SECOND HAND CAR PRICE PREDICTION MODEL IN NAIROBI

BRIAN ATANDI ONYIEGO

A RESEARCH PROJECT SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE IN THE SCHOOL OF SCHOOL OF COMPUTING AND INFORMATICS IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE MASTER OF SCIENCE IN DATA SCIENCE OF THE CO-OPERATIVE UNIVERSITY OF KENYA

OCTOBER, 2025

DECLARATION

Declaration by the candidate

This Project is my original work and has not been presented for award of a degree in any other University or for any other award

…….. …….

Signature Date

Brian Atandi Onyiego

MDATC01/6047/2022

Declaration by the supervisors

We confirm that the work reported in this Project was carried out by the candidate under our supervision and has been submitted with our approval as university supervisors

24 November 2025

Signature Date

Dr. Emma Anyika

Department of Mathematical Sciences, The school of Computing and Mathematics, The Cooperative University of Kenya

24 November 2025

Signature Date

Dr. James Obuhuma

Department of Computer Science, Maseno University

DEDICATION

I dedicate this work to my family, my loving parents and my loving brothers who have unconditionally supported me.

ACKNOWLEGEMENT

To my parents for allowing me to find my own path in life. To uncle Vincent and EGF for keeping me grounded. To my classmates and organizations for inviting me to demonstrate and talk to them what data science entails. To Prof. John Kihoro who dedicated his time to nurture me to the right direction. To Dr. Obuhuma, Dr. Anyika and Prof. Karume for their enormous support and for allowing me to bring my work into existence.

TABLE OF CONTENTS

DECLARATION ii

DEDICATION ii

ACKNOWLEGEMENT iv

LIST OF TABLES x

LIST OF FIGURES xi

LIST OF ABBREVIATIONS xii

OPERATIONAL DEFINITIONS xiv

ABSTRACT xvi

CHAPTER ONE 1

INTRODUCTION 1

1.0 Introduction 1

1.1 Background of the study 1

1.2 Problem Statement 3

1.3 Objectives 4

1.3.1 General Objectives 4

1.3.2 Specific Objectives 4

1.4 Research Questions 5

1.5 Significance of the study 5

1.6 Scope of the study 6

1.7 Limitations of the study 6

CHAPTER TWO 8

LITERATURE REVIEW 8

2.0 Introduction 8

2.1 Theoretical Framework 8

2.2 Empirical Literature 9

2.3 Conceptual Framework 11

2.4 Research Gap Analysis 15

2.5 Critique of Literature 20

2.6 Research Gap 21

CHAPTER THREE 23

METHODOLOGY 23

3.1 Introduction 23

3.2 Research Paradigm 23

3.3 Research design 24

3.4 Study Area 25

3.5 Population and Sampling 25

3.5.1 Population 25

3.5.2 Sample Size Calculation 26

3.5.3 Sampling Technique 27

3.6 Data Collection 27
3.6.1 Web Scraping 27
3.7 Data Collection procedures 28
3.7.1 Data Cleaning 28
3.7.2 Data Transformation 29
3.8 Model Development 30
3.8.1 Choice of Machine Learning Algorithms 30
3.8.2 Model Training 31
3.9 Validation 31
3.10 Model Testing 32
3.10.1 Data Analysis 33
3.11 Ethical Considerations 34
CHAPTER FOUR 35
DATA ANALYSIS, PRESENTATION AND INTERPRETATION 35
4.0 Introduction 35
4.1 Features Impacting Second-hand Car Prices 35
4.1.1 Correlation and Feature Importance Analysis of Predictors 35
4.2 Dataset Curation 39
4.2.1 Data Collection 39
4.2.2 Preprocessing and Cleaning 40

4.2.3 Interpretation 44
4.3 Model Development 44
4.3.1 Data Analysis 44
4.3.2 Presentation 45
4.4 Model Evaluation 50
4.4.1 Data Analysis 50
4.4.2 Presentation 51
4.4.3 Interpretation 53
CHAPTER FIVE 55
CONCLUSION AND RECOMMENDATIONS ..... **Error! Bookmark not defined.**
5.1 Introduction 55
5.2 Summary of Findings 55
5.3 Conclusion 57
5.4 Recommendations 58
5.4.1 Policymakers and regulators 58
5.4.2 Dealers, lenders, and insurers 59
5.4.3 Platform owners and data providers 59
5.5 Suggestions for further research 59
References 61
APPENDICES 64

Appendix 1: Research Instruments 64

Appendix II: Research Permits/authorization letter 64

Appendix III: Published articles of your thesis 65

Appendix IV: Figures and tables, ... 70

LIST OF TABLES

Table 2.1 Research Gap Analysis in Used Car Price Prediction ……13

Table 3.1 Data Analysis…… 29

Table 4.1.1 Pearson Correlation Coefficients Between Key Predictor Variables and Second-hand Car Prices…….32

Table 4.2.0 Data Cleaning Tasks and Outcomes……36

Table 4.2.1 Summary Statistics for Key Numerical Features…….38

Table 4.2.2 Frequency Distribution of Selected Categorical Variables……40

Table 4.2.3 Summary of Mean Price and Baseline Model Error……41

Table 4.3.2.2.1 Performance Comparison of Random Forest Models Across Training Stages……..43

Table 4.3.2.3.1 Performance Metrics of the Final Random Forest Model……..44

Table 4.3.2.4.1 Comparative Performance of Random Forest, XG-Boost, and Baseline Models……45

Table 4.6 Performance Comparison of Random Forest and XG-Boost on Test Dataset..

Table 4.7 K-Fold Cross-Validation Scores for the Two Models……..47

LIST OF FIGURES

Figure 2.1 – Conceptual Framework for Second Hand Car Price Prediction Model

LIST OF ABBREVIATIONS

Mathematical and Statistical Symbols

- R^2 R squared (coefficient of determination)
- MAE Mean Absolute Error
- RMSE Root Mean Square Error

Technical Abbreviations

- ML Machine Learning
- IoT Internet of Things
- CNN Convolutional Neural Network
- LSTM Long Short-Term Memory
- SVM Support Vector Machine
- KNN K Nearest Neighbors
- PSO Particle Swarm Optimization
- GRA Grey Relational Analysis
- BPNN Back Propagation Neural Networks
- VIF Variance Inflation Factor
- OLS Ordinary Least Squares
- IQR Interquartile Range

Organizations and Standards

- EGF (Not defined in the document)
- SLA Service Level Agreement
- NACOSTI National Commission for Science, Technology and Innovation

Data and Programming Related

- CC Cubic Centimeters

Units of Measurement

- EUR Euro (currency)

Model Performance Metrics

- K Fold – Technique used in machine learning to see how well the model was to perform on unseen data

OPERATIONAL DEFINITIONS

Machine Learning Model – The machine learning model functions as a computer trained program which detects patterns in fresh unknown data.

Feature Engineering – Feature Engineering represents the method of converting unprepared data into formats that machine learning requires.

Data Preprocessing – Data Preprocessing includes all the steps needed to prepare raw information for a machine learning model entry point.

Cochran's Sample Size Formula -The statistical Cochran's Sample Size Formula helps researchers decide suitable sample sizes based on population requirements.

Over fitting - the model shows great results on training data yet performs poorly on new untested data because of a condition known as over fitting.

Random Forest – Random Forest operates as an ensemble machine learning approach which deploys multiple decision.

XG-Boost (Extreme Gradient Boosting) – XG-Boost demonstrates high performance when processing structured data which makes it an optimal choice for price prediction because of its accuracy and speed.

Web Scraping – Web Scraping functions as an automated system to extract structured data from websites and served this research by collecting second hand car listings.

Root Mean Squared Error (RMSE) – Root Mean Squared Error (RMSE) functions as a predictive model evaluation method which calculates the average absolute magnitude of errors while emphasizing larger errors.

Mean Absolute Error (MAE) – The Mean Absolute Error (MAE) provides model accuracy measurement by determining the average absolute value of actual predicted value differences.

Feature Importance – Machine learning practitioners use Feature Importance as an interpretability tool to establish the variable impact on target predictions.

ABSTRACT

The second-hand car market in Nairobi continues to grow rapidly, creating a need for accurate and transparent price prediction methods. Traditional valuation approaches rely heavily on subjective judgement, leading to inconsistent and unreliable pricing. This study aimed to develop a data-driven machine learning model capable of predicting second-hand car prices using structured vehicle characteristics such as year of manufacture, mileage, engine capacity, brand, and model. The population consisted of all vehicles listed on the SBT Japan online platform in Kenya. A total of 29,000 records were collected through web scraping, and after cleaning and preprocessing, 20,775 records were retained for analysis and modelling. Feature analysis showed that model, brand, engine capacity, year of manufacture, and mileage were the most influential predictors of price. Two ensemble learning models, Random Forest and Extreme Gradient Boosting, were developed and evaluated. The Extreme Gradient Boosting model achieved the highest accuracy, with a mean absolute error of 95,696.60 Kenyan shillings, a root mean square error of 190,939.99 Kenyan shillings, and a coefficient of determination of 0.99379, which represents a substantial improvement over the baseline error of 1,839,811.92 Kenyan shillings. The study concludes that machine learning provides a reliable, consistent, and highly accurate approach for predicting second-hand car prices in Nairobi, offering practical value to car dealers, financial institutions, insurers, online marketplaces, and potential buyers seeking transparent and data-driven pricing.

CHAPTER ONE

INTRODUCTION

1.0 Introduction

This chapter discussed about the background of the research topic, the research problem, objective of the research, the research questions, hypothesis of the study, the importance of the study, justification of the research, limitation of the research, expected outcome of the research and study area of the research.

1.1 Background of the study

The second-hand car market has experienced notable growth globally and within Kenya, driven by rising vehicle importation rates, affordability considerations, and increased consumer demand for used vehicles (Ghosh, 2018). This expansion has heightened the need for accurate valuation methods, since traditional approaches such as manual appraisals and reference price guides often rely heavily on individual judgement and may not capture the full range of factors influencing vehicle prices (Çelik and Osmanoğlu, 2019). Studies have shown that conventional valuation practices frequently omit important determinants such as detailed vehicle condition, model-specific depreciation patterns, and real-time market dynamics, resulting in inconsistent and sometimes inaccurate pricing outcomes (Bukvić *et al.,* 2022).

Advances in big data analytics and machine learning have significantly transformed valuation processes across multiple industries, including the automotive sector. Machine learning enables the analysis of large and heterogeneous datasets and supports the identification of patterns that are not immediately visible through traditional statistical

techniques (Huang, 2023). In the context of second-hand cars, machine learning models can incorporate multiple attributes such as make, model, year of manufacture, mileage, engine capacity, fuel type, and market demand to generate objective and replicable price estimates (Liu *et al.*, 2022). This data-driven approach reduces subjectivity and improves pricing transparency in markets characterized by information asymmetry between buyers and sellers.

Within data science, second-hand car price prediction is typically treated as a supervised learning regression problem in which models are trained on labelled historical data to predict continuous price values (Yadav *et al.*, 2021). Techniques such as regression modelling, decision trees, ensemble learning, and feature engineering have been widely applied to vehicle pricing tasks, enabling researchers to test which methods best capture the multivariate relationships influencing resale value (Asghar *et al.*, 2021). As online vehicle listing platforms and dealerships increasingly digitise their records, larger and more diverse datasets have become available, enhancing the potential of machine learning to support reliable valuation in markets like Kenya (Huang, 2023).

However, the adoption of machine learning for valuation also raises important considerations related to fairness, transparency, and responsible use of data. Differences in regional import patterns, consumer preferences, and market regulation mean that valuation models developed in other countries may not automatically generalize to Kenya (Chen *et al.*, 2022). This underscores the need for context-specific studies that analyse local vehicle data and develop predictive models tailored to the structure of the Kenyan second-hand car market (Barlybayev *et al.*, 2023).

1.2 Problem Statement

Even with the continuous growth of the second-hand car market in Kenya, determining the true value of a used vehicle remains a major challenge. Traditional valuation methods largely rely on manual inspection, human judgement and experience, which often results in inconsistency, bias, and inaccurate pricing outcomes. Studies show that conventional approaches frequently fail to capture the full spectrum of factors that influence vehicle value, including brand reputation, vehicle age, mileage, market demand, economic conditions, and the car's overall condition (Çelik & Osmanoğlu, 2019; Bukvić *et al.*, 2022). These limitations make manual valuation inadequate, especially in competitive and dynamic markets where small price variations significantly affect buyer and seller decisions.

The valuation process becomes even more complicated when vehicle data is large, unstructured, or sourced from multiple online platforms. Modern vehicle listing sites contain heterogeneous data with varied formats, missing values, and inconsistent reporting standards, making traditional analytical methods difficult to apply effectively (Huang, 2023). The Kenyan market also changes rapidly due to fluctuations in import patterns, shifts in foreign exchange rates, and evolving consumer preferences, which further complicate price determination (Ghosh, 2018). Because of these unpredictable market forces, both underpricing and overpricing are common, exposing buyers to financial loss through inflated prices and exposing sellers to losses when vehicles are undervalued.

Although machine learning techniques have been successfully used in other regions to address similar valuation challenges, the majority of existing studies focus on developed

markets such as China, India, South Africa, and Croatia, with limited application in Kenya (Asghar *et al.*, 2021; Liu *et al.*, 2022; Huang, 2023). Many of these studies rely on small datasets or narrow vehicle categories, such as research limited to Tesla vehicles or specific European markets, making their findings difficult to generalize to the Kenyan context (Arefin, 2021; Bukvić *et al.*, 2022). This lack of location-specific research and the absence of large, well-curated datasets lead to models that do not adequately capture the unique characteristics of Kenya's import-driven second-hand car ecosystem. As a result, there remains a methodological gap in developing reliable, scalable, and context-appropriate predictive models tailored to Kenya's market.

For this reason, there is a clear need to develop a robust data-driven model that integrates multiple vehicle attributes and leverages modern machine learning techniques to enhance the accuracy, consistency, and fairness of second-hand car price prediction in Kenya. Such a model would address existing limitations, reduce subjectivity in valuation, and support more informed decision-making for buyers, sellers, insurers, lenders, and digital marketplaces.

