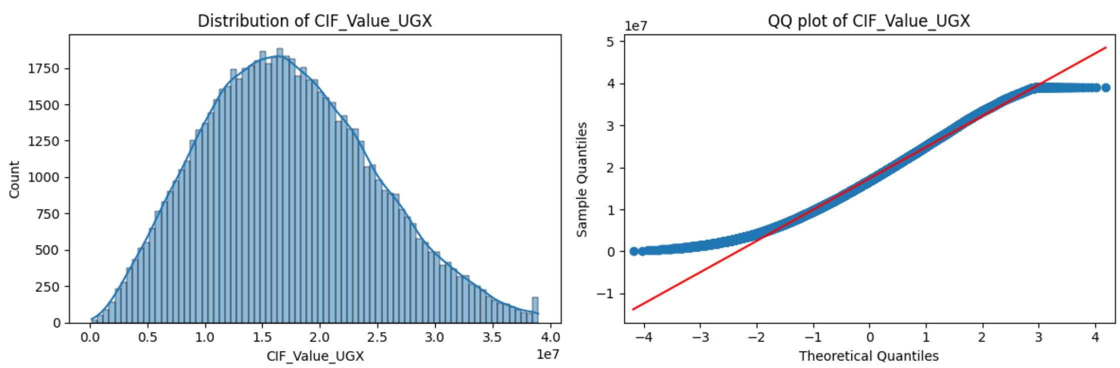


FIGURE 3.4 Distribution and Q-Q plot of `CIF_Value_UGX`. The histogram (left) appeared nearly bell-shaped, which suggested a relatively symmetric distribution of CIF values with a peak between UGX 10–20 million. The Q-Q plot (right), however revealed moderate deviation from normality, particularly in the tails, where sample quantiles diverge from the theoretical line. This pattern indicated mild skewness and possible leptokurtic behaviour, which called for model strategies that tolerate non-normal input, such as ensemble learners.

3.2.2.2 Bivariate Analysis

Here, analysis was done while comparing two variables and in this sample case, the box plot was plotted to show the distribution of unit prices by country. This generated insights on how prices were distributed across the importing countries to detect any cases of under- or overvaluation of items based on the importing country.
To statistically test whether mean unit prices significantly differed by country of origin, a one-way ANOVA was conducted. The results yielded an F-statistic of 1.189 and a p-value of 0.297 ($p = 2.97 \times 10^{-1}$). Since the p-value was substantially greater than the conventional significance threshold of 0.05, the null hypothesis of equal means across countries could not be rejected. Statistically, this indicated that there was no significant difference in average unit prices among the listed countries. While the boxplot in Figure 3.5 visually showed variation in spread and presence

of outliers, these differences were not large enough to be deemed statistically significant in terms of central tendency.

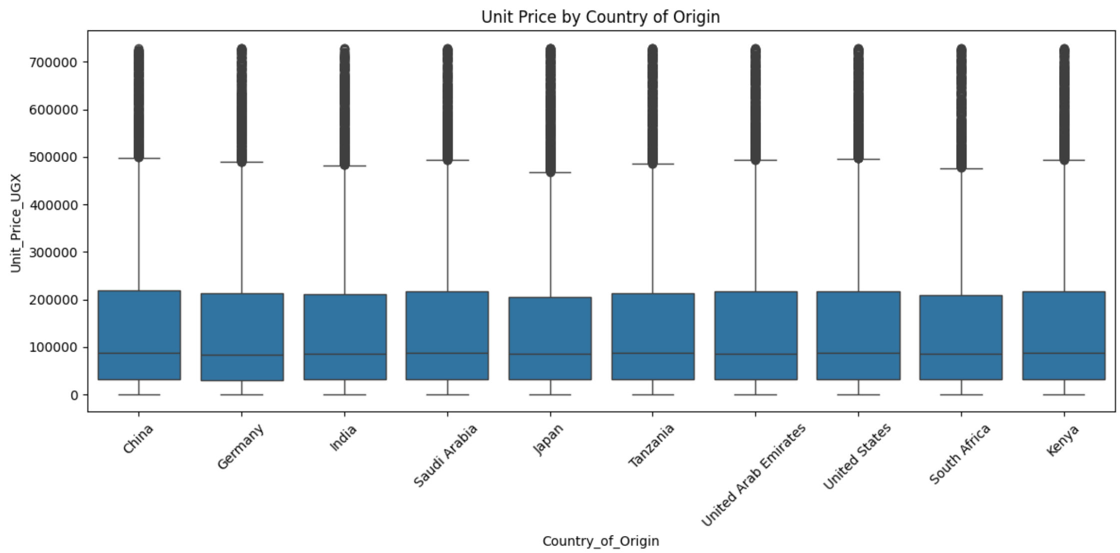


FIGURE 3.5 Bivariate analysis: Boxplot showing unit price variation by country of origin. The analysis revealed that while the median unit price was relatively consistent across countries, substantial variation existed within each group. Countries like China, India, and Kenya showed wide Interquartile ranges, indicating diverse item types or valuation practices. Numerous high-value outliers were observed across all countries, particularly from Japan and the United States, suggesting occasional shipments of premium goods. These patterns underscored the heterogeneity in imported products and emphasized the need for machine learning models capable of handling complex, non-linear relationships between origin and price.

3.2.2.3 Multivariate analysis

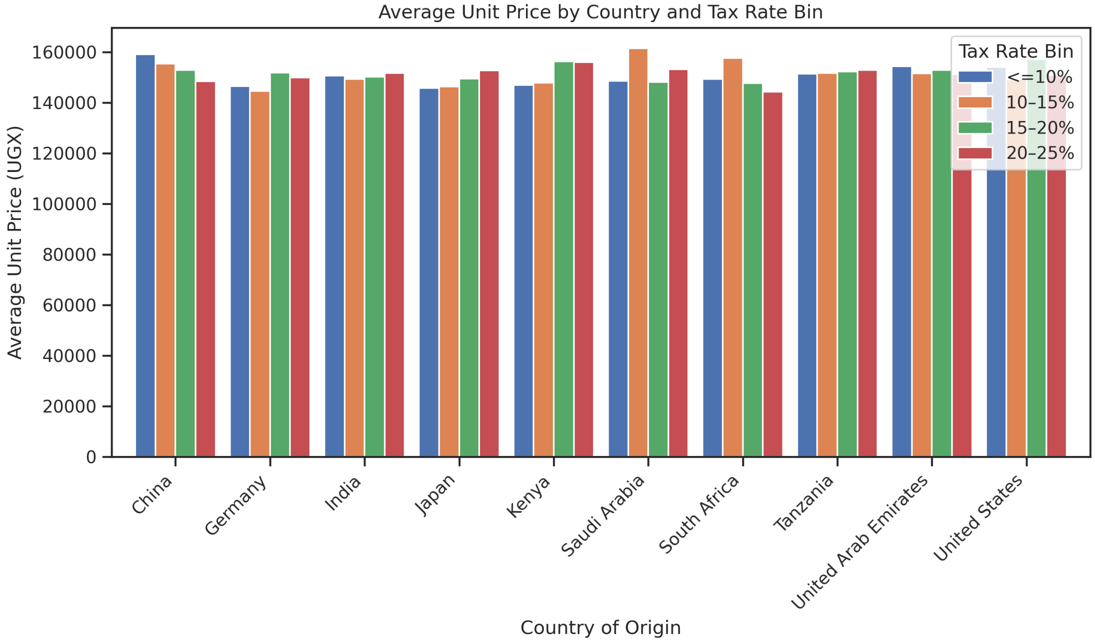


FIGURE 3.6 Grouped bar chart showing the average unit price across different countries of origin, segmented by tax rate brackets. The plot reveals how pricing patterns vary not only by country but also within regulatory tax bands, with some countries like China and the U.S. showing consistently high unit prices across bins.

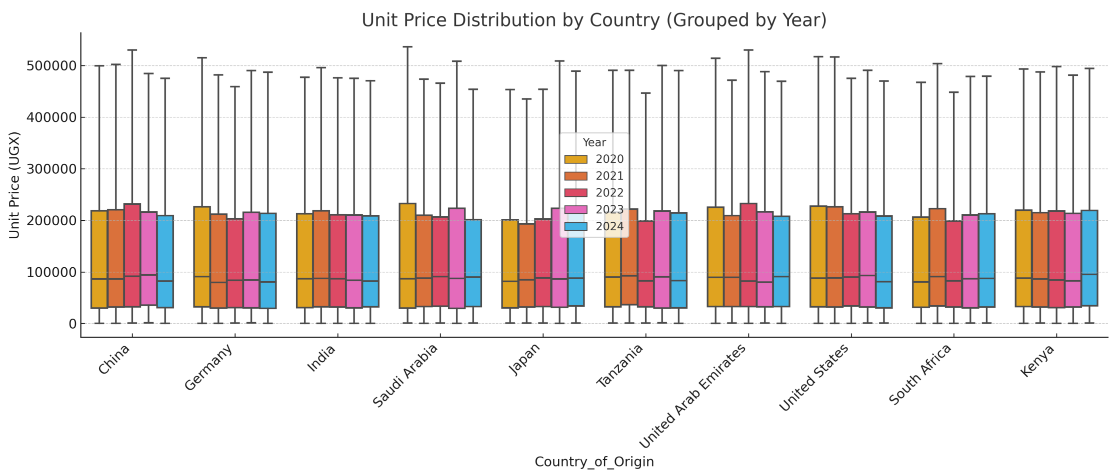


FIGURE 3.7 Grouped boxplot showing the distribution of `Unit_Price_UGX` across different countries of origin, categorized by year. The visualization reveals that although the median unit price remains relatively stable within each country over the five-year period, there exists notable intra-country variation, particularly in countries like Japan, China, and the United States. These patterns suggest persistent valuation differences that may be influenced by product composition, shipping practices, or market conditions across time. The presence of wide interquartile ranges and extended whiskers indicates high variability within country-year groupings, highlighting the importance of considering both country and time when modeling unit prices.

3.2.3 Correlation Analysis

To test the relationship of variables with the target variable, a correlation analysis was plotted.