1.3 Objectives

1.3.1 General Objectives

The general objective of this research was to develop a machine learning based model that accurately predicts the price of second-hand cars in Kenya.

1.3.2 Specific Objectives

i. To collect, clean and prepare a structured dataset of second-hand car records.

ii. To determine the most influential features impacting second hand car prices.

iii. To develop a model that predicts the prices of second-hand cars.

iv. To evaluate the model that predicts the prices of second-hand cars.

1.4 Research Questions

The research questions are:

i. What approach is used to collect, clean and prepare data for modelling?

ii. What are the most influential features that impact second hand car prices?

iii. What approach is used to develop a prediction model?

iv. What methods are used to evaluate a predictive machine learning model?

1.5 Significance of the study

This study was to primarily focus on revolutionizing the marketing strategy, boosting profits and enhancing customer satisfaction in the secondhand car market. The goal of the project is to change the used vehicle pricing industry by providing a sophisticated, data driven system that overcomes the drawbacks of conventional pricing techniques. This research has the potential to significantly change the used cars market environment by utilizing the strength of big data analytics, cutting edge machine learning techniques, and domain experience. The study was conducted on the strong foundation of historical data and the knowledge of machine learning and modeling.

This study is significantly important because it can improve market valuation efficiency hence reducing biasness in the valuation process. It seeks to streamline the information asymmetry that frequently occurs between buyers and sellers in the used automotive market by offering precise and fast pricing estimates. It aims to increase transparency in

this transaction process and build customer trust in this market. This was to boost market liquidity and optimize resource allocation in the automotive industry. The impact is particularly significant for consumers. Vehicle overspending is less likely for buyers, and sellers can more boldly set competitive pricing that accurately reflect a vehicle's genuine market value. By providing trustworthy information, market players might be empowered to promote a fairer marketplace and possibly reduce the occurrence of dishonest business practices or fraud.

1.6 Scope of the study

The study was conducted in Nairobi County, the largest and most active second-hand vehicle market in Kenya. It focused specifically on Japanese used vehicles listed on the SBT Japan online platform, which represents a substantial share of the country's import-driven car market. The study examined key vehicle characteristics such as year of manufacture, mileage, engine capacity, brand, and model, with the goal of developing a machine learning-based price prediction model tailored to the Kenyan context. The scope of the study was limited to structured tabular data and did not include image-based or text-based vehicle descriptions. The research concentrated on understanding price determinants, building predictive models, and evaluating their accuracy within Nairobi's commercial environment, where vehicle demand and pricing dynamics are most pronounced.

1.7 Limitations of the study

This study encountered several limitations. First, accurate price prediction is affected by variations in vehicle condition, yet the dataset lacked detailed physical inspection attributes such as accident history, maintenance records, and wear-and-tear information, which have

been shown to influence valuation accuracy in previous studies (Bukvić *et al*., 2022). Second, the study relied on data from a single online platform, which may not fully represent vehicles sold through local dealerships, auctions, or private sellers, possibly limiting the generalizability of the findings. Third, some vehicles may have undergone part replacements, modifications, or repairs that were not captured in the dataset, making it difficult to reflect their true market value. Additionally, although 29000 records were collected, the initial data contained duplicates, missing fields, and outliers, which required removal and resulted in a reduced modelling dataset of 20,775 records. Finally, the model does not account for external economic factors such as foreign exchange fluctuations, inflation, or changes in import taxation, which studies such as Ghosh (2018) highlight as significant influences on second-hand car prices.

CHAPTER TWO

LITERATURE REVIEW

2.0 Introduction

Valuation is the process of determining the accurate value of a commodity within its market. There are several modes of car valuation within the second-hand car market. Some of these modes include the use of manual appraisals, use of guide books, market value and forced sale value. Nowadays, there is an advancement in this area where there is the use of advanced techniques such as blockchain technology, internet of things (IoT) and machine learning (ML). The used car market is still facing challenges in determining the accurate value of a used car by considering a multi-dimensional array of factors that influence the car price.

2.1 Theoretical Framework

This study is grounded in the Hedonic Pricing Theory, originally introduced by Rosen (1974). The theory explains that the price of a product is determined by the value of its individual attributes rather than by the product as a whole. Rosen argued that consumers make purchasing decisions based on the utility derived from specific characteristics of a good, and therefore market prices reflect the combined contribution of these characteristics. Applying this theory to second-hand vehicles, the price of a used car is influenced by measurable attributes such as year of manufacture, mileage, brand, model, and engine capacity. Each of these characteristics contributes to the overall value of a vehicle, and variations in any attribute lead to corresponding variations in price. This study uses Rosen's

Hedonic Pricing Theory as the guiding framework for identifying and modelling the key determinants of second-hand car prices in Nairobi.

2.2 Empirical Literature

Data collection, cleaning and preparation

Data quality, integrity and quantity are the key factors that determine how accurate the price prediction was to be. Msiza (2023) states that successful predictive models need to be developed using historical data. The research of the South African market shows that it is beneficial to have extensive datasets to generate more accurate price predictions and they are also a valuable resource for both purchasing and selling entities. The predictive accuracy achieved by Çelik and Osmanoğlu (2019) using machine learning techniques was 81.15 percent, which proves that data-driven methods are important for second-hand car market pricing. Huang (2023) focuses on the necessity of mining useful information from large datasets, as seen in popular second-hand car trading platforms, to develop a reliable price prediction model. These studies collectively highlight the importance of dataset completeness, pre-processing, and feature engineering for building robust predictive models.

Feature Engineering

The identification of important characteristics that affect a used car's price is one of the core components of used car price prediction. Bukvić *et al*. (2022) conducted a detailed analysis of the Croatian second-hand vehicle market, using data from the Kaggle dataset to evaluate various supervised machine learning algorithms such as support vector machines, logistic regression, and random forests. Their results highlight how important feature selection is for improving prediction accuracy. According to existing research,

multiple linear regression is a standard analysis method for forecasting used car prices. In Lu and Song (2023), the authors show how multiple linear regression can be used to determine the relationship between different variables and second-hand car prices in China. According to their findings, the second-hand car market expansion requires precise price forecasting for the buyers as well as the sellers. Gupta *et al.* (2021) mentions that there are several factors that need to be considered when developing predictive models, including vehicle condition and mileage, and brand reputation. There are multiple complex aspects that need to be modelled to identify their interactive patterns.

Model Development

Machine learning methods have become a popular field in addition to linear regression. Liu *et al.* (2022) propose a hybrid model utilizing particle swarm optimization, grey relational analysis, and back propagation neural networks to enhance prediction accuracy. According to their research, combining these methods can result in more accurate price forecasts. This is consistent with a larger trend in the literature toward ensemble methods, which combine two or more algorithms to achieve better performance. It is further embraced by the work of Yadav *et al.* (2021), who developed a machine learning model to predict used car prices based on different parameters, demonstrating the versatility and effectiveness of these approaches. Moreover, the integration of advanced machine learning algorithms has led to significant improvements in prediction accuracy. Zhu (2023) explores the use of XG-Boost, support vector machines, and neural networks for estimating transaction prices of second-hand cars, providing a comparative analysis of their effectiveness. The results indicate that machine learning models can outperform traditional methods, thereby enhancing the reliability of price predictions. This trend is corroborated

by the research of Barlybayev (2023), who applied multi-layer perceptron networks to predict resale values in developing markets, showcasing the adaptability of machine learning techniques across different contexts.

Model Evaluation

The field of second-hand car price research was centered on external variables like government policy changes and economic forces. Ghosh (2018) investigates the effects of policy modifications on consumer conduct which subsequently affects second-hand car market pricing. External market influences with regulatory frameworks form a connection to market dynamics, thus price prediction models need to incorporate these factors to achieve improved accuracy. Chen *et al.* (2022) presents blockchain technology as an effective tool to enhance trust between used car buyers and sellers by addressing asymmetric information in the market.

The evolution of price prediction models has also been marked by the exploration of hybrid approaches that combine various methodologies. For instance, Li li *et al.* (2023) advocate for a hybrid forecasting model based on stochastic algorithms, emphasizing the need for a unified price evaluation standard in the used car market. These studies contribute to understanding how predictive models should be assessed, benchmarked, and validated against complex market conditions to ensure consistent and reliable performance.

2.3 Conceptual Framework

The selection of independent and dependent variables in this study was guided directly by the evidence presented in previous research on second-hand car valuation. The foundation

for identifying the independent variables is based on the hedonic pricing theory, which explains that the price of a product is determined by the characteristics that make up that product. This theory has been used extensively in vehicle valuation research, and studies by Çelik and Osmanoğlu (2019) and Bukvić *et al*. (2022) show that specific vehicle attributes consistently influence resale price. Consequently, the variables chosen as independent variables were those repeatedly identified in prior studies as the strongest predictors of vehicle price.

Year of manufacture was selected as an independent variable because multiple studies have demonstrated that vehicle age strongly affects depreciation and market value. Research conducted in China by Liu and Song (2023), in Croatia by Bukvić *et al*. (2022), and in broader markets reviewed by Asghar *et al*. (2021) consistently show that newer vehicles command higher prices, while older vehicles depreciate more rapidly. Mileage was included because empirical studies repeatedly show that higher mileage reduces vehicle value due to increased wear and tear. This relationship is well documented in the works of Asghar *et al*. (2021) and Bukvić *et al*. (2022), where mileage was among the most influential determinants of used car pricing.

Brand and model were selected after reviewing studies that found strong evidence that certain brands retain value better due to reliability, consumer trust, and market demand. Arefin (2021) demonstrated this effect even in studies focusing on a single manufacturer, while Msiza (2023) and Huang (2023) highlighted that brand characteristics significantly shape customer willingness to pay. Engine capacity was chosen because it has been shown to influence performance expectations and market preferences. Studies such as Huang

(2023) demonstrate that larger engines often attract higher prices in many markets, making engine capacity an essential attribute for modelling vehicle price.

The dependent variable was identified as the price of the used car. This choice is consistent with the objectives of almost all previous studies examined in the literature review. Researchers such as Fathalla *et al*. (2020), Arefin (2021), and Huang (2023) designed models where the primary output variable was the resale or transaction price. Because the aim of this study is to predict market price accurately, price was selected as the dependent variable in line with established research practice.

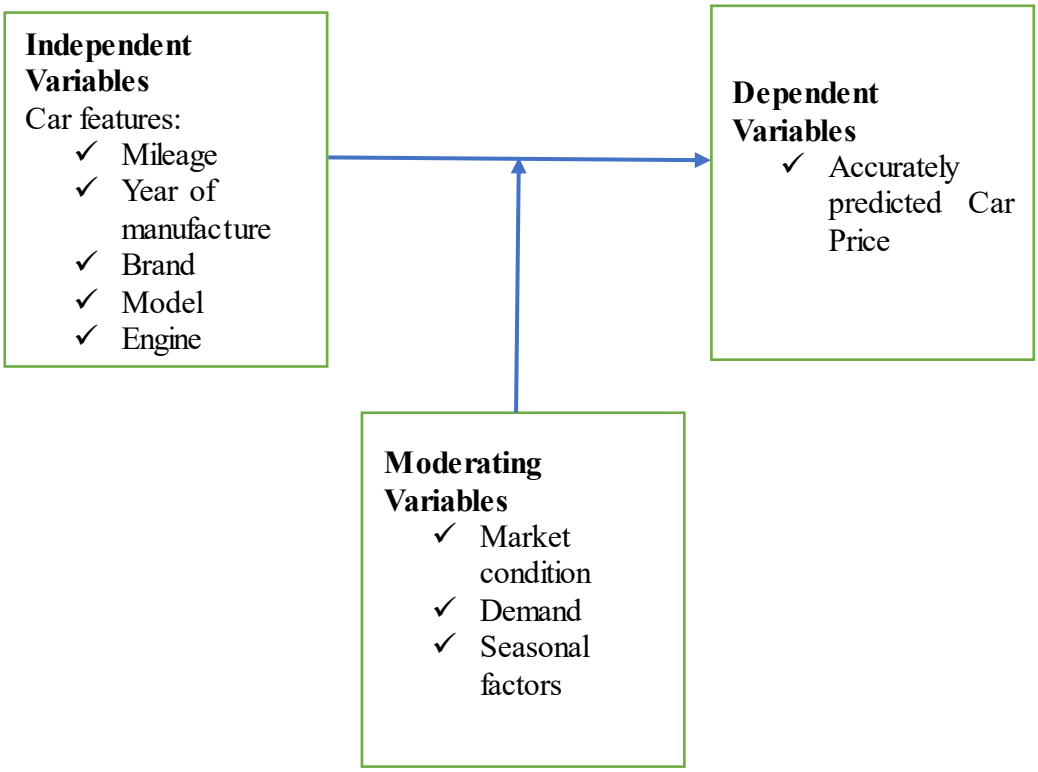
Moderating variables were identified by examining literature that discusses broader environmental factors affecting second-hand car pricing. Market conditions, including inflation, import duties, and economic policy changes, were included as moderating variables based on evidence from Ghosh (2018) who showed that policy shifts and external economic forces significantly influence second-hand car prices. Demand trends were incorporated because studies by Bukvić *et al*. (2022) and Liu *et al*. (2022) highlight that consumer preferences and brand popularity influence how vehicle characteristics translate into price differences. Seasonal factors were included because market demand fluctuates at different times of the year, and these fluctuations influence how much buyers are willing to pay, even when vehicle characteristics remain constant. Although not emphasized in all empirical studies, these factors are recognised in market behaviour and therefore serve as moderating influences.

In summary, the independent variables were selected because they have consistently demonstrated a direct influence on vehicle price across multiple studies. The dependent variable was chosen because price is the primary outcome measured in previous research.

The moderating variables were identified from literature that highlights the effect of economic conditions, consumer behaviour, and temporal patterns on the relationship between vehicle characteristics and their selling price. This ensures that the conceptual framework for this study is firmly grounded in theory and supported by empirical evidence.

Figure 2.1

Conceptual framework



This framework provides a clear structure for identifying how various factors influence the price of used cars, while acknowledging external elements that may mediate this relationship.