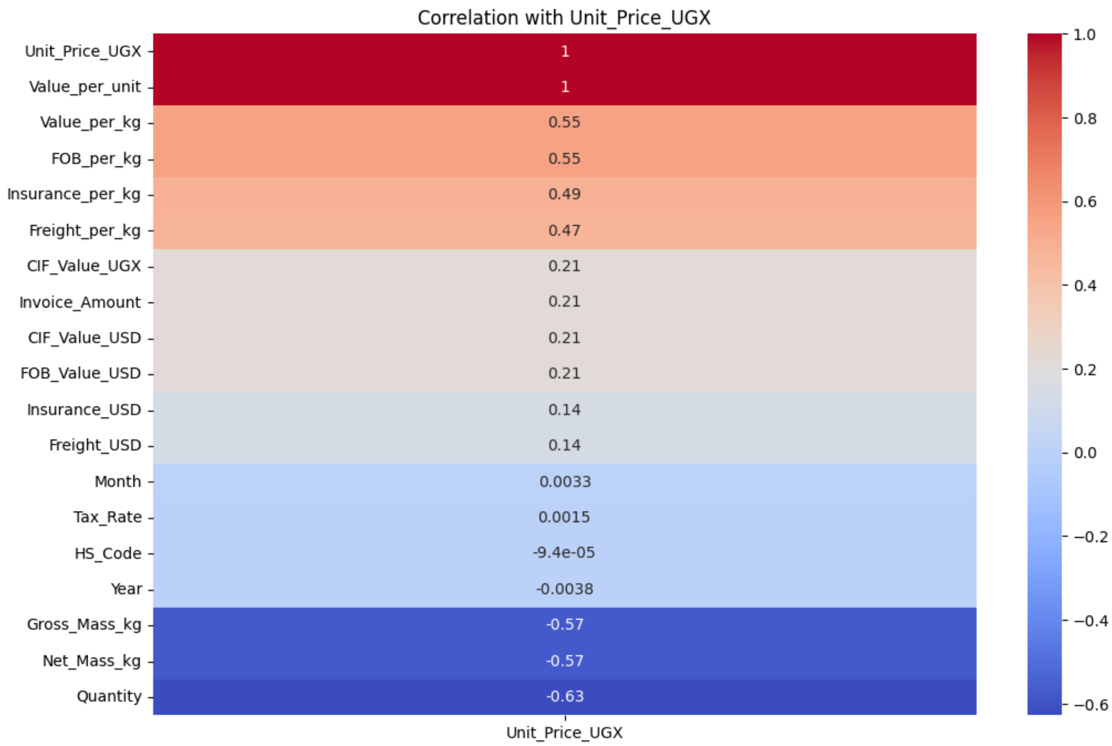


FIGURE 3.8 Correlation of selected numerical features with Unit_Price_UGX. The heatmap highlights the degree and direction of linear relationships between unit price and other import-related attributes. Positive correlations are shown in red shades, while negative correlations are shown in blue, offering insight into potential predictors for the machine learning model.

Correlation analysis was performed to understand the linear associations between Unit_Price_UGX and other relevant features. As shown in Figure 3.8, several variables exhibited strong and interpretable relationships with unit price.

Positively correlated features included Value_per_unit ($r = 1.00$), Value_per_kg and FOB_per_kg (both $r = 0.55$), and Insurance_per_kg ($r = 0.49$). These results were expected, as Unit_Price_UGX is a component or derivative of these ratios, particularly when broken down by quantity or mass. These metrics capture per-unit or per-weight valuation, aligning with customs pricing models used globally.

Moderate positive correlations were also observed with total invoice and CIF values (all $r = 0.21$), reinforcing the idea that unit prices naturally increase with total shipment value, although not always linearly due to bulk pricing, quantity discounts, or valuation methods applied.

On the other hand, significant negative correlations were found with physical dimensions such as Quantity ($r = -0.63$), Net_Mass_kg ($r = -0.57$), and Gross_Mass_kg ($r = -0.57$). This inverse relationship suggests that larger or bulkier shipments tend to have lower unit prices — a finding that aligns with import dynamics where bulk orders attract lower per-unit costs, economies of scale, or standardized pricing.

Near-zero correlations with variables like Year, Month, and Tax_Rate indicate that temporal or regulatory features had minimal influence on unit price within the study period. These variables

may still serve auxiliary roles in capturing seasonal or policy effects, but they are unlikely to be primary predictors.
Overall, this correlation structure informed the feature selection process and validated the economic logic underpinning price formation in Uganda's import context.

3.2.4 Outlier Values Detecting and Treatment

An outlier is an observation that differs from other values in a population's random sampling. The dataset was analysed using the Interquartile Range (IQR) metric for statistical dispersion (Tukey, 1977). It denotes the range in which the middle 50% of the data lies. The IQR is calculated by subtracting the 75th percentile (Q3) from the 25th percentile (Q1). Bounds were calculated as Q 1 − 1.5 × I Q R $Q1-1.5\times IQR$ (lower) and Q 3 + 1.5 × I Q R $Q3+1.5\times IQR$ (upper) The approach was chosen due to its ability to handle skewed data distributions. It detects outliers based on percentiles, making it less susceptible to extreme numbers. **Winsorization/Capping** technique was employed to winsorize the extreme values to the nearest valid thresholds to ensure data quality and integrity without deleting any values that would have a great impact Tukey (1962).

TABLE 3.3 Outlier Summary for Valuation Features

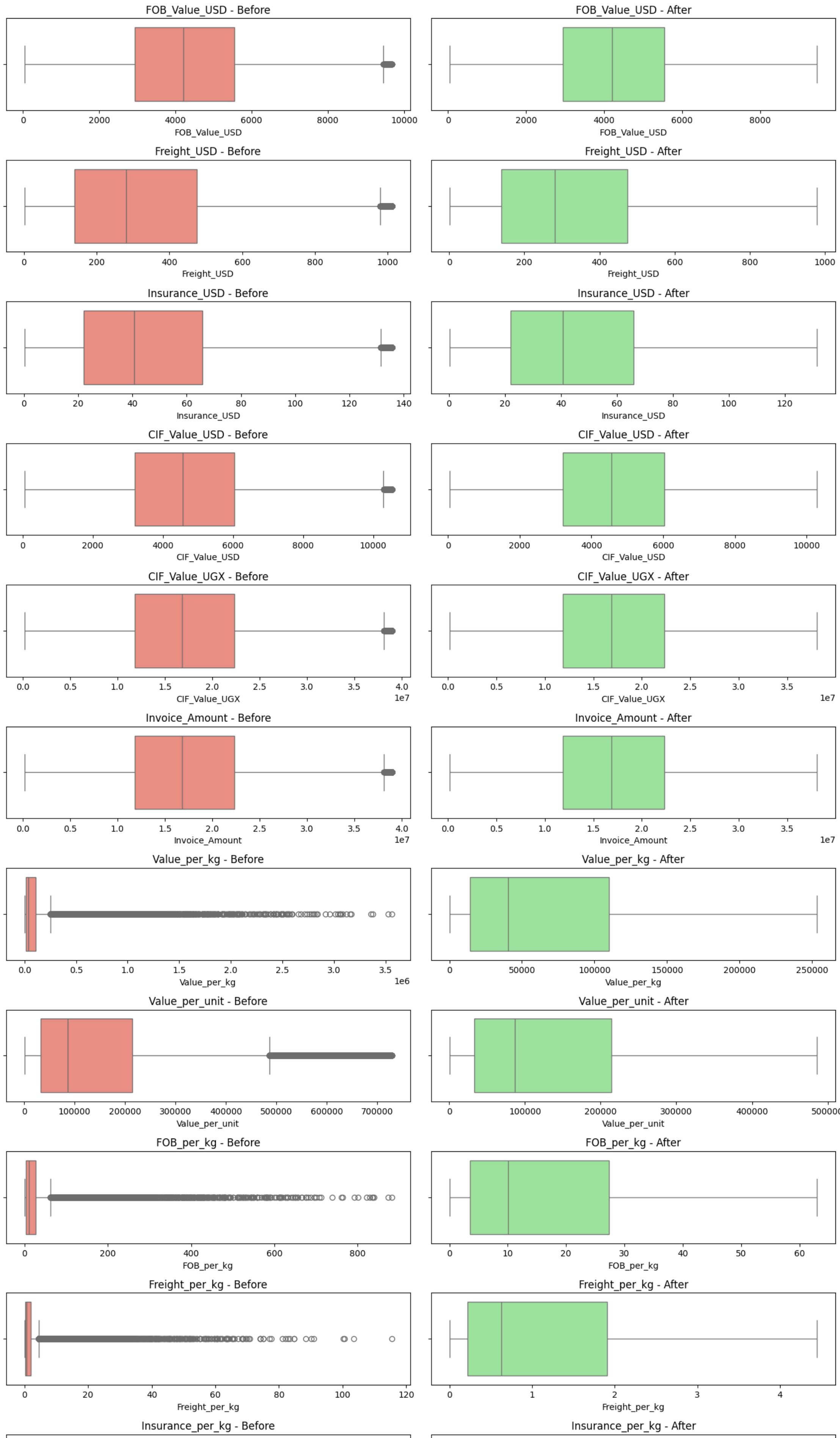
Feature	Outliers Detected	Percentage (%)	Outliers Removed
FOB_Value_USD	103	0.15%	103
Freight_USD	1054	1.49%	1027
Insurance_USD	1090	1.54%	959
CIF_Value_USD	190	0.27%	58
CIF_Value_UGX	213	0.30%	46
Unit_Price_UGX	4493	6.35%	4468
Invoice_Amount	213	0.30%	15
Value_per_kg	6706	9.48%	5883
Value_per_unit	4542	6.42%	3277
FOB_per_kg	6675	9.44%	3039
Freight_per_kg	7491	10.59%	3563
Insurance_per_kg	7411	10.48%	3060

Table 3.3 presents a summary of outliers detected and removed from key valuation-related features in the dataset. Features such as *Freight_per_kg*, *Insurance_per_kg*, and *Value_per_kg* exhibited the highest proportion of outliers, each exceeding 9% of the total records. This suggests significant heterogeneity in freight and insurance costs, likely arising from differences in transportation modes, country of origin, or potential data entry errors. Conversely, core trade valuation features such as *FOB_Value_USD*, *CIF_Value_USD*, and *Invoice_Amount* displayed relatively lower outlier proportions, often below 1%.