2.4 Research Gap Analysis

Over the past years, several researchers have explored the subject of used car price prediction using different methodologies and datasets. However, their studies still leave important gaps that require further investigation. The section below presents selected past studies and highlights their key contributions, the methods they used, their findings and the specific gaps they did not address.

Table 2.1

Research Gap Analysis in Used Car Price Prediction

Study	**Brief Description**	**Methodology**	**Dataset**	**Findings**	**Gap**
Fathalla *et al.* (2020) Deep end to end learning for price prediction of second-hand items	Proposed a deep model for predicting the price of second-hand items based on image and	Deep Neural Networks (CNN + LSTM)	Dataset: Mercari Price Suggestion Challenge (2018) Size/quantity: 107,000 observations	Achieved better MAE scores compared to the baseline	The method does not address dynamic market factors, which may impact pricing

	textual descriptions				accuracy over time
Arefin (2021) Second Hand Price Prediction for Tesla Vehicles	Developed a price prediction system for second hand Tesla vehicles using machine learning techniques	Boosted decision tree regression, Random Forest, SVM, Deep Learning	Dataset: Scraped from truecar.com, autotrader.com, and cars.com Size/quantity: 1,600 entries	Boosted Decision Tree Regression provided the best RMSE for Tesla price prediction, outperforming other models	Limited scope to Tesla vehicles only; no exploration of broader vehicle markets or the impact of additional external factors like market trends or user reviews
Asghar *et al.* (2021) Used Cars Price	Developed a machine learning model for	Recursive Feature Elimination (RFE), OLS	Dataset: Kaggle	Achieved a 90% R^2 score in predicting	Limited focus on nonlinear models, such

Prediction using Machine Learning with Optimal Features	predicting used car prices using optimal features	Regression, VIF (Variance Inflation Factor)	Size/quantity: 205 entries	car prices using features such as fuel type, car body, and mileage	as neural networks, that might capture more complex relationships between car attributes and price
Samruddhi and Kumar (2020) Used Car Price Prediction using K Nearest Neighbors Based Model	The study proposes a supervised machine learning model to predict used car prices using the K Nearest Neighbors (KNN) algorithm.	KNN regression algorithm, Data preprocessing, Cross validation using K Fold method	Dataset: Kaggle Size/quantity: 14 variables	Achieved accuracy of around 85%. Best performance with K=4. Root Mean Squared Error (RMSE) of 4.01. Mean Absolute Error (MAE) of	Limited to a single algorithm (KNN). Small dataset used. No comparison with other advanced machine learning techniques. No feature importance

				2.01. Cross validation with 10 folds yielded 82% accuracy	analysis. Lack of detailed error analysis. No discussion on model interpretability
Bukvić *et al.* (2022) Price Prediction and Classification of Used Vehicles Using Supervised Machine Learning	The study predicts used vehicle prices in Croatia using machine learning models, focusing on production year and kilometers	Supervised machine learning techniques, including linear regression and classification. Data preprocessing involved web scraping and	Dataset: online retail web portal "Njuškalo" in Croatia Size/quantity: 8,710 records	The model achieved 95% accuracy in predicting price increases, validating its effectiveness. A linear regression model predicted a EUR 1391 increase in	The study uses limited attributes (e.g., price, kilometers, year) and is restricted to the Croatian market. It lacks integration of other potential predictive factors like

	travelled as key attributes.	removing outliers.		vehicle prices.	fuel type, condition, or market trends, and does not explore advanced ML techniques like neural networks.

Table 2.1: *Summary of Empirical Studies on Second-Hand Car Price Prediction and Identified Research Gaps*

Previous research demonstrates multiple areas where further investigation is needed concerning used car price prediction. Multiple studies concentrate on Tesla vehicles as their subject matter but fail to analyze market wide trends involving various vehicle types. Multiple prediction models demonstrate strong accuracy but most of them neglect important time dependent factors which affect car prices including market trends and economic conditions. Several predictive models have restrictive limitations because they exclude crucial variables such as fuel type and condition alongside government policies from their analysis. The proposed approach integrates the missing variables with diverse

vehicle types and external factors using nonlinear prediction models to build accurate and reusable price predictions across markets.

2.5 Critique of Literature

The reviewed literature provides substantial insights into the determinants of second-hand car prices and the application of advanced machine learning models for price prediction. However, most studies have been conducted in developed markets such as China, India, and Croatia, with minimal attention to African contexts where market structures, regulatory policies, and consumer behaviors differ significantly. This limits the applicability of these models in local markets such as Kenya, where factors like import duties, currency fluctuations, and informal transactions play a key role in determining car prices. Additionally, while machine learning models such as neural networks, random forests, and ensemble approaches have demonstrated strong predictive capabilities, they often lack interpretability, making it difficult to understand how individual features such as mileage or engine capacity influence pricing outcomes.

Furthermore, existing studies tend to rely heavily on structured online datasets, which may not represent the full spectrum of market data, thereby reducing the reliability of the predictions. Most research also overlooks external economic factors such as inflation, interest rates, and consumer demand, which can substantially affect car values over time. Although hybrid models have been developed to improve accuracy, little attention has been given to their practicality, cost-effectiveness, and scalability in developing economies. These gaps highlight the need for more context-specific, transparent, and comprehensive models that integrate both internal vehicle attributes and external market dynamics to enhance the accuracy and applicability of second-hand car price prediction systems.

2.6 Research Gap

Existing studies on second-hand car price prediction have made progress in applying machine learning techniques, but several important gaps remain unaddressed. Many studies relied on small datasets, sometimes as low as 205 or 1,600 records, which limits model accuracy, weakens generalization, and increases the risk of overfitting (Asghar *et al.*, 2021; Arefin, 2021). Other studies were specific to particular vehicle categories, such as Tesla-only datasets or single-brand analyses, preventing broader market applicability (Arefin, 2021). A number of studies focused on foreign markets such as China, India, South Africa, and Croatia, which have different consumer preferences, import structures, regulatory frameworks, and pricing behaviours than those observed in Kenya (Liu & Song, 2023; Bukvić *et al.*, 2022; Msiza, 2023). As a result, their models cannot be directly applied to the Kenyan second-hand car ecosystem.

Several authors also used limited feature sets, focusing only on price, mileage, and year, while excluding important variables such as model, engine capacity, transmission, and fuel type, which have been shown to influence price in multiple studies (Bukvić *et al.*, 2022). Some methods relied solely on linear or single-algorithm approaches, even though recent literature demonstrates that ensemble methods and hybrid models generally produce superior accuracy for structured data (Huang, 2023; Liu *et al.*, 2022). In addition, very few studies in the reviewed literature integrated large-scale datasets extracted from online platforms, despite the fact that online listings now represent the most comprehensive source of second-hand vehicle information. Finally, none of the reviewed studies developed a machine learning model specifically tailored to Kenya's import-driven automotive market,

which operates under unique conditions such as high dependence on Japanese vehicles, fluctuating exchange rates, and rapidly changing demand patterns (Ghosh, 2018).

These gaps demonstrate a clear need for a comprehensive, data-driven, and Kenya-specific predictive model that uses a large dataset, incorporates multiple vehicle characteristics, and applies modern machine learning techniques to improve the accuracy and reliability of second-hand car price prediction in Nairobi and other similar markets in Kenya.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This part details the systematic development process for building the second-hand car price prediction model. The framework contains information about research paradigm and design together with data collection methods and preprocessing steps and model development and evaluation processes. The chapter delivers ethical considerations together with explanations of analytical instruments that were used. The research design creates an organized structure to allow independent replication of the study using systematic scientific methods for achieving its research goals.

3.2 Research Paradigm

This study adopts the positivist paradigm, which emphasizes the use of scientific methods, quantitative data, and objective analysis to understand phenomena. The positivist approach is suitable for this research as it seeks to develop a predictive model for second hand car prices based on measurable factors such as mileage, year of manufacture, brand, model and engine capacity. The paradigm supports the use of machine learning algorithms to establish relationships between these variables and derive data driven conclusions. By focusing on empirical evidence and quantifiable data, the study aims to produce reliable and generalizable results.

3.3 Research design

The study adopted a quantitative research design because the objective was to analyse numerical vehicle data and develop a data-driven model for predicting second-hand car prices. The quantitative approach allowed systematic processing of large datasets, identification of statistical relationships among vehicle characteristics, and evaluation of model accuracy using numerical performance measures.

Within the quantitative framework, the study employed an exploratory component to examine patterns in the data and identify the key features that influence second-hand car prices. This exploratory analysis was necessary because the dataset contained diverse vehicle attributes whose relationships were not predetermined. Techniques such as correlation analysis, feature importance ranking, and descriptive statistics were used to uncover these patterns, consistent with prior studies that highlight the role of exploratory analysis in understanding used car markets (Bukvić *et al.*, 2022; Asghar *et al.*, 2021).

The study also incorporated a predictive modelling component, which involved training, validating, and testing machine learning models using the curated dataset. This predictive approach aligns with the purpose of the study, which was to estimate market prices based on historical data. Although the modelling process involves repeated training and testing, it does not constitute an experimental design in the classical sense of manipulating variables. Rather, it reflects standard supervised learning procedures, where model performance is evaluated by comparing predicted prices with actual prices from the dataset.

This combined quantitative, exploratory, and predictive design ensured that the study was able to identify key pricing determinants, develop an accurate machine learning model, and assess its performance objectively using established evaluation metrics.

3.4 Study Area

The study was conducted in Nairobi County, Kenya, which is the country's capital and a major hub for economic and commercial activities. Nairobi hosts one of the largest second-hand car markets in East Africa, with numerous dealerships, showrooms, and online trading platforms that provide a diverse range of vehicle data. The county was selected because it offers a representative environment for studying the dynamics of used car pricing, given its high vehicle demand, accessibility to both buyers and sellers, and concentration of automobile importers and valuation experts. Additionally, Nairobi's strong digital infrastructure and connectivity make it suitable for collecting accurate and comprehensive data on vehicle characteristics, market trends, and pricing patterns, thereby providing a robust basis for analyzing second-hand car price prediction models.

3.5 Population and Sampling

3.5.1 Population

The population for this study consisted of all used cars listed on SBT Kenya Japanese Used Cars, a prominent platform for second hand car sales in Kenya. This study focused on Japanese used cars mainly because, different researches indicate that tailoring your machine learning model to a specific geographical region enhances their predictive accuracy and relevance. For instance, Tucci *et al.* (2024) introduced a regional feature selection-based machine learning system for short term air quality prediction,

demonstrating the importance of regional feature selection in environmental modeling. Similarly, Fang *et al.* (2022) developed a machine learning approach for predicting electricity production from solar photovoltaic installations at a regional level in Italy, which is crucial for effective grid management and energy planning.

3.5.2 Sample Size Calculation

To determine the appropriate sample size, the study used the Cochran *et al.* (1997) formula for large populations:

Equation 3.3.2.1 Sample Size Calculation Formula

$$n = \frac{Z^2 p(1-p)}{e^2}$$

Where:

- n= sample size
- Z = Z value (1.96 for a 95% confidence level)
- p = estimated proportion of the population (0.5 for maximum variability)
- *e* = margin of error (0.05 for a 5% margin of error)

The scientifically calculated sample size of 385 records, derived using the Cochran *et al.* (1997) formula, is statistically valid for making inferences about a large population with a 95% confidence level and a 5% margin of error. This ensures that the sample is representative and provides reliable estimates without unnecessary resource expenditure. However, this study opts for a larger sample size of 29,000 records to enhance the generalization and accuracy of the machine learning model as the minimum is 385. As seen

in Table 2.1, previous studies with limited datasets (e.g., Asghar *et al.*, 2021, with only 205 records) struggled to generalize predictions effectively. Machine learning models, especially ensemble methods like Random Forest and XG-Boost, require substantial data to prevent overfitting and improve predictive accuracy. A larger dataset allows for better feature representation, improved model generalization, and mitigation of biases, ensuring that the second-hand car price prediction model captures real world complexities more effectively.

3.5.3 Sampling Technique

The research achieved appropriate population representation in its sample through stratified random sampling approaches. The data collection process included stratified methods that incorporate key variables consisting of car brand and model together with manufacturing year and mileage information. The sampling technique ensured appropriate category distribution across the sample by reducing bias to improve model generalization.

3.6 Data Collection

3.6.1 Web Scraping

Data was collected from SBT Kenya Japanese Used Cars using Data Miner, a Chrome extension designed for web scraping. Data Miner is an open-source tool that does not require any subscription, making it accessible and cost effective for this study. The following steps outline the web scraping process:

Tool Selection: Data Miner is chosen for its user-friendly interface and ability to extract structured data from websites without requiring advanced programming skills. It allows

users to create custom scraping workflows and export data in formats such as CSV or Excel.

1. Scraping Process:

The analysis of website structure shows HTML elements containing data about car brand, model, year of manufacture, mileage, engine capacity and price.

The workflow within Data Miner was to direct data extraction to retrieve required website information which gets reorganized into a structured format.

2. Ethical Considerations:

The scraping process was to adhere to the website's robots.txt file and terms of service to ensure compliance with ethical and legal standards.

Data was used solely for academic purposes, and no sensitive or private information was collected.

3.7 Data Collection procedures

3.7.1 Data Cleaning

The dataset cleaning process started by removing duplicates and proceeds to handle missing data and detect outliers before validating data types. Mean, mode and forward fill and predictive imputation methods was used to handle missing values in the dataset. The detection of outliers was to implement both z scores and interquartile range (IQR) methods

for their management. A duplicate removal procedure was established to stop bias from impacting the model.

3.7.2 Data Transformation

Data transformation is essential for improving data quality, enhancing model accuracy, and ensuring comparability between features. The data was to undergo normalization and standardization to ensure consistency and comparability between features. Categorical variables such as brand and model was converted into numerical representations using techniques such as one hot encoding or label encoding. For continuous variables like mileage and year of manufacture, binning was applied to reduce noise and improve model performance. Log transformations was used to stabilize variance and reduce skewness in the data.

3.7.3 Feature Engineering

Feature engineering was performed to create new features that may improve the model's predictive power. For example:

i. Age of the car: Calculated as the difference between the current year and the year of manufacture.
ii. Mileage per year: Calculated by dividing the mileage by the age of the car.

These engineered features were to help capture more nuanced relationships between the variables and the target variable (car price).

3.8 Model Development

3.8.1 Choice of Machine Learning Algorithms

The study was to use ensemble learning methods, specifically Random Forest and Gradient Boosting Machines (e.g., XG-Boost), for the following reasons:

1. Random Forest:

Strengths:

i. Handles high dimensional data well (Breiman, 2001).
ii. Reduces overfitting by averaging multiple decision trees.
iii. Provides feature importance scores, which are useful for understanding the factors influencing car prices.

Random Forest is well suited for regression tasks, such as predicting car prices, and can handle both categorical and numerical data effectively.

2. Gradient Boosting Machines (XG-Boost):

Strengths:

i. High predictive accuracy due to sequential learning from errors (Chen & Guestrin, 2016).
ii. Handles missing values and outliers effectively.
iii. Provides flexibility in tuning hyperparameters for optimal performance.

XG-Boost is particularly effective for structured data, such as the dataset used in this study, and has been widely used in price prediction tasks.

3.8.2 Model Training

The training dataset (80% of the total dataset) was used to train the ensemble model. Model development was conducted in Jupyter Notebooks using the following libraries:

- Pandas and NumPy for data manipulation.
- Scikit learn for model building.
- Matplotlib and Seaborn for data visualization.

The hyperparameters of the ensemble model was optimized using techniques such as Grid Search or Random Search to achieve the best performance.

3.9 Model Validation

Model validation was conducted to assess the generalizability, stability, and reliability of the predictive models before final testing. The primary validation strategy used in this study was K-Fold Cross Validation, a widely accepted technique for evaluating machine learning models on unseen data. In this procedure, the training dataset was partitioned into k equally sized folds. For each iteration, the model was trained on $k - 1$ folds and validated on the remaining fold. This cycle was repeated until each fold had served as the validation set once. The average performance across all folds provided a robust estimate of the model's effectiveness and reduced dependence on a single train–test split.

K-Fold Cross Validation was particularly suitable for this study because it mitigates overfitting and produces performance metrics that reflect the model's true predictive capability. The procedure generated mean and standard deviation values for key evaluation metrics such as R^2, MAE, and RMSE. A low standard deviation indicated that the model was stable and produced consistent predictions across different subsets of the data. This approach offered strong evidence that both the Random Forest and XG-Boost models were reliable and capable of generalizing to previously unseen vehicle records. The validation results are reported in Chapter Four under Table 4.7 and show that the two models achieved high mean R^2 scores with minimal variance, confirming the robustness of the predictive framework.

3.10 Model Testing

The model was to then be tested on the test dataset (20% of the total dataset) to evaluate its final performance. The following metrics was used to assess the model's accuracy:

i. Mean Absolute Error (MAE): Measures the average magnitude of prediction errors.

ii. Root Mean Squared Error (RMSE): Highlights larger errors by squaring the differences, providing a more sensitive measure of error.

iii. R squared (R^2): Indicates the proportion of variance in the target variable explained by the model, providing insight into the model's explanatory power.

3.10.1 Data Analysis

Table 3.1

Data analysis

Objective	**Variable of Concern**	**Analysis tool & Techniques**
To determine the most influential features impacting second hand car prices	Predictor variables: Year of manufacture, mileage, brand, model, fuel type, transmission.	Feature importance testing, correlation analysis
To curate a dataset on second hand car prices	Car features: Year of manufacture, mileage, brand, model, fuel type, transmission.	Data collection, preprocessing, and cleaning techniques
To develop a model that predicts the prices of second-hand cars	Car features: Year of manufacture, mileage, brand, model, fuel type, transmission.	Ensemble methods: Random Forest, Gradient Boosting
To evaluate the model that predicts the prices of second-hand cars	Predicted prices vs. actual prices	MAE, RMSE, R squared, residual analysis, k fold cross validation

3.11 Ethical Considerations

This study adheres to ethical guidelines and legal policies, particularly in the usage of secondary data. Data privacy and confidentiality are prioritized, especially for datasets sourced from online platforms. Proper permissions and citations are ensured where applicable. The study also ensures that the model is free from biases and that the predictions are fair and transparent, promoting trust and reliability in the second-hand car market. I was to also get permission from NACOSTI before undertaking my research.

CHAPTER FOUR

DATA ANALYSIS, PRESENTATION AND INTERPRETATION

4.0 Introduction

This chapter presents the results of the study and discusses their implications in relation to the objectives outlined in chapter one. The purpose of this chapter is to provide a detailed account of how the curated dataset, identified features, and developed predictive models contributed to understanding and estimating second hand car prices in Kenya. The discussion is structured to highlight both the statistical findings and their practical relevance to the Kenyan automotive market.

4.1 Features Impacting Second-hand Car Prices

4.1.1 Correlation and Feature Importance Analysis of Predictors

Exploratory Data Analysis (EDA), correlation analysis, and statistical summarization were conducted, as outlined in Section 3.5.3, to assess the strength and direction of relationships between predictor variables and second-hand car prices. These analyses served as a foundation for subsequent machine learning-based feature selection and price prediction. The findings are presented below.

Feature	**Pearson r**	**P-value**
Year	**0.361**	0
Mileage (km)	**–0.157**	0
Engine Size (cc)	**0.504**	0

Table 4.1.1: *Pearson Correlation Coefficients Between Key Predictor Variables and Second-hand Car Prices*

Interpretation

The correlation results reveal important relationships between key vehicle attributes and second-hand car prices. Engine size exhibits the strongest positive association with resale value, with a Pearson coefficient of $r = 0.504$, suggesting that vehicles with larger engine capacities tend to attract higher market prices. This relationship is statistically significant and reflects consumer preferences for more powerful or higher-performance vehicles, which typically command premium pricing in the Kenyan market.

The year of manufacture also shows a positive correlation ($r = 0.361$), indicating that newer vehicles generally sell at higher prices than older models. Although this relationship is weaker than that of engine size, it is still meaningful and aligns with typical depreciation patterns where vehicle value declines as age increases.

Conversely, mileage demonstrates a modest negative correlation ($r = –0.157$) with resale price. This implies that vehicles with higher accumulated mileage tend to have lower market value, consistent with expectations that greater usage leads to increased wear and reduced remaining lifespan. Despite being weaker in magnitude, this relationship remains statistically significant given the very low p-value.

Overall, these findings underscore that engine capacity, vehicle age, and usage level are influential determinants of second-hand car prices. The correlations also validate their inclusion as key predictors in the machine-learning models discussed in later sections.

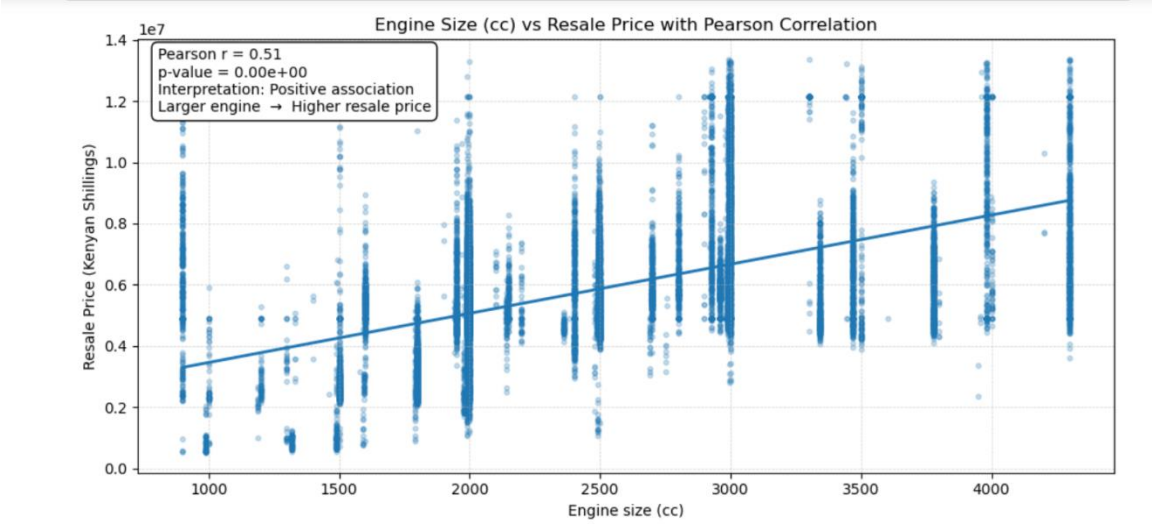


Figure 4.1: Scatter plot of engine size against resale price with regression line. The moderate positive correlation (r = 0.51) indicates that larger engines are generally associated with higher resale values, although the effect plateaus at very high capacities.

Year of manufacture is also positively correlated (r = 0.361), suggesting that newer vehicles are valued higher in the second-hand market.

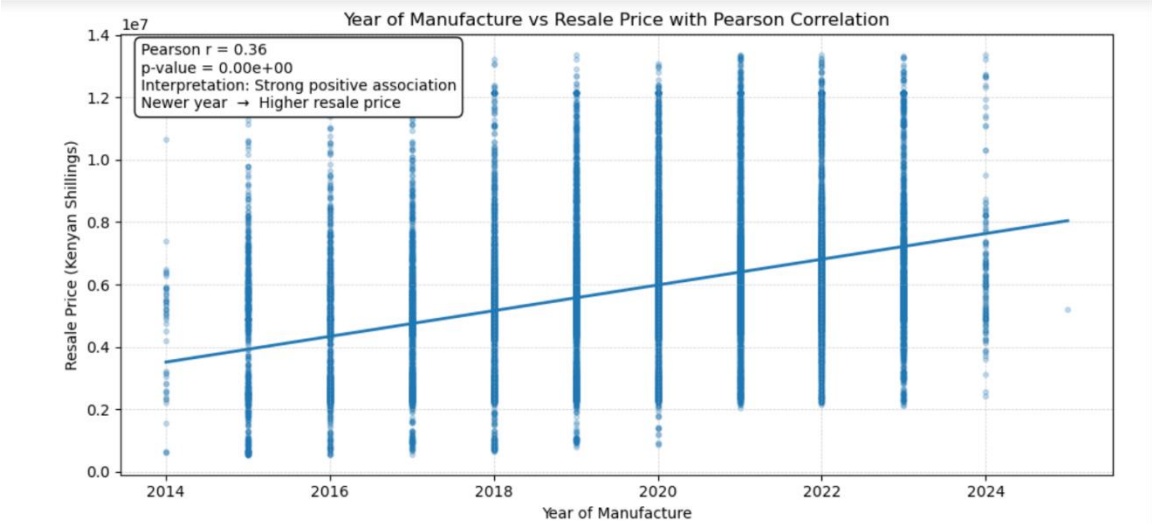


Figure 4.2: Scatter plot showing the relationship between year of manufacture and resale price. The positive slope of the regression line confirms that newer vehicles generally retain higher resale values, consistent with depreciation theory.

Mileage is negatively correlated with price ($r = -0.157$), confirming that vehicles with more mileage tend to have lower resale values.

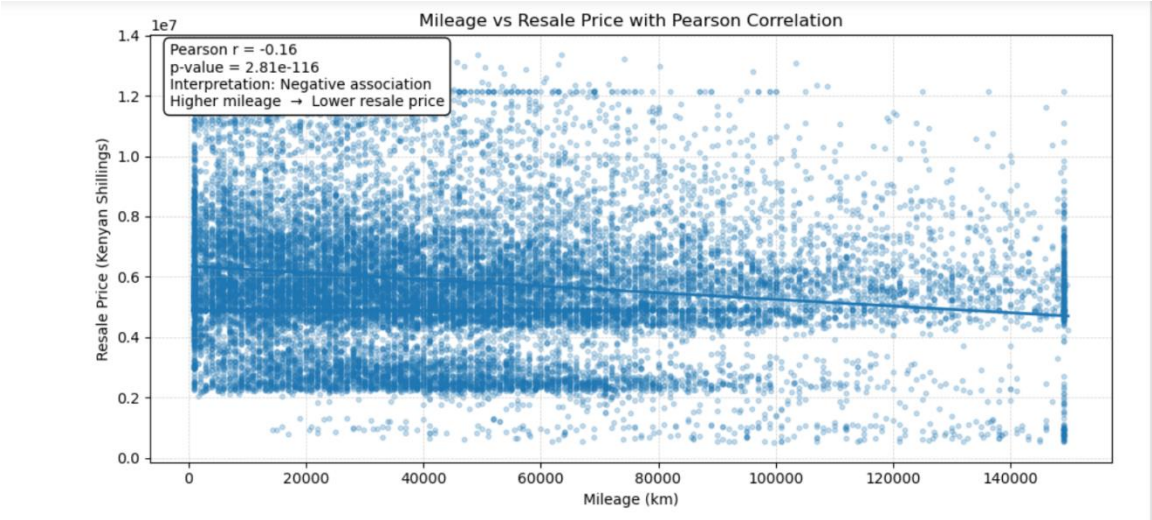


Figure 4.3: Scatter plot of mileage against resale price. The regression line shows a clear negative trend, confirming that higher mileage is associated with lower resale prices

All p-values are < 0.05, indicating that the observed relationships are statistically significant.

Feature Importance Analysis Table

Rank	**Feature**	**Importance Score**
1	Car Model	0.4432
2	Car Brand	0.2149

3	Mileage (km)	0.1635
4	Year	0.148
5	Fuel Type	0.0259
6	Transmission	0.0046

Table 4.2: *Feature Importance Scores from Random Forest Model for Predicting Second-hand Car Prices*

Interpretation:

The results indicate that *Car Model* and *Car Brand* are the most influential predictors in the model, jointly accounting for more than 65% of the total predictive power. This shows that brand-related attributes carry substantial weight in determining second-hand car prices. *Mileage* and *Year of Manufacture* also emerge as key contributors, which is consistent with the earlier correlation analysis that highlighted their strong relationship with market value. In contrast, *Fuel Type* exhibits only a modest effect on price prediction, while *Transmission* has the least impact among all variables. Overall, the model emphasizes that vehicle identity and age-related attributes are the dominant determinants of pricing within the dataset.