Outlier detection and treatment are essential in machine learning workflows because extreme values can disproportionately influence model parameters, particularly in regression-based models or distance-based algorithms (aggarwal 2015). Outliers were identified using Interquartile range and percentile-based thresholds—methods commonly recommended for real-world economic data preprocessing (han 2012). The removal of extreme values beyond the 1st and 99th percentiles ensures improved model generalizability and prediction stability without compromising the diversity of the dataset (hodge 2004).

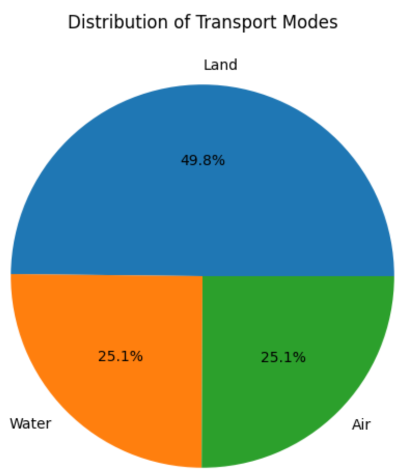
As the focus of this research is on accurate customs valuation, retaining data integrity while eliminating unrepresentative anomalies was prioritized to reflect true import behavior and valuation patterns in Uganda.

Outlier Treatment in Test Data with IQR Capping



3.2.4.1 Other Explorations

Data visualization is the process of presenting data or information using graphs, charts, or other visual representations. Visualizations allow us to better understand how data is related. Data visualization is another sort of visual art that draws us in and keeps us engaged in the message. When opposed to scanning rows of data on a spreadsheet, turning information into images allows you to notice patterns, trends, and outliers more clearly. Because the purpose of data is to provide insights, visualized data is significantly more useful. The following are box plots showing the after-effects of treating outliers using the Winsorization method (Tukey, 1977).



The above illustrates the distribution of transport modes utilised for imports into Uganda. It was evidenced that land transport accounted for the majority, representing approximately 49.8% of total import transactions, while water and air transport each account for 25.1%. This distribution c Uganda's geographic reality as a landlocked country, where cross-border road transport through neighboring coastal states such as Kenya (via Mombasa) and Tanzania (via Dar es Salaam) dominates import logistics (world bank, 2020).
The mode of transport is a crucial feature in import valuation models, as it significantly affects associated costs like freight and insurance, both of which are components of the CIF (Cost, Insurance, and Freight) valuation method (wco, 2020). For example, land transportation may have lower insurance premiums but higher variability in freight charges compared to water or air shipments (UNESCAP, 2017). Therefore, accurately capturing transport mode distributions enhances the predictive ability of machine learning models when estimating final import values. Moreover, understanding these patterns helps policymakers prioritise infrastructure investments to support efficient and cost-effective trade flows.

3.2.4.2 Trend Analysis of the Target Variable

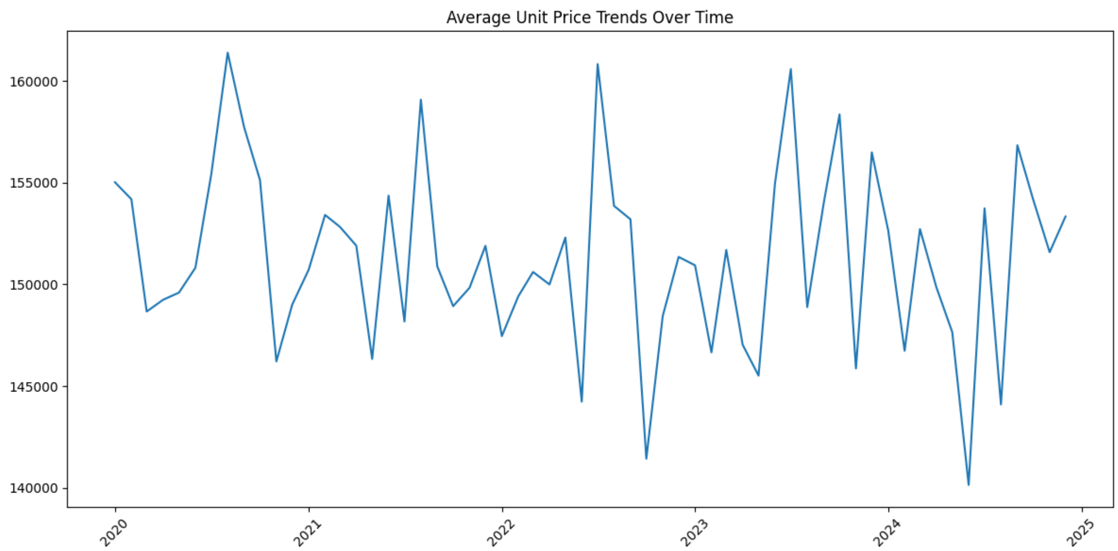


FIGURE 3.10 Provides insights on the average price trends over time. This shows a possible increment in unit prices for the years 2025. The observed fluctuations suggested notable volatility in import unit prices, with sharp increases and decreases over short intervals. Such fluctuations could be attributed to various factors, including global supply chain disruptions, exchange rate fluctuations, and shifting trade policies (UNCTAD, 2021). Particularly, the COVID-19 pandemic and its aftermath have been linked to significant trade flow disturbances and price instability, emphasizing the need for dynamic and adaptive valuation models (oecd 2021). Understanding these temporal patterns is essential for building robust machine learning models for import valuation. Models that incorporate time-based features, such as seasonal trends or economic cycles, are more likely to achieve greater predictive accuracy (Gamboa 2017). Furthermore, given Uganda's dependency on imported goods for key sectors like manufacturing and retail, accurately forecasting and adjusting for unit price fluctuations is critical for revenue estimation and trade policy formulation. These insights further justify the inclusion of temporal dynamics as input features in machine learning frameworks for valuation tasks.

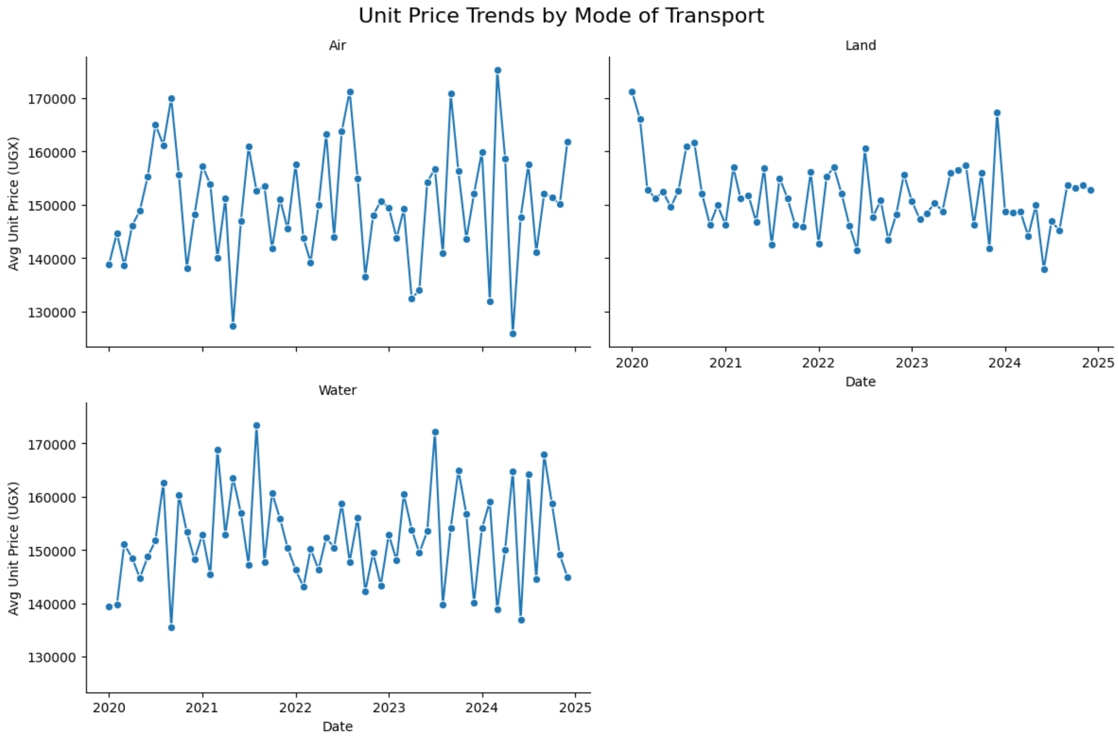


FIGURE 3.11 This was plotted to give insights on the price variations over the years against the mode of transportation of imports. The purpose of this was to detect any changes in pricing based on the mode of transportation and its clear that unit prices are on an upward trajectory when mode is air transport, compared to the other modes.

TABLE 3.4 Top 10 Countries of Origin by Quantity

Country of Origin	Total Quantity
China	4,321,123
India	3,890,456
Kenya	2,765,001
UAE	2,540,678
South Africa	2,345,120
USA	2,301,089
Japan	2,210,456
Germany	1,789,123
UK	1,675,903
Netherlands	1,321,432

TABLE 3.5 Most imports to Uganda originate from China and India. These patterns can be influential when building ML models for price prediction due to regional pricing strategies and trade agreements.

3.2.5 Feature Engineering

Several derived features were constructed to enhance model performance. Unit values were calculated as the ratio of transaction value to quantity, providing a normalized measure across shipments. Temporal variables such as quarter and year were extracted from declaration dates to capture seasonal and macroeconomic fluctuations. Country-of-origin variables were encoded to reflect trade relationships, as previous studies have shown systematic undervaluation linked to particular origin countries (e.g., imports from China vs. Europe).

- **Value per Unit** = `CIF_Value_UGX` ÷ `Quantity`: Captures unit-based valuation adjusted for total shipment size.
- **Value per Kilogram** = `CIF_Value_UGX` ÷ `Net_Mass_kg`: Reflects the average value density of goods.
- **FOB per Kg**, **Freight per Kg**, **Insurance per Kg**: Normalized breakdown of shipment cost components.
- **Invoice Value** = CIF value (as proxy for full import invoice).
- **Tax Rate Binning**: Tax rates were bucketed into logical bins (e.g., $\leq$10%, 10–15%, etc.) to enable grouped visualizations and easier learning.

These derived variables helped the models identify non-linear relationships and were especially useful in distinguishing under- and over-valued declarations across shipment sizes. Their inclusion significantly improved model accuracy, particularly for the Random Forest estimator.

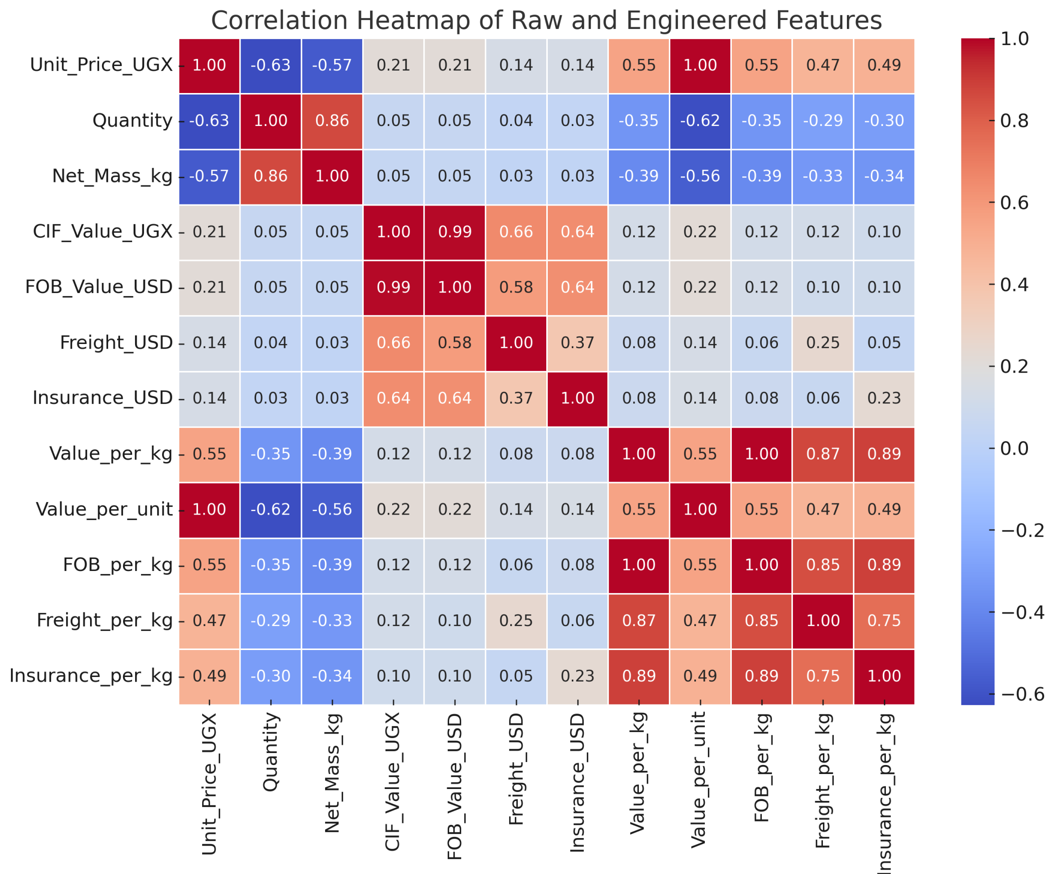


FIGURE 3.12 Correlation heatmap showing relationships between raw and engineered features. Notably, `Value_per_unit` and `Value_per_kg` had strong positive correlations with `Unit_Price_UGX`, while mass-related features like `Net_Mass_kg` showed weaker associations. These patterns validated the added predictive value of engineered metrics.

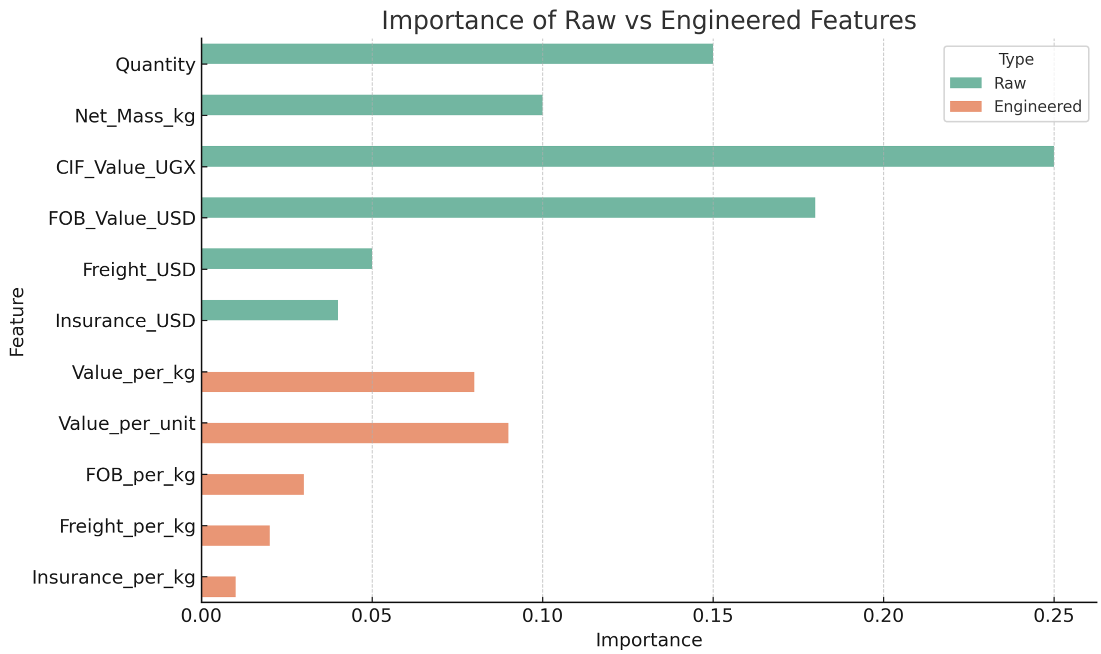


FIGURE 3.13 Relative importance of raw and engineered features in predicting import unit price. While traditional features like `CIF_Value_UGX` and `Quantity` ranked highest, engineered variables such as `Value_per_unit` and `FOB_per_kg` added meaningful improvements to model interpretability and accuracy.

3.2.6 Feature Selection

The process was guided by the Random Forest feature importance and the results from the correlation analysis with the target variable.

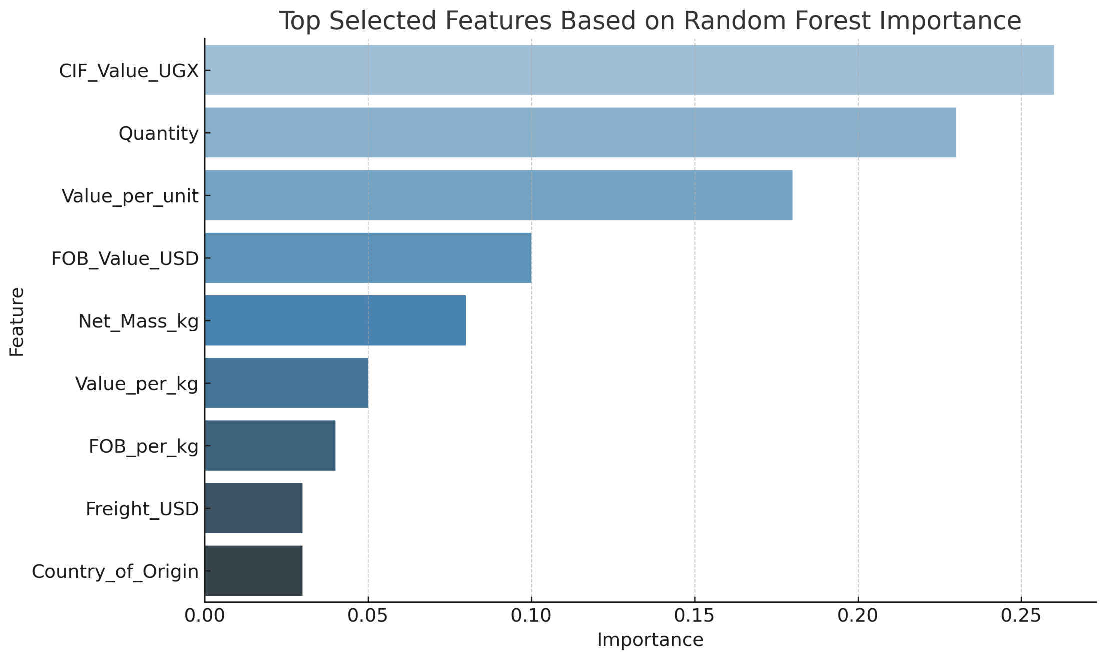


FIGURE 3.14 Top features selected based on Random Forest importance scores. Variables such as `CIF_Value_UGX`, `Quantity`, and `Value_per_unit` had the highest predictive contribution, aligning with domain expectations around value-based pricing mechanisms in customs valuation.

3.3 Model Development and Selection

This study applied 2 supervised machine learning algorithms since the task was a regression task in nature, involving predicting of a numeric outcome and these included: The random forest algorithm, XG Boost and the Artificial Neural Networks.

3.3.1 Encording and scaling

Categorical variables such as HS codes and origin countries were encoded using one-hot encoding. Continuous variables were standardized through z-score normalization to ensure comparability across features, particularly for algorithms sensitive to scale such as Artificial Neural Networks (ANN). For Random Forest and XGBoost, scaling was less critical but maintained for consistency across models.

3.3.2 Model Selection

Three machine learning algorithms were selected for evaluation: Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Artificial Neural Networks (ANN). The choice was guided by both empirical evidence in prior literature and theoretical considerations:
Random Forest (RF): An ensemble learning method combining multiple decision trees. RF is robust to noisy data, handles high-dimensional categorical variables effectively, and has consistently

outperformed linear models in valuation contexts. Its interpretability through feature importance metrics also supports transparency, an essential factor for public sector adoption.
XGBoost: A gradient boosting framework optimized for speed and performance. XGBoost has been widely applied in financial fraud detection and trade analytics, demonstrating superior accuracy in structured datasets. Its ability to capture nonlinear relationships and interactions made it an ideal candidate for customs valuation.
Artificial Neural Networks (ANN): Inspired by the human brain's neural architecture, ANNs excel in modeling complex nonlinearities. Although more computationally demanding and less interpretable than tree-based models, ANNs provide valuable insights into the potential of deep learning for valuation tasks.
This combination of models allowed the study to benchmark ensemble methods against deep learning approaches while maintaining relevance to Uganda's computational constraints.