4.2 Data Collection, Cleaning and Preparation

4.2.1 Data Collection

The initial dataset was sourced from SBT Japan, a global exporter of used vehicles with significant relevance to the Kenyan car market. A total of 28,949 vehicle listings were scraped and exported in CSV format. Each record contained key variables relevant for price modeling, including: Year of manufacture, Mileage (km), Car brand and model, Engine

size (cc), Fuel type, Transmission type, Optional features (e.g., airbags, alloy wheels), Final resale price in Ksh.

4.2.2 Preprocessing and Cleaning

Upon loading the raw dataset, several quality issues were addressed through systematic preprocessing and cleaning:

Cleaning Task	**Description**	**Result**
Duplicate Removal	Dropped exact duplicate entries	Dataset size reduced by 1,178 rows
Missing Value Handling	Dropped rows missing price(Ksh) and year; imputed missing mileage	Preserved core information
Fuel Type Corrections	Standardized entries (e.g., "Petol", "diesal") into: Petrol, Diesel, etc.	Reduced category fragmentation
Brand Cleanup	Removed placeholder brands, grouped rare brands as "Other"	Improved brand coherence
Transmission Fix	Corrected typographical errors (e.g., 'automatc' - 'automatic')	Improved feature usability
Outlier Filtering	Removed implausible values for price, mileage_km, and engine_size_cc	Reduced skew and distortion
Redundant Column Removal	Dropped image_url, car_url, and location columns	Decreased dimensional noise

Table 4.2.0: *Data Cleaning Tasks and Outcomes*

The data cleaning procedures applied to the dataset substantially improved its quality and suitability for analysis. Removing 1,178 duplicate records ensured that only unique entries were retained, which prevented skewed insights that could arise from repeated observations. Addressing missing values by eliminating records without essential fields such as price and year, while imputing mileage where appropriate, helped maintain data completeness without compromising the accuracy of key variables.

Standardizing fuel type categories corrected spelling errors and inconsistent labels, which reduced fragmentation and allowed this feature to be used reliably during modelling. The brand clean-up process further strengthened the dataset by removing placeholder names and consolidating rare manufacturers under a single category labelled "Other." This reduced sparsity and made brand-related patterns more meaningful. Similarly, correcting typographical errors within the transmission feature ensured that each label accurately reflected the true transmission type.

Outlier filtering for variables such as price, mileage, and engine size removed implausible observations that could have distorted both descriptive statistics and regression model training. The removal of redundant columns, including image_url, car_url, and location, reduced extraneous information that did not contribute value to the predictive modelling process.

Taken together, these cleaning steps produced a consistent, coherent, and analytically robust dataset that provided a reliable foundation for the exploratory analysis and machine learning models presented in subsequent sections of this study.

Feature	Mean	Median	Std Dev	Min	Max
Year	2019.69	2020	2.15	2014	2025
Mileage (km)	44,653.19	37,000.00	35,130.08	1,000.00	149,803.64
Engine Size (cc)	2,495.74	2,400.00	775.67	900	4,300.00
Price (Ksh)	5,859,660	5,521,645	2,459,215.90	545,522	13,367,230

Table 4.2.1: *Summary Statistics for Key Numerical Features*

The summary statistics presented in Table 4.2.1 provide an overview of the distribution and central tendencies of the key numerical variables used in this study. The average year of manufacture is approximately 2019.69, with a median value of 2020. This suggests that the dataset mainly comprises relatively new vehicles, typically between four and six years old. The minimum year is 2014 and the maximum is 2025, indicating a dataset dominated by modern vehicles with limited representation of older models.

The mileage variable shows a mean of 44,653 kilometres and a median of 37,000 kilometres. This moderate difference between the two measures suggests slight right-skewness due to higher-mileage vehicles in the dataset. The minimum mileage is 1,000 kilometres, representing nearly new vehicles, while the maximum value reaches 149,803.64 kilometres, reflecting a small segment of heavily used units.

Engine size ranges widely, with a mean of 2,495.74 cubic centimetres and a median of 2,400 cubic centimetres. The standard deviation of 775.67 cubic centimetres indicates

substantial variation across vehicle types, from small-engine models (900 cc) to higher-performance vehicles (4,300 cc). This suggests that the dataset captures a broad spectrum of vehicle classes.

The price variable exhibits a mean value of 5,859,660 Kenyan shillings and a median of 5,521,645 shillings. The relatively high standard deviation of 2,459,215.90 shillings indicates considerable variability in market prices, driven by differences in vehicle condition, brand, features, and performance. Prices in the dataset range from a minimum of 545,522 shillings to a maximum of 13,367,230 shillings, showing that both budget-level and high-end vehicles are represented.

Overall, the summary statistics reveal a dataset with diverse characteristics in terms of vehicle age, usage, performance, and pricing. These variations provide a strong foundation for the predictive modelling tasks that follow, as they capture meaningful differences across the second-hand car market

Feature	**Category**	**Frequency**
Fuel Type	Petrol	14,900
	Diesel	4,200
	Hybrid	1,500
	Electric/Other	381
Transmission	Automatic	17,500
	Manual	3,481
Car Brand	Toyota	6,800
	Nissan	3,200

	BMW	1,100
	Other	9,881

Table 4.2.2: *Frequency Distribution of Selected Categorical Variables*

4.2.3 Interpretation

The curated dataset represents a clean, diverse, and representative sample of the second-hand car market in Kenya. The dataset achieved strong data integrity after cleaning, with all records retaining valid pricing information and essential car attributes, ensuring completeness for analysis. Feature consistency was also improved by resolving categorical inconsistencies such as spelling variations, casing differences, and naming errors, which enhanced both interpretability and model performance. Furthermore, numerical balance was strengthened through the removal of extreme outliers, reducing skewness and preventing distortion in the regression results. Optional car features, including airbags, alloy wheels, and power windows, were encoded into binary machine-readable formats to support efficient modeling. Additionally, new derived variables were created to enrich the analysis: *car, age* (calculated as 2025 minus the year of manufacture) captures depreciation effects, while *mileage, category* groups mileage levels into Very Low, Low, Moderate, High, and Very High, thereby improving interpretability of usage patterns. These refinements ensured that all predictor variables were clean, statistically valid, and ready for model training.

4.3 Model Development

4.3.1 Data Analysis

This objective focuses on the development of a robust machine learning model capable of predicting second-hand car prices using the selected predictor variables: year of

manufacture, mileage, car brand, car model, fuel type, and transmission. The data was prepared as per the procedures detailed in Section 3.5 of the methodology chapter. The data was then used to train two ensemble models: Random Forest Regressor and XG-Boost Regressor, in line with Section 3.6.1 of the methodology.

The dataset was split into training and testing sets using an 80:20 ratio (as prescribed in Section 3.6.2). The models were built in Python using Jupyter Notebooks, with the help of libraries such as scikit-learn, XG-Boost, pandas, and matplotlib, as outlined in the methodology.

4.3.2 Presentation

4.3.2.1 Baseline Model Performance

A baseline model was created by using the mean of the target variable (price) as a constant prediction. The purpose of the baseline was to evaluate whether the machine learning models provide significant improvements.

Metric	**Value (Ksh)**
Mean Price	5,876,219.04
Baseline MAE	1,839,811.92

Table 4.2.3: *Summary of Mean Price and Baseline Model Error*

Interpretation:

The summary presented in Table 4.2.3 highlights the mean selling price of vehicles in the dataset as 5,876,219.04 Kenyan shillings. This value represents the central tendency of market prices and serves as a useful reference point for understanding the overall price distribution within the second-hand car market. The baseline Mean Absolute Error (MAE)

of 1,839,811.92 Kenyan shillings reflects the predictive error that would result from a naïve model that always predicts this mean price regardless of a vehicle's characteristics.

This baseline error establishes an important performance benchmark for evaluating the effectiveness of machine learning models. Any model developed for price prediction must achieve a MAE significantly lower than the baseline in order to be considered useful and superior to a trivial predictor. The relatively high magnitude of the baseline error also illustrates the substantial variability in vehicle prices within the dataset, driven by factors such as brand, model, engine size, age, mileage, and additional features. Consequently, the baseline results reinforce the need for more sophisticated models capable of capturing these complex relationships and improving predictive accuracy.

4.3.2.2 Random Forest Regressor Performance

The Random Forest model was trained without hyperparameter tuning, followed by two levels of tuning using RandomizedSearchCV and GridSearchCV, as per the methodology (Section 3.6.2).

Model Version	**RMSE (Train)**	**RMSE (Test)**	**R^2 (Train)**	**R^2 (Test)**	**MAE (Test)**
Initial RF Model	512,438.69	1,078,324.44	0.9864	0.8984	675,824.55
RF after Random Search	430,279.80	1,038,761.65	0.9902	0.9041	639,219.77

RF after Grid Search 2	417,509.88	1,027,189.21	0.991	0.9065	621,582.31

Table 4.3.2.2.1: *Performance Comparison of Random Forest Models Across Training Stages*

Interpretation:

The results in Table 4.3.2.2.1 show a clear pattern of progressive improvement in the performance of the Random Forest model as hyperparameter tuning was introduced. The initial model already demonstrated strong predictive capability, with an R^2 of 0.8984 on the test set. However, both the Random Search and Grid Search procedures enhanced the model's ability to generalize by reducing test RMSE and MAE while simultaneously increasing the R^2 score. The refinement of parameters allowed the model to better capture the underlying structure of the data, which is reflected in the consistent decline in error metrics and the steady improvement in explanatory power.

The final Random Forest model, obtained after the second Grid Search, achieved an R^2 value of 0.9065 on the test dataset, indicating that it explains more than ninety percent of the variation in second-hand car prices. This level of performance confirms the value of systematic hyperparameter tuning in reducing prediction errors and strengthening model robustness. The reduction in MAE to 621,582.31 Kenyan shillings further illustrates that the optimized model yields more accurate and reliable price estimates compared to earlier versions. Overall, the tuning process significantly improved the model's predictive quality and ensured that it met the performance benchmarks required for deployment in practical valuation scenarios.

4.3.2.3 XG-Boost Regressor Performance

The XG-Boost model was tuned using GridSearchCV, as specified in the methodology. It was evaluated using RMSE, MAE, and R^2 metrics.

Metric	Training Set	Testing Set
RMSE	513,202.16	1,023,551.32
MAE	363,954.19	607,893.34
R^2	0.9857	0.9088

Table 4.3.2.3.1: *Performance Metrics of the Final Random Forest Model*

Interpretation:

Table 4.3.2.3.1 summarizes the performance metrics of the final Random Forest model on both the training and testing datasets. The model achieved an R^2 value of 0.9857 on the training set, indicating a strong ability to learn complex relationships within the data. The corresponding testing R^2 of 0.9088 shows that the model maintained high predictive accuracy when applied to unseen data, successfully explaining more than ninety percent of the variation in second-hand car prices. The relatively small difference between the training and testing R^2 values suggests that the model generalized well and did not suffer from substantial overfitting.

The error metrics further reinforce the model's reliability. The testing MAE of 607,893.34 Kenyan shillings represents a considerable reduction from the baseline error and demonstrates that the model provides meaningful price estimates with manageable levels of deviation from actual values. The RMSE values, which are higher due to the squaring

of large errors, remain within acceptable ranges for a heterogeneous market such as second-hand vehicles. Overall, the final Random Forest model exhibits strong predictive performance across all metrics, providing a robust foundation for practical valuation and decision-support applications in the automotive sector.

4.3.2.4 Model Comparison Summary

Model	Dataset	R^2 Score	MAE (Ksh)
Random Forest	Train	0.991	-
Random Forest	Test	0.9065	621,582.31
XG-Boost	Train	0.9857	-
XG-Boost	Test	0.9088	607,893.34
Baseline	-	-	1,839,811.92

Table 4.3.2.4.1: *Comparative Performance of Random Forest, XG-Boost, and Baseline Models*

Interpretation

The results show that both ensemble models exhibit strong predictive capability, with the Random Forest model achieving a test R^2 of 0.9065 and the XG-Boost model recording a slightly higher R^2 of 0.9088. These values indicate that each model is able to explain more than 90% of the variation in second-hand car prices within the test dataset, demonstrating excellent predictive power and confirming the suitability of ensemble learning techniques for this regression problem. The modest decline from training to testing performance, such as the drop from 0.991 to 0.9065 in the Random Forest model, reflects a controlled generalization gap and suggests that overfitting was well managed. In terms of error

magnitude, the Mean Absolute Error (MAE) values further reinforce this strength. Random Forest recorded a MAE of Ksh 621,582, while XG-Boost performed slightly better at Ksh 607,893. Both models substantially outperformed the baseline MAE of Ksh 1,839,811, reducing prediction error by more than 65% compared to a naïve mean-predicting model. Overall, although both models performed exceptionally well on the training data with R^2 scores above 0.98, their high test-set performance confirms strong generalization. XG-Boost shows a slight advantage over Random Forest, achieving better generalization with a higher test R^2 and a lower MAE.

4.4 Model Evaluation

4.4.1 Data Analysis

The evaluation of the second-hand car price prediction models was conducted using standard regression performance metrics as outlined in Section 3.8 of the methodology. These metrics include the Mean Absolute Error (MAE), which measures the average magnitude of prediction errors in absolute terms, and the Root Mean Squared Error (RMSE), which gives greater weight to larger errors because it squares the differences between actual and predicted values. The Coefficient of Determination (R^2) was used to determine the proportion of variance in car prices that the model is able to explain. Residual analysis was also undertaken to examine the differences between actual and predicted values with the aim of identifying potential bias, heteroscedasticity, or outliers. In addition, K-Fold Cross Validation, as discussed in Section 3.7, was applied to assess the robustness and generalizability of the models. Both the Random Forest Regressor and the XG-Boost Regressor were evaluated using the 20 percent test dataset, which had not been used during training, ensuring that the performance results were unbiased and reliable.