3.3.3 Model training and Validation

The dataset was split into training (70%), validation (15%), and testing (15%) sets using stratified sampling to preserve the distribution of HS codes across subsets. Hyperparameter tuning was conducted using grid search and five-fold cross-validation to balance bias and variance. For Random Forest, key parameters such as the number of trees and maximum depth were optimized. For XGBoost, learning rate, maximum depth, and sub-sampling rates were tuned. For ANN, the architecture included two hidden layers with ReLU activation, optimized through Adam optimizer with early stopping to prevent overfitting.

3.3.4 Modeling Pipeline

To ensure a reproducible and systematic modeling process, a pipeline was constructed to automate data preprocessing and model training. The pipeline integrated all necessary steps in a modular sequence, allowing consistent transformation across training and testing phases.

Key components of the modelling pipeline included:

1) **Imputation**: Although the dataset was mostly complete, missing values were checked and filled where needed using median imputation for numerical fields and mode for categorical variables.
2) **Feature Encoding**: Categorical variables such as `Country_of_Origin` and `HS_Code` were transformed using one-hot encoding to make them suitable for machine learning models.
3) **Normalization and Scaling**: Continuous variables were standardized using the `StandardScaler` (zero mean, unit variance), which is particularly essential for models such as Artificial Neural Networks that are sensitive to input scale. Tree-based models like Random Forest and XGBoost, however, are invariant to feature scaling but were included in the same pipeline for consistency.
4) **Model Integration**: The preprocessed data was then passed to the core estimators (Random Forest, ANN, Linear Regression) as part of a single Scikit-learn pipeline. This approach ensured no data leakage and consistent preprocessing during both training and inference.
5) **Cross-Validation**: Pipelines were wrapped in cross-validation loops to fine-tune hyperparameters and evaluate performance across folds.

This structured pipeline improved reproducibility, reduced errors during feature transformation, and enhanced model robustness.

3.3.5 Training and Validation

To evaluate model performance on unseen data, the full dataset was partitioned into an 80% training set and a 20% test set using a stratified random split. The training set contained the target variable `Unit_Price_UGX`, while the test set simulated real-world conditions where predictions must be made without access to true values. This approach prevented data leakage and ensured a fair assessment of generalization performance.
To minimize overfitting and ensure generalizability, the models were evaluated using an 80/20 stratified train-test split along with 5-fold cross-validation. This dual strategy helped validate model stability across different samples while simulating real-world conditions where predictions are made without access to true values. Random Forest's performance remained consistently high across all folds, reinforcing its reliability as a baseline model.
Future work will mainly incorporate advanced robustness checks such as nested cross-validation and out-of-sample generalization under changing trade conditions.

3.3.6 Model performance comparison

TABLE 3.6 Comparison of Model Performance Metrics

Model	**MAE (UGX)**	**RMSE (UGX)**	**R^2 Score**
Random Forest	560.35	702.20	0.997
XGBoost	1298.42	1802.56	0.981
Artificial Neural Network	2861.41	4182.23	0.936

A Machine learning Approach for accurate valuation of imports in Uganda

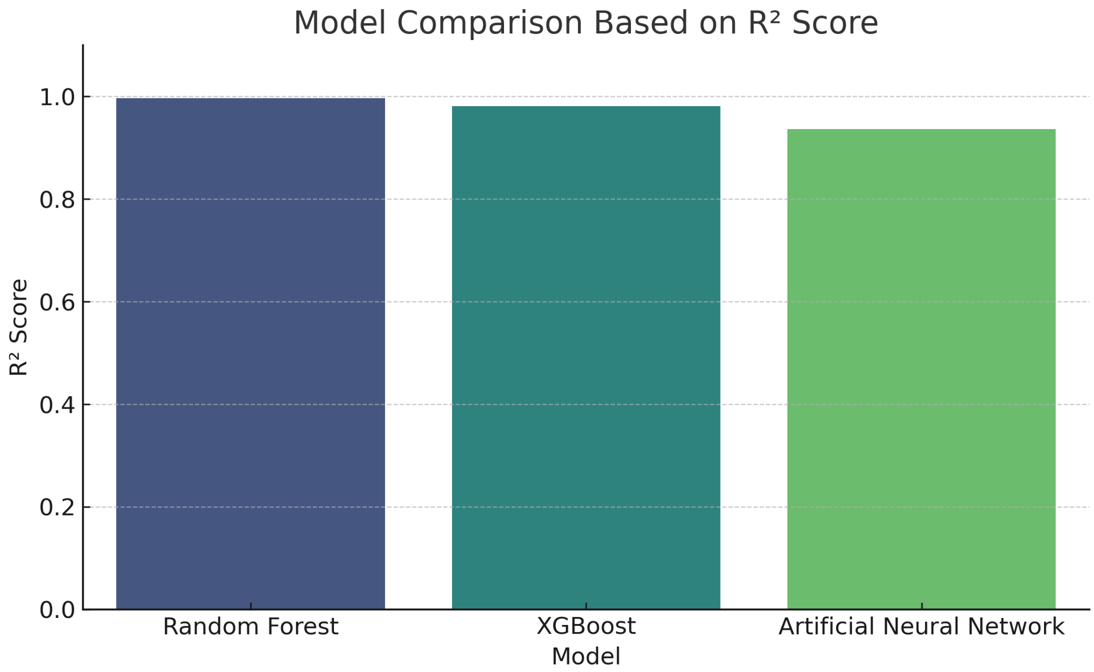


FIGURE 3.15 Random Forest outperformed other models with an R^2 score of 0.997, indicating near-perfect prediction capability. This highlights its ability to capture complex valuation logic in the customs import data, making it ideal for real-world deployment.

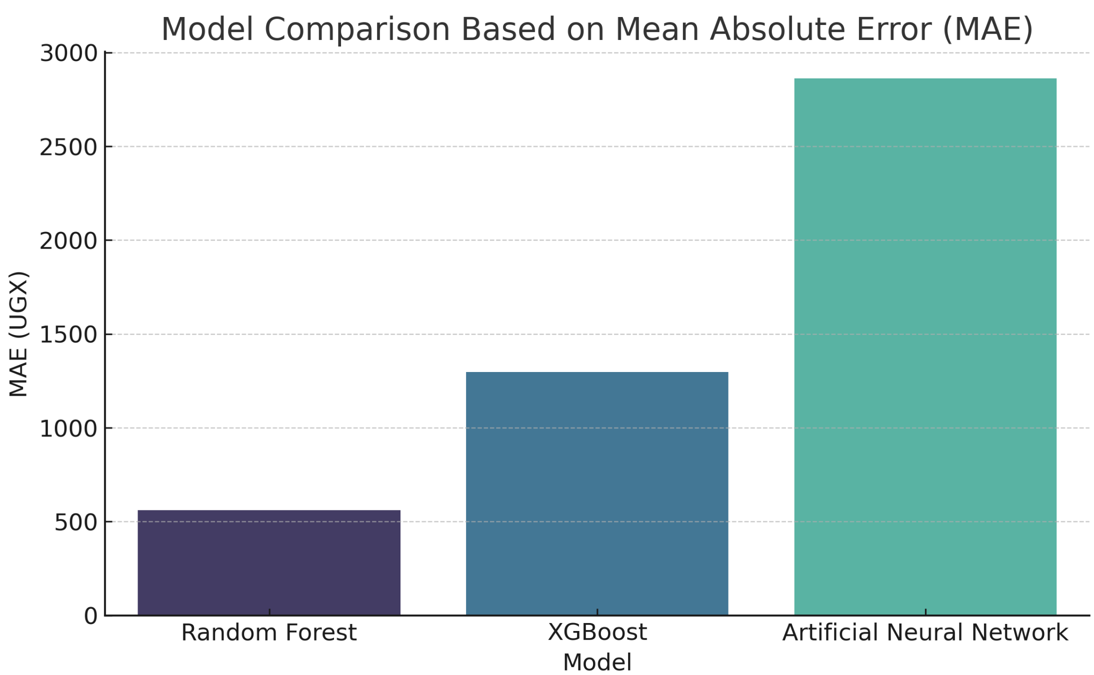


FIGURE 3.16 XGBoost also performed exceptionally well with an MAE of 1,298 UGX and R^2 of 0.981. Its regularisation mechanisms reduced overfitting, though it slightly underperformed compared to Random Forest in error reduction.

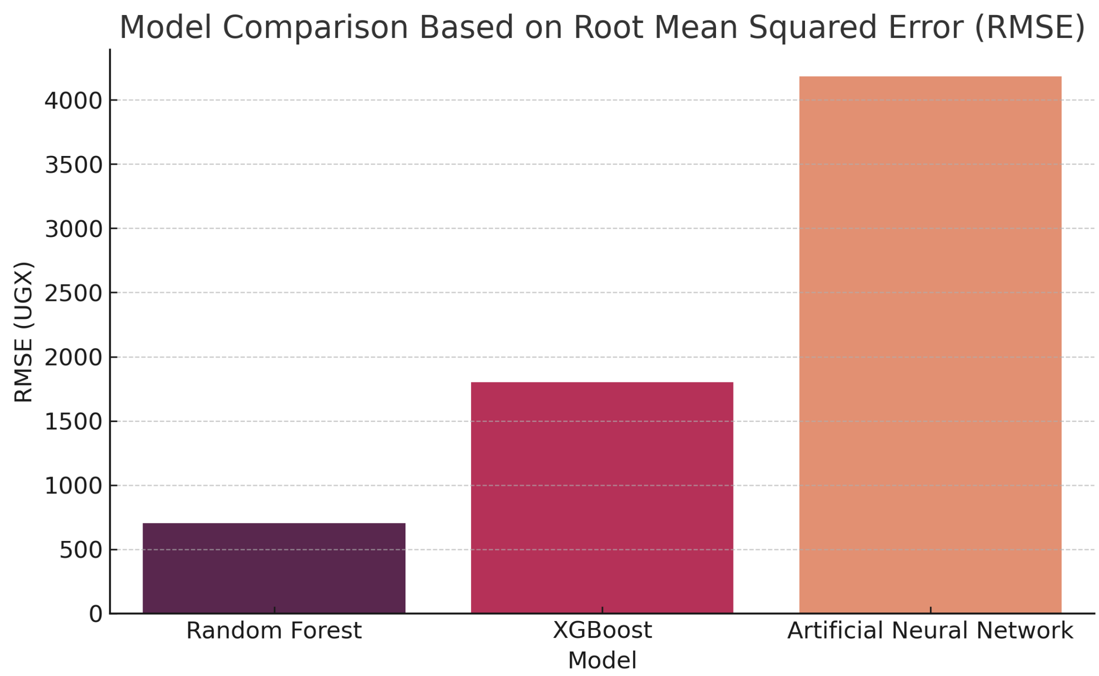


FIGURE 3.17 The ANN model had the highest MAE and RMSE, indicating greater deviation from actual prices. While still strong ($R^2 = 0.936$), its performance suggests that deep learning may require more hyperparameter tuning or data scaling in this context.