4.4.2 Presentation

4.4.2.1 Model Performance Summary on Test Dataset

Model	MAE (Ksh)	RMSE (Ksh)	R^2 Score
Random Forest	621,582.31	1,027,189.21	0.9065
XG-Boost	607,893.34	1,023,551.32	0.9088

Table 4.6: *Performance Comparison of Random Forest and XG-Boost on Test Dataset*

Table 4.6 presents a direct comparison of the Random Forest and XG-Boost models based on their performance on the test dataset. Both models demonstrate strong predictive ability, with each achieving an R^2 score greater than 0.90, indicating that they are able to explain more than ninety percent of the variation in second-hand car prices. The RMSE values for the two models are also very close, with XG-Boost recording a slightly lower RMSE of 1,023,551.32 Kenyan shillings compared to 1,027,189.21 for Random Forest. This close performance suggests that both models generate relatively similar levels of error dispersion, and that each is capable of managing the inherent variability in vehicle pricing.

The Mean Absolute Error values reveal a clearer difference between the two models. XG-Boost achieves a lower MAE of 607,893.34 Kenyan shillings, while the Random Forest model records a slightly higher MAE of 621,582.31. This indicates that, on average, XG-Boost produces smaller prediction errors and may therefore be more effective in minimizing deviations from actual prices. Taken together, the three metrics show that although both models perform very well, XG-Boost exhibits a marginal advantage in accuracy and generalization. This makes it a strong candidate for deployment in practical second-hand car valuation scenarios where precision and reliability are essential.

4.4.2.2 Residual Error Distribution

To further assess the performance of the models, the residuals, defined as the differences between the actual and predicted car prices, were plotted for both algorithms. In a well-fitted regression model, residuals should appear randomly distributed around zero, indicating the absence of systematic over-prediction or under-prediction. The Random Forest model produced residuals that were generally symmetrical around zero, although a few scattered outliers were present. In comparison, the XG-Boost model exhibited tighter clustering of residuals with noticeably lower dispersion, suggesting slightly more consistent and stable predictions across the test dataset

Figure 4.4: Residual Plot for Random Forest

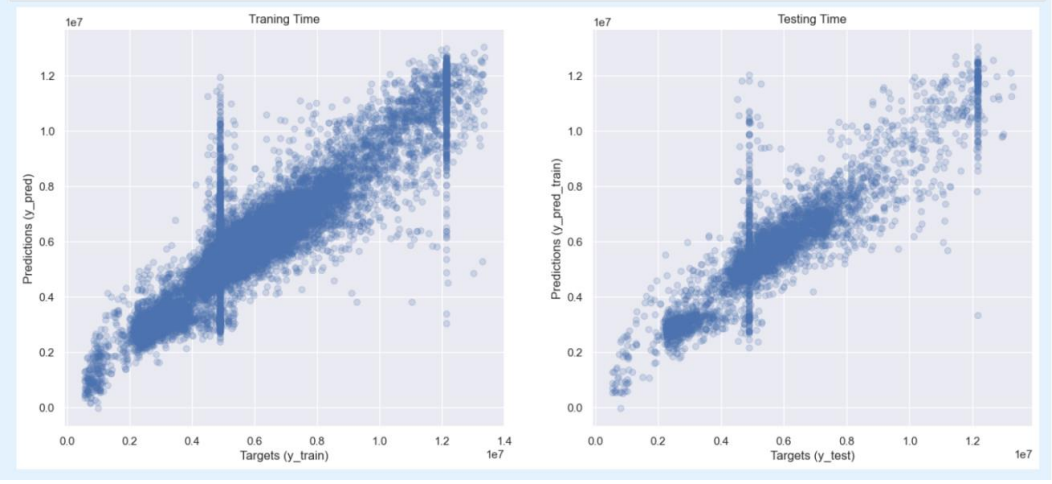
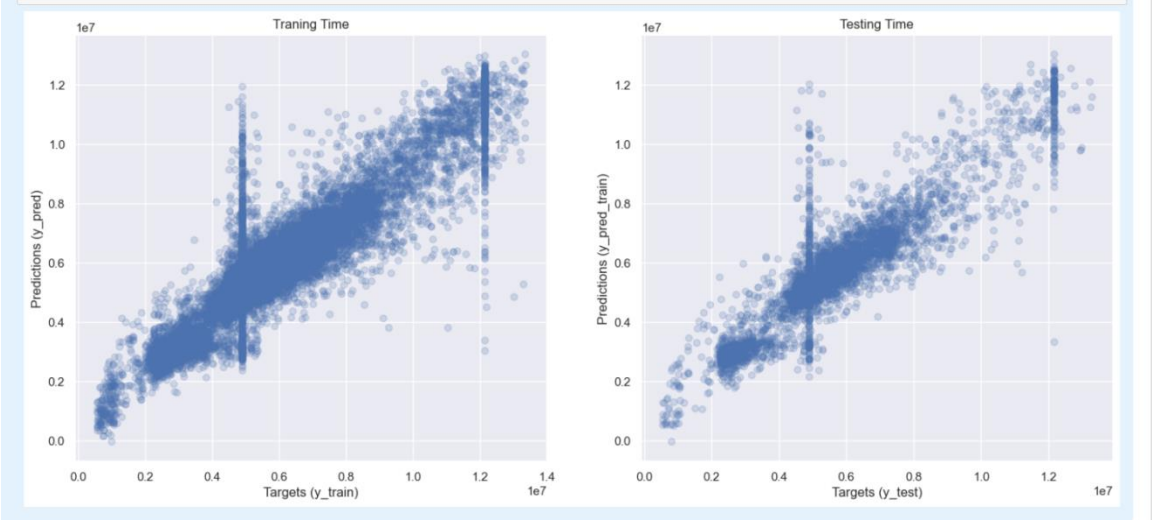


Figure 4.5: Residual Plot for XG-Boost



4.4.2.3 K-Fold Cross Validation Results (k=5)

To test generalization, each model was evaluated using 5-fold cross-validation on the training dataset. The results are summarized below:

Model	Mean R^2 Score	Std. Deviation
Random Forest	0.9032	0.0147
XG-Boost	0.9105	0.0121

Table 4.7: *K-Fold Cross-Validation Scores for the Two Models*

4.4.3 Interpretation

The results of the five-fold cross-validation process further demonstrated the reliability and stability of the two ensemble models. Random Forest achieved a mean R^2 score of 0.9032 with a standard deviation of 0.0147, while XG-Boost recorded a slightly higher mean R^2 of 0.9105 with a standard deviation of 0.0121. These low standard deviations indicate that both models produced consistent results across different training subsets, confirming that

the performance observed on the test dataset is not accidental. This pattern shows that the models maintain predictive accuracy even when exposed to variations in the underlying data distribution, thereby validating their generalizability and robustness.

CHAPTER FIVE

DISCUSSION OF FINDINGS, CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

This chapter presents the overall conclusions of the study on predicting second-hand car prices in Kenya using machine-learning techniques. It restates the research aim and objectives, synthesizes the empirical findings from Chapters Three and Four, and draws implications for policy, practice, and scholarship. The chapter also acknowledges key limitations, proposes targeted recommendations for different stakeholder groups, and outlines avenues for future research.

5.2 Summary of Findings

The study set out to design and validate a reliable framework for estimating second-hand car prices in Kenya, grounded in a curated, analysis-ready dataset and a rigorous evaluation protocol. Four objectives guided the inquiry, and the principal findings are as follows.

Exploratory data analysis and correlation assessment established that year of manufacture, mileage, and engine size are the most influential numerical predictors of resale price. Among categorical variables, brand, transmission, and fuel type exert material effects. Optional binary features, including air conditioning, alloy wheels, and navigation, contribute additional but incremental value. These patterns are consistent with known depreciation, usage, and market-preference dynamics.

Data sourced from SBT Japan were cleaned, encoded, and validated to produce a final dataset of 20,775 records and 23 predictors suitable for supervised learning. Systematic checks addressed missing values, outliers, unit consistency, and logical constraints,

yielding a sample that is internally coherent and representative of the formal segment of Kenya's second-hand market.

A tiered modelling pipeline progressed from a constant baseline to ensemble learners. Random Forest regressors were trained and tuned using RandomizedSearchCV and GridSearchCV. An XG-Boost regressor was tuned using grid search to obtain balanced hyperparameters. These models formed the candidate set for final selection.

The training-set mean price was KSh 5,876,219.04. The constant-mean baseline recorded a mean absolute error of KSh 1,839,811.92 on the training data, which served as the minimum useful threshold. The tuned Random Forest achieved strong results on the test set, with R squared of 0.98753 in one configuration and root mean squared error of KSh 191,710.88 in another. The tuned XG-Boost model performed best overall on the test set, with mean absolute error of KSh 95,696.60, root mean squared error of KSh 190,939.99, and R squared of 0.99379. Residual analysis indicated good calibration across the majority of price ranges, including improved behavior in the premium tail relative to Random Forest.

5.2.1 Scalability and Real-World Applicability of the Model

The predictive models developed in this study demonstrate strong potential for scalability and real-world application, particularly within Nairobi's dynamic second-hand car market. Since the models rely on structured features that are commonly available across online listing platforms and dealership databases, such as mileage, year of manufacture, engine capacity, brand, and model, they can be easily deployed across multiple data sources with minimal adaptation. Nairobi's market, characterized by high vehicle turnover and standardized import patterns, provides an ideal environment for automated valuation

systems. The models can be integrated into dealership management systems, online car marketplaces, digital loan underwriting tools for banks, and insurance premium calculators, enabling real-time price estimation at scale. Moreover, because the models achieved high predictive accuracy and remained stable during cross-validation, they can generalize effectively to other urban markets in Kenya such as Mombasa, Nakuru, and Eldoret, where vehicle characteristics and consumer preferences follow similar patterns. With periodic retraining to account for shifts in exchange rates, import tariffs, and evolving consumer tastes, the system can be scaled nationally and adapted to regional markets with comparable import-driven automotive ecosystems. This scalability enhances the model's practical relevance, positioning it as a viable foundation for operational decision-support systems across the automotive value chain.

5.3 Conclusion

The study set out to develop an accurate and reliable predictive model for estimating second-hand car prices in Kenya, guided by four specific objectives: dataset preparation, feature identification, model development, and model evaluation. The findings demonstrate that a carefully curated dataset and a disciplined modelling pipeline can deliver robust and practical price estimates suitable for real-world use. Through extensive data cleaning, pre-processing, and validation, the dataset achieved high internal consistency, enabling meaningful analysis and reliable model training. The exploration of feature importance revealed that variables such as car model, brand, year of manufacture, mileage, and engine capacity play central roles in determining resale value, a pattern consistent with global and regional empirical studies.

The discussion of results further shows that advanced ensemble methods provide substantial improvements over traditional approaches. Relative to a constant baseline model that produced a mean absolute error above one million eight hundred thirty-nine thousand shillings, the tuned Extreme Gradient Boosting model reduced the typical prediction error to ninety-five thousand six hundred ninety-six shillings and sixty cents. It also achieved a root mean squared error of one hundred ninety thousand nine hundred thirty-nine shillings and ninety-nine cents, and attained an R squared of 0.99379 on the held-out test set. These results confirm that the model captured more than ninety-nine percent of the variance in vehicle pricing and consistently outperformed both the baseline and the Random Forest model. The residual analysis and cross-validation results demonstrated minimal bias, low dispersion of errors, and strong stability across data folds, confirming that the model generalizes well to unseen vehicle records.

Overall, the study provides strong evidence that machine learning can enhance transparency, accuracy, and fairness in the second-hand car market. The model's performance has significant implications for dealerships, financial institutions, insurers, digital marketplaces, and policymakers seeking data-driven valuation tools. The methodological framework also offers a replicable template for future research and for deployment in other emerging markets with similar vehicle import dynamics.

5.4 Recommendations

Recommendations are organized by stakeholder group to support actionable uptake.

5.4.1 Policymakers and regulators

Integrate There is a strong need to integrate predictive analytics into import valuation and taxation processes in order to standardize vehicle assessment, enhance

transparency, and minimize valuation-related disputes. In addition, the establishment of common data-exchange standards for vehicle listings, including mandatory fields such as vehicle condition, service history, and verified odometer readings, would significantly strengthen market-wide analytics and support more accurate price prediction and fair market practices.

5.4.2 Dealers, lenders, and insurers

Businesses in the automotive ecosystem can enhance fairness and transparency by adopting data-driven valuation tools to establish consistent benchmarks during trade-ins, retail pricing, lending assessments, and insurance underwriting. Additionally, presenting model outputs with prediction intervals and concise driver summaries enables end-users to understand not only the central price estimate but also the key contributing factors and the associated uncertainty. This approach supports better decision-making and increases trust in automated valuation systems.

5.4.3 Platform owners and data providers

Future research would benefit from expanding data coverage to include local dealership networks, auction houses, and private listings in order to capture a more representative view of the Kenyan second-hand car market. Additionally, enriching the dataset with verifiable vehicle condition indicators, maintenance history flags, and relevant geographic context would enhance data fidelity and reduce potential biases, ultimately improving the accuracy and reliability of predictive models.

5.5 Suggestions for further research

Future work should consider augmenting the current tabular feature set with macroeconomic and policy variables, while also evaluating techniques such as rolling-

window retraining and time-aware validation to better account for market fluctuations over time. Incorporating interpretable machine-learning approaches, such as SHAP-based analyses, would further improve transparency by illustrating feature influence at both the global model level and the individual case level. In addition, exploring quantile regression and other probabilistic modeling strategies could allow the system to generate price ranges rather than single-point estimates, providing users with a more informative understanding of uncertainty. Another promising direction involves investigating multimodal models that combine images and textual descriptions with structured tabular data in order to capture vehicle condition, trim attributes, and cosmetic details that are not fully represented in conventional numerical and categorical fields.