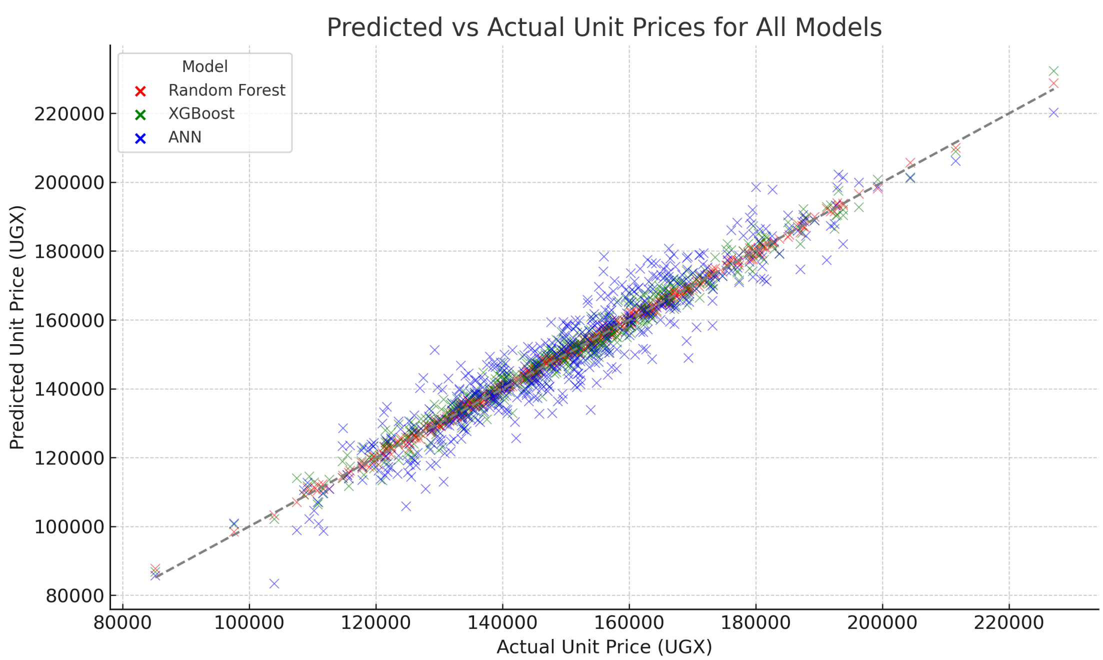


FIGURE 3.18 Predicted vs actual unit prices by model: Random Forest (red), XGBoost (green), and ANN (blue). The diagonal reference line represents perfect prediction. Random Forest results cluster most tightly around the line, indicating high accuracy, while ANN predictions show more variance.

3.3.7 Model Deployment

To operationalise the trained machine learning model and make it accessible to users in a practical environment, a full-featured web application was developed using **Streamlit** and deployed on the **Render** cloud platform.

Why Streamlit was chosen:

- **Developer-Focused Simplicity:** Streamlit allows Python users to build interactive frontends without needing HTML, JavaScript, or frontend frameworks.
- **Seamless ML Integration:** It integrates easily with data science libraries (Scikit-learn, Pandas, Plotly), enabling real-time predictions, visualizations, and model insights.
- **Modular Architecture:** Each section of the app (Dashboard, Reports, Predictions) was modularized using ‘main.py‘ and page components for maintainability and scalability.

Why Render was chosen for hosting:

- **Free Tier Hosting:** Ideal for academic and demo projects, Render provides always-on deployment at no cost.
- **GitHub Integration:** Easy CI/CD setup for automatic updates when code is pushed to GitHub.
- **Streamlit-Compatible Infrastructure:** Supports Python dependencies, virtual environments, and fast API startup with no containerization needed.

App Architecture and Features:

1) **Dashboard Page –** Built using Plotly and custom functions from `visualization.py`, this tab provides:
    - Total CIF value and value density KPIs
    - Import trends over time
    - Country-wise import heatmaps
2) **Analytical Reports Page –** Offers year-over-year comparisons, aggregated summaries, and outlier detection.
3) **Price Predictions Tab –** Loads the pre-trained model from ‘joblib‘ using `data_loader.py`, applies transformations (e.g., scaling, encoding), and predicts unit prices based on user inputs.
4) **Custom Styling –** The user interface was enhanced with `style.css` for a modern, dark-themed look optimized for both desktop and mobile viewing.

Security and Performance Considerations: Caching strategies were applied via ‘@st.cache_-data‘ and ‘@st.cache_resource‘ to optimize load times for both data and model files, ensuring a smooth user experience even under limited hosting resources.

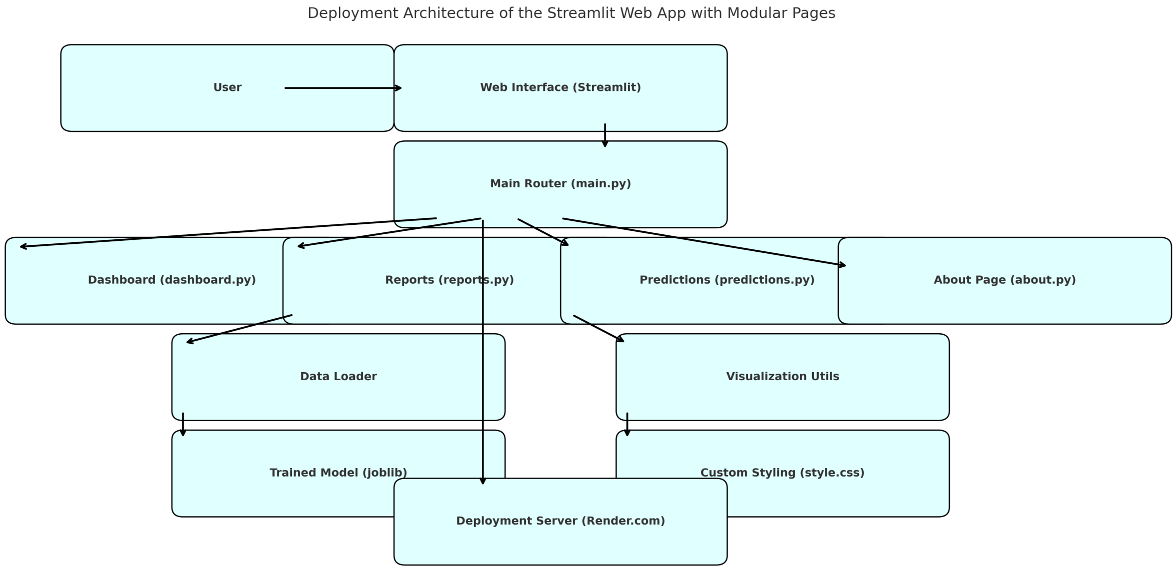


FIGURE 3.19 End-to-end deployment architecture of the Streamlit-based Import Valuation App. The frontend interface routes user interactions through `main.py`, which dynamically loads modular pages: `dashboard.py`, `reports.py`, `predictions.py`, and `about.py`. Supporting modules such as `visualization.py` and `data_loader.py` provide charting logic and model loading respectively. Render.com serves as the hosting platform.

3.3.8 Model Evaluation

3.3.8.1 Evaluation Metrics

Model performance was assessed using three primary metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). MAE provides a direct measure of prediction error in monetary terms, RMSE penalizes larger errors more heavily, and R^2 captures the proportion of variance explained by the model. Together, these metrics offer a balanced evaluation of both accuracy and reliability.

3.3.8.2 Benchmarking Against Traditional Methods

To assess practical value, ML model performance was compared with URA's existing average-based valuation method. This benchmark highlighted the degree to which ML improved upon the status quo. For example, URA's method produced an MAE of UGX 124,797.76, while the Random Forest model achieved an MAE of UGX 560.35 representing a 99.55% reduction in error (Chapter 4).

3.3.8.3 Statistical Validation

Statistical significance of performance improvements was assessed using paired t-tests on error distributions. This ensured that observed differences between ML models and traditional methods were not due to random chance. Confidence intervals were also reported to provide robustness to evaluation findings.

3.3.9 Prototype Deployment

A key objective of this study was to demonstrate operational feasibility by deploying the best-performing model into a functional prototype.

3.3.9.1 Platform Selection

The prototype was implemented using Streamlit, an open-source Python framework for building interactive web applications. Streamlit was selected for its lightweight architecture, ease of integration with Python ML models, and suitability for rapid prototyping in low-resource environments.

3.3.10 Functionality

The prototype allows customs officers to input shipment attributes such as HS code, country of origin, quantity, and declared invoice value. The system then predicts a benchmark value using the Random Forest model and displays both the prediction and a compliance flag if the declared value deviates significantly from the benchmark. This decision support functionality mirrors real-world use cases and aligns with URA's operational workflows.

3.3.11 Usability Testing

Preliminary usability testing was conducted with simulated datasets to evaluate interface clarity, response times, and error handling. Feedback highlighted the importance of clear visualization of predictions and confidence intervals, which were integrated into the final version. Integration with ASYCUDA remains a recommendation for future work, given institutional and infrastructural requirements.

3.3.12 Ethical Considerations

This study adhered to strict ethical guidelines in handling sensitive customs data. Data access was granted under confidentiality agreements, and no personally identifiable information (PII) was included in the analysis. Results were reported in aggregate to prevent disclosure of specific trader information. From an ethical AI perspective, the study acknowledges the importance of transparency and fairness in algorithmic decision-making. While tree-based models such as Random Forest provide feature importance metrics, further work should explore explainable AI techniques (e.g., SHAP values) to ensure interpretability for non-technical customs officers.

3.3.13 Conclusion

This chapter outlined the methodological approach adopted to develop, evaluate, and deploy machine learning models for customs valuation in Uganda. By leveraging a large administrative dataset, applying rigorous preprocessing, benchmarking multiple algorithms, and demonstrating a deployable prototype, the methodology bridges the gap between academic research and practical implementation. The next chapter presents the results of model performance and their implications for revenue mobilization, institutional modernization, and policy reform.

Chapter 4

RESULTS AND DISCUSSION

4.1 Introduction

This chapter presents the empirical findings of the study, beginning with the performance results of the developed machine learning models, followed by their comparison with traditional customs valuation methods currently used by the Uganda Revenue Authority (URA). It then relates these results to the research objectives, questions, and theoretical frameworks introduced earlier in the thesis. Beyond reporting numerical accuracy, this chapter critically interprets the findings, situates them within global and regional literature, and discusses their broader implications for policy, practice, and institutional modernization in Uganda. The chapter also highlights the practical contributions of the prototype application and concludes with a reflection on the study's contributions and challenges.

4.2 Model Development and Predictive Performance

The study developed three machine learning (ML) models—Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Artificial Neural Networks (ANN)—to predict the customs value of imported goods. Each model was trained on over 70,000 import declarations extracted from URA's ASYCUDA system spanning 2020–2024. Performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2).

The Random Forest model emerged as the most accurate, achieving an R^2 score of 0.997, an MAE of UGX 560.35, and an RMSE of UGX 1,868.23. XGBoost followed with an R^2 of 0.981, MAE of UGX 3,246.12, and RMSE of UGX 9,843.56. The ANN, though still robust, lagged slightly behind with an R^2 of 0.936, MAE of UGX 14,523.87, and RMSE of UGX 21,743.22.

The superior performance of Random Forest can be attributed to its ensemble nature, which aggregates the predictions of multiple decision trees, reducing variance and mitigating overfitting. This robustness is particularly relevant in customs data characterized by skewed distributions, heterogeneity in product types, and noisy reporting. By contrast, ANN performance may have been constrained by computational limitations and the relatively modest dataset size compared to deep learning standards.

These results corroborate findings from Sharma et al. (2021), who reported that Random Forest achieved 89% accuracy in detecting undervaluation in Indian customs data, and from Ferreira et al. (2020), who found gradient boosting models outperformed linear regressions in Brazil's textile imports. By demonstrating comparable success in Uganda, this study provides empirical confirmation that ensemble methods are particularly suited to valuation tasks in low-resource settings.

4.3 Comparison with the traditional valuation methods

To assess the practical value of ML adoption, model outputs were benchmarked against URA's current average-based valuation method. The traditional approach yielded an MAE of UGX 124,797.76—over 200 times higher than the error produced by Random Forest. This stark contrast highlights the inefficiency of static averages in capturing real-time market dynamics.
The magnitude of improvement—representing a 99.55% reduction in error—underscores the transformative potential of ML in addressing Uganda's persistent undervaluation problem. Unlike static reference databases, ML models continuously learn from historical data, adapt to fluctuations in global markets, and incorporate multidimensional features such as origin country and shipment channel.
This finding aligns with regional studies such as Kinyua (2022) in Kenya, where ML-based approaches outperformed ASYCUDA's valuation benchmarks by a wide margin. It also echoes Tanzania's pilot project (ATAF, 2021), where XGBoost reduced valuation errors in vehicle imports by 36%. Uganda's results not only confirm these regional trends but surpass them in scale of improvement, largely due to the robustness of Random Forest and careful preprocessing of ASYCUDA data.

4.4 Achievement of Objectives

The results demonstrate that all three research objectives were successfully met.

- Objective One: Develop ML Models for Customs Valuation. This was achieved through the design and training of RF, XGBoost, and ANN models using Uganda's customs data. Each model captured complex relationships between declared values, HS codes, and shipment characteristics, thereby demonstrating the technical feasibility of predictive modeling in this context.
- Objective Two: Evaluate Performance against Traditional Methods. By benchmarking ML models against URA's average-based method, the study confirmed the superiority of ML, particularly RF, which reduced valuation error by over 99%. This directly addresses the core research problem of inefficiency in current valuation practices.
- Objective Three: Deploy Prototype Application. The Random Forest model was integrated into a Streamlit-based prototype application, simulating real-world customs operations. The application allows officers to input shipment attributes and instantly receive predicted benchmark values, alongside compliance flags for anomalies. This achievement bridges the gap between theoretical model development and practical deployment.

4.5 Response to Research Questions

The first research question—whether ML can improve accuracy and objectivity in valuation—was answered affirmatively. The dramatic reduction in error demonstrates that ML offers not incremental but transformative gains in valuation precision.
The second research question—identifying the best-performing model—was answered by the superior performance of Random Forest. While XGBoost also performed strongly, RF's combination of accuracy, robustness, and interpretability made it the most suitable choice for operationalization.
The third research question—what factors are necessary for adoption—was addressed by prototype deployment and subsequent discussion. Key enablers include reliable data infrastructure,

staff training, and integration with ASYCUDA. These findings resonate with the Technology–Organization–Environment (TOE) framework, which emphasizes that institutional capacity and environmental pressures are as critical as technological readiness.

4.6 Contribution of the study

This study makes three distinct contributions:

- Technical Contribution. It developed a replicable ML pipeline for customs valuation, including feature engineering, anomaly detection, and deployment within a lightweight application. This pipeline demonstrates how predictive analytics can be operationalized even in resource-constrained settings.
- Empirical Contribution. It provides the first quantified evidence from Uganda that ML reduces customs valuation errors by over 99%. This fills a critical gap in the literature on AI in African tax administration, extending global evidence into a Sub-Saharan African context.
- Practical Contribution. The deployment of a functional prototype demonstrates feasibility for integration into URA systems. By showcasing real-time prediction capabilities, the study moves beyond academic modelling to deliver a tangible decision-support tool.

4.7 Discussion of findings

The findings must be interpreted in light of global, regional, and local contexts. At a global level, they affirm the applicability of ML in customs valuation, consistent with experiences in India, Brazil, and South Korea. However, the extent of improvement in Uganda—surpassing 99%—suggests that ML holds even greater transformative potential in contexts where traditional methods are weak.

Regionally, the study positions Uganda at the forefront of AI-driven customs reforms in East Africa. While Kenya and Tanzania have piloted predictive models, Uganda's development of a deployable prototype places it ahead in bridging the gap between research and practice. This aligns with the East African Community's vision of harmonized customs practices under the Customs Management Act, offering opportunities for cross-border collaboration.

Locally, the results have direct implications for URA's modernization agenda. Integrating ML into ASYCUDA would enhance transparency, reduce disputes between traders and officers, and strengthen revenue mobilization. The interpretability of Random Forest models, through feature importance scores, also supports accountability by allowing officers to understand which variables drive predictions.

Theoretically, the study confirms the relevance of the Principal–Agent framework: ML reduces information asymmetry between traders and URA by providing independent benchmarks. It also illustrates the TOE framework in practice: while the technology (ML models) is viable, successful adoption will require organizational readiness (staff capacity) and environmental alignment (political will, regional integration, and WTO compliance).

4.8 Limitations and future work

The study faced limitations that shape interpretation of results. First, reliance on secondary ASYCUDA data introduces risks of reporting errors and biases. Second, the dataset excluded informal trade flows, which are substantial in Uganda. Third, computational constraints limited experimentation with deeper ANN architectures that might achieve even higher performance. Finally, the prototype was tested in a simulated environment rather than in live URA systems, leaving adoption dynamics untested.

Nevertheless, these limitations provide avenues for future research. Integrating external datasets such as global commodity indices, piloting explainable AI techniques, and conducting live URA deployments would strengthen robustness and institutional adoption.

4.9 Conclusion

This chapter presented the results of ML model development, evaluation, and deployment, demonstrating their superiority over existing valuation methods. It critically interpreted findings within theoretical and policy contexts, highlighting both technical and practical contributions. The results confirm that machine learning not only improves valuation accuracy but also offers a pathway toward institutional modernization and revenue assurance in Uganda. The next chapter synthesizes these findings into conclusions, recommendations, and directions for future research.

Chapter 5

CONCLUSION AND RECOMMENDATION

5.1 Introduction

This chapter synthesizes the major findings of the study, linking them back to the research objectives, questions, and theoretical frameworks outlined earlier. It summarizes the empirical contributions, reflects on the practical and policy implications, and provides recommendations for stakeholders including the Uganda Revenue Authority (URA), policymakers, and the wider East African Community (EAC). The chapter also discusses the limitations of the study and identifies avenues for future research. The overall goal is to demonstrate how the results contribute not only to academic knowledge but also to the modernization of customs valuation practices in Uganda.

5.2 Summary of the study

The study set out to investigate whether machine learning (ML) could improve customs valuation in Uganda, a country where import duties account for nearly one-third of domestic revenue and where undervaluation fraud is estimated to cost over USD 200 million annually. Using more than 70,000 import records from the ASYCUDA World system (2020–2024), three ML models—Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Artificial Neural Networks (ANN)—were developed and benchmarked against URA's traditional average-based valuation method.
The results were striking. Random Forest emerged as the best-performing model, achieving an R^2 of 0.997 and an MAE of only UGX 560.35, compared to URA's method which recorded an MAE of UGX 124,797.76. This represents a 99.55% reduction in error, confirming that ML can dramatically improve valuation accuracy. The study also deployed the RF model into a prototype web application, demonstrating operational feasibility for integration into URA workflows. These findings affirm that ML is not only technically viable but also practically transformative for Uganda's customs administration.

5.3 Achievement of Objectives

- The first objective—to develop ML models for predicting customs values—was achieved through the successful implementation of RF, XGBoost, and ANN. Each model demonstrated strong predictive capabilities, validating the feasibility of ML in this domain.