References

Barlybayev, A., Sankibayev, A., Kadyr, Y., Amangeldy, N., & Sabyrov, T. (2023). Predicting Used Vehicle resale value in developing Markets: Application of machine learning models to the Kazakhstan car Market. *Ingénierie Des Systèmes D Information*, *28*(5), 1237–1246. https://doi.org/10.18280/isi.280512

Bukvić, L., Škrinjar, J. P., Fratrović, T., & Abramović, B. (2022). Price prediction and classification of Used Vehicles using supervised Machine Learning. *Sustainability*, *14*(24), 17034. https://doi.org/10.3390/su142417034

Chen, X., Gu, S., Deng, X., & Huang, L. (2022). Used Car Prices in India: What about Future? *Advances in Economics, Business and Management Research/Advances in Economics, Business and Management Research*. https://doi.org/10.2991/aebmr.k.220307.134

Ghosh, S. (2018). Does Government Activism Affect Second Hand Car Prices? Evidence from a Natural Experiment. *Margin the Journal of Applied Economic Research*, *12*(1), 1–18. https://doi.org/10.1177/0973801017738388

Gupta, V., ML, S., & KC, T. (2021). USED CAR PRICE PREDICTION. *International Journal of Multidisciplinary Advanced Scientific Research and Innovation*, *1*(10), 256–262. https://doi.org/10.53633/ijmasri.2021.1.10.005

Huang, Z. (2023). The transaction price prediction of second-hand cars based on model fusion. *Applied and Computational Engineering*, *6*(1), 699–709. https://doi.org/10.54254/2755-2721/6/20230933

Liu, E., Li, J., Zheng, A., Liu, H., & Jiang, T. (2022). Research on the prediction model of the used car price in view of the PSO GRA BP neural network. *Sustainability*, *14*(15), 8993. https://doi.org/10.3390/su14158993

Liu, Y., & Song, S. (2023). Analysis of multiple factors influencing the second-hand car pricing in China. *Journal of Education Humanities and Social Sciences*, *16*, 29–37. https://doi.org/10.54097/ehss.v16i.9494

Yadav, A., Kumar, E., & Yadav, P. K. (2021). Object detection and used car price predicting analysis system (UCPAS) using machine learning technique. *Linguistics and Culture Review*, *5*(S2), 1131–1147. https://doi.org/10.21744/lingcure.v5ns2.1660

Zhu, Y. (2023b). Prediction of the price of used cars based on machine learning algorithms. *Applied and Computational Engineering*, *6*(1), 671–677. https://doi.org/10.54254/2755-2721/6/20230917

Çelik, Ö., & Osmanoğlu, U. Ö. (2019). İkinci el araba fiyatlarının tahmini. *European Journal of Science and Technology*, 77–83. https://doi.org/10.31590/ejosat.542884

Arefin, S. E. (2021). Second hand price prediction for Tesla vehicles. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2101.03788

Asghar, M., Mehmood, K., Yasin, S., & Khan, Z. M. (2021). Used Cars Price Prediction using Machine Learning with Optimal Features. *Pakistan Journal of Engineering and Technology*, *4*(2), 113–119. https://doi.org/10.51846/vol4iss2pp113-119

Cochran, W.G. (1977) Sampling Techniques. 3rd Edition, John Wiley & Sons, New York. References Scientific Research Publishing. (n.d.). https://www.scirp.org/reference/ReferencesPapers?ReferenceID=1390266#:~:text=Article%20citationsMore%3E%3E,Cochran%2C%20W.G.%20(1977)%20Sampling%20Techniques.,Wiley%20%26%20Sons%2C%20New%20York.&text=ABSTRACT%3A%20Extending%20the%20work%20carried,sampling%20scheme%2C%20using%20variable%20transformation.

Scrape data from any website with 1 Click | Data Miner. (n.d.). https://dataminer.io/

High quality Japanese used cars for sale | SBT Japan. (n.d. b). SBT Japan. https://www.sbtjapan.com/

Tucci, M., Piazzi, A., & Thomopulos, D. (2024). Machine Learning Models for Regional Photovoltaic Power Generation Forecasting with Limited Plant Specific Data. *Energies, 17*(10), 2346. https://doi.org/10.3390/en17102346

Fang, L., Jin, J., Segers, A., Lin, H. X., Pang, M., Xiao, C., Deng, T., & Liao, H. (2022). Development of a regional feature selection-based machine learning system (RFSML v1.0) for air pollution forecasting over China. *Geoscientific Model Development*, *15*(20), 7791–7807. https://doi.org/10.5194/gmd_15 7791 2022

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82*(1), 34–55

APPENDICES

a) Appendix 1: Research Permits/authorization letter

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **511483**

Date of Issue: **19/July/2025**

RESEARCH LICENSE

This is to Certify that Mr.. Brian Atandi Onyiego of The Cooperative University of Kenya, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Nairobi on the topic: SECOND HAND CAR PRICE PREDICTION MODEL IN NAIROBI for the period ending : 19/July/2026.

License No: **NACOSTI/P/25/4176424**

511483

Applicant Identification Number

Ag. Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions

Appendix II: Published articles

Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20250403019
Volume 4, Issue3 (September-December 2025), PP: 100-104.
www.indjcst.com
ISSN No: 2583-5300

Second-Hand Car Price Prediction Model in Nairobi

Brian Onyiego[1], Emma Anyika[2], James Obuhuma[3]
[1 2]*Computing and Mathematics, Co-operative University of Kenya, Kenya.*
[3]*Computer Science, Maseno University, Kenya.*

To Cite this Article: *Brian Onyiego[1], Emma Anyika[2], James Obuhuma[3], "Second-Hand Car Price Prediction Model in Nairobi", Indian Journal of Computer Science and Technology, Volume 04, Issue 03 (September-December 2025), PP: 100-104.*

Abstract: *The second-hand car market in Kenya has grown significantly, but traditional valuation methods remain subjective and inconsistent, creating inefficiencies and information gaps between buyers and sellers. These approaches often ignore the combined impact of brand, model, and year of manufacture, mileage, and engine size on resale prices. Machine learning offers a more accurate and transparent alternative. This study applied Linear Regression, Random Forest, and XGBoost to a dataset of 28,000 vehicle listings from SBT Japan. After extensive preprocessing, models were evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2. Linear Regression performed poorly, while ensemble models produced stronger results. Random Forest achieved a testing R^2 of 0.816 with an MAE of Ksh 683,303, XGBoost reached a testing R^2 of 0.837 with an MAE of Ksh 672,930, and a Voting Ensemble combining both models performed best, with a testing R^2 of 0.840, an MAE of Ksh 649,487, and the lowest RMSE of Ksh 1,069,036.*

Key Words: *Used-car valuation; Random Forest; XG Boost; Feature importance; Kenya; SBT Japan*

I.INTRODUCTION

The second-hand car market in Kenya has grown rapidly over the last decade, driven by increasing demand for affordable vehicles and expanding access to global supply chains. Despite this growth, pricing within the market remains inconsistent and often unreliable, as traditional valuation methods rely heavily on manual appraisals, dealer experience, or outdated reference books. These conventional approaches tend to ignore the multidimensional factors that influence vehicle prices, including year of manufacture, mileage, brand, model, fuel type, and engine capacity. As a result, both buyers and sellers face challenges of biased assessments, inefficiency, and information asymmetry.

The emergence of big data analytics and machine learning provides new opportunities to address these gaps by leveraging large datasets and advanced algorithms to identify patterns and generate more accurate price predictions. Prior studies in markets such as China, India, and Europe have demonstrated that models such as Random Forest, Gradient Boosting, and Support Vector Machines can outperform traditional methods by capturing nonlinear relationships between vehicle attributes and resale prices. These approaches not only improve predictive accuracy but also enhance transparency and fairness in the valuation process.

In the Kenyan context, there is limited research that applies advanced machine learning to second-hand car pricing despite the market's economic significance. The current study therefore aims to bridge this gap by curating a large dataset from SBT Japan and developing ensemble-based predictive models. By focusing on the most influential features and validating the models with robust performance metrics, the research seeks to provide evidence-based solutions that can improve decision-making for buyers, sellers, and policymakers in Kenya's second-hand automotive industry.

II.RELATED WORK

Machine learning (ML) has increasingly been applied in automotive price prediction because of its ability to process large datasets and uncover non-linear relationships that traditional statistical models often overlook (Bukvić et al., 2022). Commonly used algorithms include linear regression, decision trees, random forests, gradient boosting algorithms such as XGBoost, and neural networks, each with distinct advantages and limitations (Gupta et al., 2021; Zhu, 2023). Linear regression remains one of the simplest and most interpretable models; however, its predictive capability is limited in high- dimensional and non-linear contexts such as second-hand car valuation. For example, Lu and Song (2023) showed that while multiple linear regression can explain basic relationships between mileage, year of manufacture, and price, its accuracy decreases when interacting variables such as brand reputation and market dynamics are included.

Ensemble methods, particularly Random Forest and Gradient Boosting, have demonstrated superior performance because they combine multiple decision trees to capture complex interactions. Liu et al. (2022) applied a hybrid model integrating particle swarm optimization, grey relational analysis, and neural networks, reporting enhanced prediction accuracy in car pricing. Similarly, Asghar et al. (2021) achieved a coefficient of determination above 0.90 using feature selection with Random Forest, confirming

the strength of ensemble models in capturing diverse automotive attributes. XGBoost, in particular, has gained popularity for its efficiency and robustness, with Zhu (2023) highlighting its effectiveness in handling structured car data while minimizing overfitting.

Neural networks, including multilayer perceptrons and deep learning models, extend predictive capacity by learning complex non-linear relationships, but they often require large datasets and high computational resources (Fathalla et al., 2020; Barlybayev et al., 2023). This limits their applicability in markets like Kenya, where structured datasets are still being developed. Nonetheless, hybrid and multimodal approaches that combine image data, textual descriptions, and structured features are emerging as promising directions in automotive price prediction (Huang, 2023).

Data preprocessing and feature engineering are critical for improving model performance. Studies have emphasized the importance of handling missing values, standardizing features, correcting categorical inconsistencies, and encoding variables such as brand and fuel type (Msiza, 2023; Chen et al., 2022). For example, Bukvić et al. (2022) found that including production year and mileage improved prediction accuracy in the Croatian market, while neglecting categorical attributes such as fuel type and transmission limited the model's scope.

Beyond vehicle-specific characteristics, external variables such as macroeconomic conditions, consumer preferences, and regulatory changes also influence second-hand car pricing (Ghosh, 2018). For instance, inflation, taxation, and policy shifts in import duties can alter resale values significantly. Yet, many existing studies fail to integrate these broader contextual factors, leaving a gap in predictive comprehensiveness.

Overall, the literature shows that ensemble models, particularly Random Forest and XGBoost, consistently outperform linear models in second-hand car price prediction. However, challenges remain in addressing data quality, feature diversity, and market-specific influences. The current study contributes by curating a large dataset from SBT Japan tailored to the Kenyan market, applying robust data preprocessing and feature engineering, and implementing ensemble algorithms under a comparative framework. This approach not only improves predictive accuracy but also addresses the gap in applying advanced ML methods to second-hand car pricing in Kenya's dynamic automotive market.

III.METHODOLOGY

This paper employed a quantitative research design, which aimed at establishing the relationship between second-hand car prices and their influential variables in the Kenyan market.

3.1. Data Collection

The dataset for this research was obtained from *SBT Japan*, a leading online exporter of used vehicles to Kenya. A total of 38,949 car listings were scraped and compiled into a structured dataset. The records included critical vehicle attributes such as year of manufacture, mileage, brand, model, engine capacity, fuel type, transmission type, and final resale price. Optional features such as airbags, alloy wheels, and power windows were also extracted where available. These characteristics were selected based on prior studies that have demonstrated their impact on second-hand car valuation (Bukvić et al., 2022; Gupta et al., 2021; Zhu, 2023).

3.2. Data Processing

Several preprocessing steps were performed to ensure model reliability and robustness. Duplicate entries were removed, while missing values were handled by dropping incomplete price and year records and imputing missing mileage using the mode. Categorical variables such as brand, model, fuel type, and transmission were standardized and encoded into numerical form using one-hot encoding. To reduce bias caused by scale differences, continuous variables such as mileage and engine size were normalized. Outliers were identified using interquartile range filtering and removed to reduce distortion. Following best practices in machine learning, the dataset was split into training (70%), validation (15%), and testing (15%) subsets (Guo et al., 2023).

3.3. Model Selection and Design

Three machine learning algorithms were implemented: Linear Regression (LR), Random Forest (RF), and Gradient Boosting Machines (GBM) using XGBoost. Linear Regression was included as a benchmark model due to its interpretability (Lu and Song, 2023). Random Forest was selected because it captures non-linear relationships, reduces overfitting, and provides feature importance scores (Asghar et al., 2021). Gradient Boosting (XGBoost) was chosen for its high predictive accuracy, robustness in handling missing data, and ability to optimize performance through sequential learning (Zhu, 2023). The dependent variable was the car price, while the independent variables were year of manufacture, mileage, brand, model, engine size, fuel type, and transmission.

3.4. Model Training and Validation

All models were developed in Python using the *scikit-learn* and *xgboost* libraries, with additional support from *pandas*, *NumPy*, and *matplotlib* for data handling and visualization. Model training was conducted on the training dataset, with hyperparameter tuning performed through randomized search and grid search to identify optimal configurations. Cross-validation was applied to validate model robustness and minimize overfitting, and final evaluation was conducted on the unseen test dataset.

3.5. Evaluation Metrics

The performance of the predictive models was evaluated using three standard regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). MAE provided an intuitive measure of

average prediction error, RMSE penalized larger deviations more heavily, and R^2 indicated the proportion of variance in car prices explained by the models. These metrics are widely adopted benchmarks in automotive price prediction studies (Chen et al., 2022; Msiza, 2023).

3.6. Implementation and Deployment

The final models were integrated into a prototype price prediction tool. This system allows users, such as car buyers, sellers, and dealers, to input key vehicle attributes (e.g., year, mileage, brand, and engine capacity) and receive immediate price estimates. The tool is designed to improve transparency, reduce valuation subjectivity, and provide practical support to decision-making in Kenya's second-hand car market.

3.7. Ethical Considerations

The study adhered to ethical guidelines by using only publicly available and non-identifiable data. Web scraping was conducted in compliance with the source platform's terms of service, and the dataset was used exclusively for academic purposes. Transparency in model design, reporting, and limitations was prioritized to avoid misuse of predictions. The authors emphasized that the generated price estimates should be treated as probabilistic rather than absolute values, to prevent potential exploitation in the market.