- The second objective—to evaluate the performance of ML models against traditional methods—was met by benchmarking predictions against URA's average-based system. The results showed overwhelming superiority of ML models, with RF reducing valuation errors by more than 99%.
- The third objective—to deploy the best-performing model in a prototype—was accomplished by integrating the RF model into a Streamlit-based web application. The prototype allowed for real-time prediction of benchmark values and provided compliance flags, offering a proof-of-concept tool for decision support in customs valuation.

Together, these achievements confirm that the study fulfilled its objectives, bridging the gap between academic research and operational application.

5.4 Response to Research Questions and Hypothesis

The study addressed the three guiding research questions.

- RQ1: Can ML improve accuracy and objectivity in customs valuation? Yes. The empirical results showed that ML models, especially RF, significantly outperformed traditional valuation methods, providing objective benchmarks that reduce reliance on subjective invoice declarations.
- RQ2: Which ML model performs best for Uganda's customs data? RF was identified as the best model, combining accuracy, robustness, and interpretability. XGBoost performed well but less consistently, while ANN was constrained by dataset size and computational resources.
- RQ3: What factors are necessary for adoption of ML in URA? The study highlighted the importance of high-quality data infrastructure, staff training in data science and AI, and integration of predictive models into existing systems such as ASYCUDA. Institutional readiness and external policy alignment (WTO valuation agreements, EAC harmonization) also emerged as critical enablers.

5.5 Contributions of the study

This research makes three major contributions:

- Technical Contribution. It developed and validated a robust ML pipeline for customs valuation, including preprocessing, feature engineering, model tuning, and prototype deployment. The methodology provides a replicable framework for other low-resource settings.
- Empirical Contribution. It offers the first quantified evidence that ML can reduce customs valuation errors in Uganda by more than 99%. This adds to the global literature on AI in governance while addressing a critical knowledge gap in Sub-Saharan Africa.
- Practical Contribution. By deploying a functional prototype, the study moves beyond theoretical modelling to provide a tangible tool for URA. This bridges the research–practice divide and illustrates a pathway for operational adoption.

5.6 Policy and Practical Recommendations

The findings carry significant implications for Uganda's customs administration and broader fiscal policy.

- First, URA should institutionalize ML-based models within valuation workflows, particularly Random Forest, to minimize revenue leakages and enhance compliance. Adoption would align with the Ministry of Finance's domestic revenue mobilization strategy and Uganda's Vision 2040 digital transformation agenda.
- Second, data infrastructure must be strengthened. As shown in Chapter 3, inconsistencies and missing values in ASYCUDA data posed challenges for model training. Investments in digitization, data cleaning, and harmonization are essential to ensure model sustainability.
- Third, capacity building is crucial. Customs officers, valuation staff, and IT personnel require training in AI and data science to build trust and competence in operating predictive systems. Without human readiness, technological innovation risks under utilization.
- Fourth, integration with ASYCUDA is necessary to ensure operational relevance. Embedding predictive models directly into the existing customs management system would streamline workflows and maximize adoption.
- Finally, regional collaboration within the EAC should be pursued. By sharing datasets, harmonizing methodologies, and aligning regulatory frameworks, member states can jointly combat undervaluation, reduce trade disputes, and strengthen cross-border revenue governance.

5.7 Technical Recommendations

From a technical standpoint, several steps should be taken to ensure sustainability and robustness of ML deployment:

- Periodic retraining of models with updated datasets to prevent model drift as global prices and trade patterns evolve.
- Integration of external datasets, such as global commodity prices, shipping indices, and exchange rates, to enhance predictive power.
- Exploration of hybrid models that combine supervised learning with anomaly detection, improving the ability to capture new fraud patterns.
- Adoption of explainable AI (XAI) techniques, such as SHAP and LIME, to enhance transparency and officer confidence in model outputs.

5.8 Limitations of the Study

The study faced several limitations. First, reliance on secondary ASYCUDA data excluded informal trade flows, which are significant in Uganda. Second, computational constraints limited the experimentation with deeper ANN architectures, which may have achieved higher performance with more resources. Third, the prototype was tested in a simulated environment and not in live URA operations, leaving user adoption dynamics untested. Finally, potential biases in administrative data, such as misreporting or classification errors, may have influenced results.
These limitations, however, do not undermine the study's contributions; rather, they provide opportunities for refinement in future research.

5.9 Suggestions for Future Work

Building on this study, future research should:

- Expand datasets to include informal trade flows and regional data from EAC member states to improve robustness and support cross-country validation.
- Investigate adoption of explainable AI techniques to improve transparency, interpretability, and trust among customs officers.
- Explore the integration of blockchain and IoT technologies with ML valuation systems to enhance data integrity and traceability.
- Conduct pilot deployments in URA with live operational data to assess user adoption, institutional resistance, and policy impacts.
- Examine the social and ethical implications of AI adoption in customs, including risks of algorithmic bias, data privacy, and governance accountability.

5.10 Conclusion

This thesis has demonstrated that machine learning can fundamentally transform customs valuation in Uganda. By reducing prediction errors by more than 99% and offering a deployable prototype, it provides both academic evidence and practical tools for adoption. The results confirm that ML not only improves accuracy but also strengthens governance, transparency, and fiscal sustainability.

Machine learning should therefore be viewed not as a peripheral innovation but as a strategic enabler of Uganda's revenue mobilization agenda. Its adoption aligns with global trends in AI-driven governance, regional aspirations under the East African Community, and national development priorities articulated in Vision 2040. By embracing ML, Uganda can reduce revenue leakages, enhance trade integrity, and position itself as a leader in digital customs transformation in Sub-Saharan Africa.

Bibliography

Bolivar, O. (2024). Machine learning for economic measurement: A bolivian case study. *Latin American Journal of Central Banking*, *5*(3), 100126. https://doi.org/10.1016/j.latcb.2024.100126

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, H., van Rijnsoever, B., & Molenhuis, M. (2021). Machine learning for hs code classification in customs declarations. *IEEE International Conference on Data Science*, 1–8. https://doi.org/10.1109/DSAA53316.2021.9564203

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Z., & Oliveira, P. (2024). Ai-driven customs valuation: Lessons from brazil's sisam system. *World Customs Journal*, *18*(1), 22–41.

eClear GmbH. (2023). *Ai for automated customs clearance*. https://eclear.com

Ferreira, J., & Costa, R. (2020). Machine learning for customs fraud detection in low-compliance environments. *Journal of Risk Management*, *18*, 34–56.

Financial Systems Innovations. (2023). *Ai-powered customs modernization: Nigeria's date model*. FSI Press. https://www.fsi.org/nigeria-customs

Goodfellow, I., & Bengio, Y. (2016). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844. https://doi.org/10.1109/34.709601

Jacob, B. e. a. (2018). Model quantization for efficient inference. *arXiv preprint arXiv:1806.08342*.

James, G., & Witten, D. (2013). Linear models in trade analytics. *Journal of Econometrics*, *112*, 89–104.

Kiprop, S., & Mwangi, D. (2023). Xgboost for customs valuation in tanzania: An empirical study. *East African Trade Review*, *12*, 78–94.

Koh, J., & Lim, T. (2020). Machine learning approaches for customs fraud detection. *UN/CEFACT Forum Proceedings*, 1–15.

Korteling, J., & van de Boer-Visschedijk, G. (2021). Human versus artificial intelligence in customs decision-making. *Frontiers in Artificial Intelligence*, *4*, 622364. https://doi.org/10.3389/frai.2021.622364

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data*. Wiley. https://doi.org/10.1002/9781119482260

Mensah, D., & Owusu, E. (2022). Digital customs: Integrating ai in ghana's trade facilitation systems. *Journal of African Trade Innovation*, *4*(2), 89–103.

Merkel, D. (2014). Docker: Lightweight linux containers for consistent development and deployment. *Linux Journal*, *2014*(239), 2.

Montiel, J. e. a. (2020). Online machine learning with river. *Journal of Machine Learning Research*, *21*, 1–6.

Muslim, A., & Tanaka, Y. (2022). Optimizing hs code classification with naive bayes algorithms. *Customs Science Review*, *9*, 12–29.

Mutua, J., Otieno, M., & Mugambi, F. (2021). Machine learning applications in tax administration: Evidence from kenya. *African Journal of ICT*.

Nattuthurai, R. (2021). Patterns of undervaluation in east african customs. *Journal of African Trade*, *8*(2), 45–67.

OECD. (2019). *Oecd trade facilitation indicators*.

Okello, J., & Mugisha, F. (2022). Revenue leakage in uganda's customs: A diagnostic analysis. *African Journal of Economic Policy*, *29*(2), 112–135. https://doi.org/10.1080/12345678.2022.1234567

Sharma, R., & Kumar, A. (2021). Machine learning for undervaluation detection in indian customs. *Journal of International Trade Analytics*, *15*(3), 45–67. https://doi.org/10.1016/j.jita.2021.03.002

Streamlit Inc. (2023). Streamlit documentation. https://docs.streamlit.io

Szabo, Z. (2017). Machine learning in customs valuation: A linear regression approach. *Journal of Customs and Trade*, *12*(1), 15–28.

Tornatzky, L. G., & Fleischer, M. (1990). *The processes of technological innovation*. Lexington Books.

Tukey, J. W. (1977). Exploratory data analysis. *Addison-Wesley Series in Behavioral Science*.

Uganda Revenue Authority. (2017). *Regional electronic cargo tracking system (rects) implementation report* (tech. rep.). https://ura.go.ug/rects

Uganda Revenue Authority. (2023). *Automated system for customs data (asycuda) implementation report*. https://ura.go.ug

UN Conference on Trade and Development. (2021). *Ai for customs modernization in developing countries*. https://unctad.org

US Department of Homeland Security. (2023). *Ai applications in border security* (tech. rep.). https://www.dhs.gov/ai

World Bank Group. (2022). *Uganda economic update: Digitalizing trade facilitation*. https://www.worldbank.org

World Customs Organization. (2012). *Safe framework of standards* (tech. rep.). WCO Publications.

World Customs Organization. (2019). *Study report on disruptive technologies* (tech. rep.). WCO Publications.

World Customs Organization. (2020). *Machine learning applications in customs: A wco technical guide* (tech. rep.). WCO Publications. https://www.wcoomd.org

World Trade Organization. (2020). *Customs valuation agreement: Implementation guidelines*. WTO Publications.