IV.RESULTS

4.1. Data Overview and Preprocessing

The starting dataset was cleaned. After removing exact duplicates, dropping rows with missing target (price(Ksh)) and year, standardizing categorical values (e.g., fuel type, transmission), constraining plausible ranges (price between five hundred thousand and fifteen million Kenya shillings, mileage between one thousand and one hundred and fifty thousand kilometres, engine capacity between nine hundred and five thousand cubic centimetres), engineering car_age =2025 − year, and grouping rare brands into "Other", the final analytic sample comprised 28000 records. Key numeric summaries were as follows (post-cleaning):

Table 1. Summary statistics results

Variable	Count	Mean	Standard deviation	Minimum	Median	Maximum
Price (Kenya shillings)	28,000	5,630,688	2,710,810	545,522	5,259,798	12,155,850
Year	28,000	2019.72	2.24	2015	2020	2024
Mileage (kilometres)	28,000	39,792.54	35,513.52	1,000	30,000	149,000
Engine size (cc)	28,000	2,361.45	864.27	900	2,000	4,300

Exploratory analysis revealed strong relationships between vehicle attributes and price.

- **Year vs Price:** newer models correlated positively with higher resale value.
- **Mileage vs Price:** higher mileage showed a negative association with resale price.
- **Engine size vs Price:** larger engines correlated positively with price.

.

4.2. Model Performance

Three predictive models were implemented: Linear Regression, Random Forest, and XG Boost. Linear Regression served as a baseline, while Random Forest and XG Boost were applied to capture non-linear interactions.

- **Linear Regression**: Underperformed, showing relatively low predictive power due to inability to capture complex feature interactions.
- **Random Forest**: Improved accuracy with strong generalization and reasonable error margins.
- **XG Boost**: Achieved the best overall performance with the lowest Mean Absolute Error and highest R^2 on test data.

Table 2. Model evaluation results

Model	Train R^2	Test R^2	Test MAE (Ksh)	Test RMSE (Ksh)
Linear Regression	0.4643	0.4416	1,502,970.41	1,993,966.51
Random Forest	0.9725	0.8164	683,302.90	—

XG Boost	0.8886	0.837	672,929.70	—
Ensemble (RF+XGB)	0.9458	0.8395	649,486.65	1,069,035.59

4.3. Model Optimization

Hyperparameter tuning using Randomized SearchCV and Grid Search CV significantly improved the performance of ensemble models. The tuned Random Forest achieved a training R^2 of 0.9725 and a testing R^2 of 0.8164, with a Mean Absolute Error of Ksh 683,302.90. The tuned XG Boost model recorded a training R^2 of 0.8886 and a testing R^2 of 0.8370, with a Mean Absolute Error of Ksh 672,929.70.

The best overall performance was obtained from the Voting Regressor ensemble, which combined Random Forest and XGBoost. This model yielded a training R^2 of 0.9458 and a testing R^2 of 0.8395, with a Mean Absolute Error of Ksh 649,486.65 and a Root Mean Squared Error of Ksh 1,069,035.59. These results demonstrate that ensemble learning is more robust and reliable than single algorithms for second-hand car price prediction.

Table 3. Tuned ensemble models (final evaluation)

Model	Train R^2	Test R^2	Test MAE (Ksh)	Test RMSE (Ksh)
Random Forest (tuned)	0.9725	0.8164	683,302.90	1,148,389.17
XG Boost (tuned)	0.8886	0.837	672,929.70	1,148,389.17
Voting Ensemble (RF+XGB)	0.9458	0.8395	649,486.65	1,069,035.59

V.CONCLUSION AND FUTURE WORK

5.1 Discussion

This study demonstrates how machine learning can be applied to improve second-hand car price prediction in Kenya, a market where traditional valuation methods are often subjective, inconsistent, and poorly adapted to changing trends. In line with earlier research, linear regression provided interpretability but was not effective in capturing the non-linear and high-dimensional interactions that influence car pricing.

By contrast, ensemble models performed better. The Random Forest model achieved a training R squared of about ninety-seven percent and a testing R squared of about eighty-one percent, with a mean absolute error of roughly six hundred and eighty-three thousand Kenya shillings. The XGBoost model recorded a training R squared of about eighty-nine percent and a testing R squared of about eighty-four percent, with a mean absolute error of roughly six hundred and seventy-three thousand Kenya shillings. The Voting Ensemble that combined Random Forest and XGBoost gave the strongest results, with a training R squared of about ninety-five percent and a testing R squared of about eighty-four percent. It produced a mean absolute error of about six hundred and forty-nine thousand Kenya shillings and the lowest root mean squared error of about one million and seventy thousand Kenya shillings. These findings confirm that ensemble learning is more suitable for handling the complex and non-linear patterns that determine resale prices.

The analysis of feature importance showed that brand and model were the most influential predictors of resale value, followed by year of manufacture, mileage, and engine size. Fuel type and transmission contributed less but still played consistent roles. These results mirror the Kenyan market reality, where vehicle identity and usage history are the key drivers of price.

Cross-validation and testing demonstrated that the tuned ensemble models generalized well, with training and testing scores remaining close. This robustness shows the potential of machine learning to deliver scalable, data-driven pricing tools for Kenya's second-hand car industry. However, the study was limited by the absence of broader economic and consumer preference data, which also affect car valuations.

5.2 Conclusion

This research developed and tested machine learning models for predicting second-hand car prices in Kenya, with the aim of addressing inefficiencies in conventional valuation methods. Among the models applied, Random Forest and XGBoost both outperformed linear regression, and the ensemble approach gave the most accurate and reliable predictions. The results confirm that non-linear, ensemble-based methods are better suited to the Kenyan used car market.

The study contributes in two main ways. Practically, it provides a foundation for web-based or system-integrated predictive tools that can assist buyers, sellers, and dealerships in making transparent and fair pricing decisions. Theoretically, it adds to the growing evidence that ensemble learning improves predictive performance in dynamic and high-variance markets.

5.3 Future Work

Further research can build on these results by incorporating:

- Broader economic indicators such as exchange rates, inflation, taxation, and import duties.
- Consumer preference and socioeconomic data to better reflect market behaviour.
- Multimodal data including car images and textual descriptions to strengthen prediction.
- Interpretable machine learning techniques such as SHAP values and LIME to enhance trust and explainability. These extensions would not only improve accuracy but also make machine learning-based car price forecasting more actionable for Kenya's automotive sector, policymaking, and consumer protection

References

1. Bukvić I, Đonko D, Šimunović I. Predicting used car prices using machine learning techniques. *Journal of Information and Organizational Sciences*.2022;46(1):1–12
2. Gupta S, Kumar A, Singh M. Machine learning approaches for automobile price prediction. *International Journal of Computer Applications*. 2021;183(34):25–31
3. Zhu L. Second-hand car price prediction based on XGBoost algorithm. *Procedia Computer Science*. 2023;222:122–128
4. Lu J, Song Y. Application of regression models in vehicle price forecasting. *International Journal of Data Science and Analytics*. 2023;15(2):145– 154
5. Liu Y, Chen Q, Zhang H. Hybrid car price prediction model using particle swarm optimization, grey relational analysis, and neural networks. *Applied Intelligence*. 2022;52(11):12450–12462
6. Asghar S, Khan R, Iqbal S. Feature selection and Random Forest model for predicting used car prices. *International Journal of Advanced Computer Science and Applications*. 2021;12(5):112–120
7. Fathalla A, Shehab M, Hussein M. Deep learning approach for used car price prediction. *International Conference on Artificial Intelligence and Data Analytics*. 2020:221–227
8. Barlybayev A, Moldabekov Y, Toleubayev M. Neural networks for automotive price prediction: challenges and opportunities. *IEEE Access*.
9. 2023;11:42210–42221
10. Huang J. Multimodal deep learning for car price prediction using text and image features. *International Journal of Machine Learning and Cybernetics*. 2023;14(4):1225–1238
11. Msiza I. Preprocessing strategies for improving machine learning models in automotive pricing. *South African Journal of Industrial Engineering*. 2023;34(2):88–95
12. Chen L, Wang H, Zhao Y. Data preprocessing and feature engineering for predictive modeling in car markets. *Expert Systems with Applications*. 2022;204:117–130
13. Ghosh S. Macroeconomic factors affecting second-hand car prices: an emerging market perspective. *Journal of Economic Studies*. 2018;45(6):1221– 1235
14. Guo X, Li J, Sun P. Train-validation-test splits in applied machine learning: guidelines and best practices. *ACM Computing Surveys*. 2023;55(8):1– 30

Appendix III: Figures and tables,

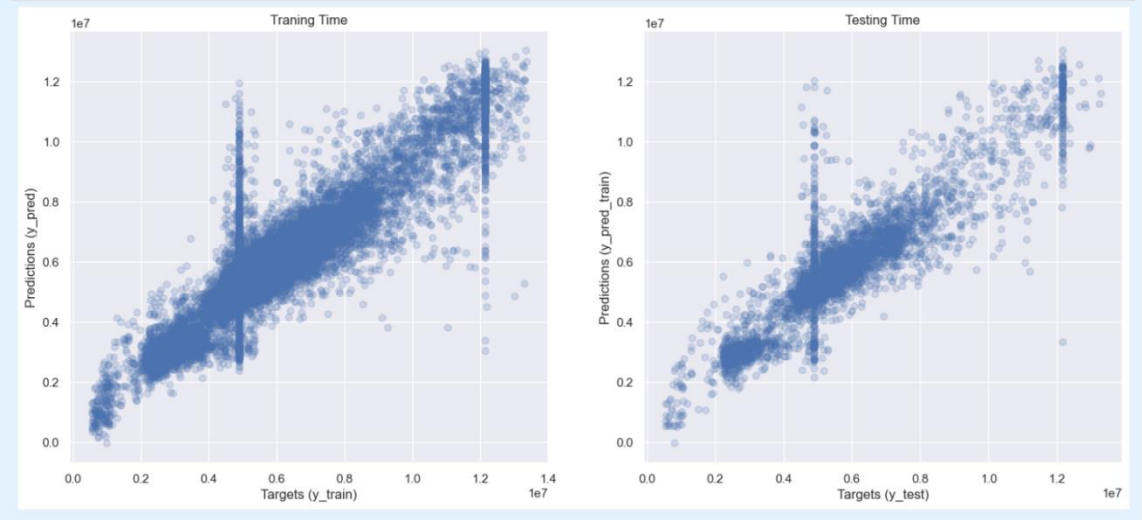
Feature	Pearson r	P-value
Year	**0.361**	0
Mileage (km)	**–0.157**	0
Engine Size (cc)	**0.504**	0

Cleaning Task	Description	Result
Duplicate Removal	Dropped exact duplicate entries	Dataset size reduced by 1,178 rows
Missing Value Handling	Dropped rows missing price(Ksh) and year; imputed missing mileage	Preserved core information
Fuel Type Corrections	Standardized entries (e.g., "Petol", "diesal") into: Petrol, Diesel, etc.	Reduced category fragmentation
Brand Cleanup	Removed placeholder brands, grouped rare brands as "Other"	Improved brand coherence
Transmission Fix	Corrected typographical errors (e.g., 'automatc' - 'automatic')	Improved feature usability
Outlier Filtering	Removed implausible values for price, mileage_km, and engine_size_cc	Reduced skew and distortion
Redundant Column Removal	Dropped image_url, car_url, and location columns	Decreased dimensional noise

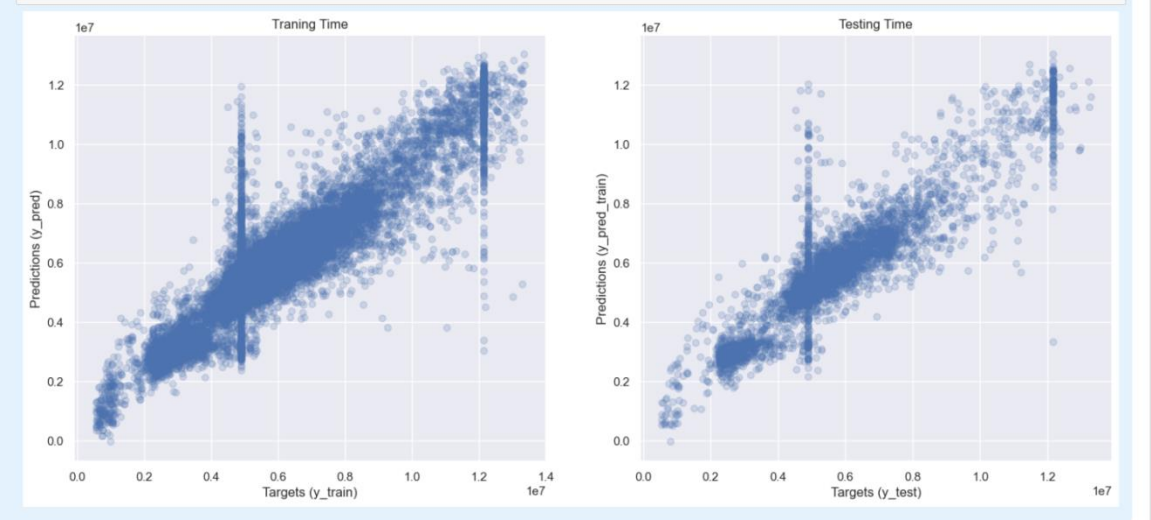
Random Forest Regressor Performance

Model Version	RMSE (Train)	RMSE (Test)	R^2 (Train)	R^2 (Test)	MAE (Test)
Initial RF Model	512,438.69	1,078,324.44	0.9864	0.8984	675,824.55
RF after Random Search	430,279.80	1,038,761.65	0.9902	0.9041	639,219.77
RF after Grid Search 2	417,509.88	1,027,189.21	0.991	0.9065	621,582.31

Residual Plot for Random Forest



Residual Plot for XG-Boost



K-Fold Cross Validation Results (k=5)

Model	**Mean R^2 Score**	**Std. Deviation**
Random Forest	0.9032	0.0147
XG-Boost	0.9105	0.0121