ENSEMBLE MACHINE LEARNING MODEL FOR PREDICTING USED CAR PRICES IN KENYA.

MOSES ONSERIO JAMES

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

2025

DECLARATION

Declaration by the candidate

This project is my original work and has not been presented for any other degree or award in any university.

21/11/2025

…….. ……

Signature

Moses Onserio James : MDATC01/6051/2022 Date

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

21/11/2025

……. ……..

Signature

Dr. Fidelis Mukudi Date

Department of Mathematical sciences,

School of computing and Mathematics,

The Cooperative University of Kenya.

21/11/2025

…… ……

Signature

Dr. Ronald Ojino Date

Computing and Informatatics,

School of Sciences and Technology,

Open University of Kenya.

DEDICATION

I dedicate this work to my family my loving parents Mr. James Onserio and Mrs. Perpetual Onserio and my siblings Esther and Macvivian.

ACKNOWLEDGEMENT

First, I take this opportunity to thank God for the gift of life. Secondly, the management of the University for the Financial Support and a conducive environment for learning. Special thanks to my supervisors, Dr. Fidelis Mukudi and Dr. Ronald Ojino for guidance, and encouragement that shaped the research. Special thanks to Dr. Shem Mbandu, Dean School of Computing and Mathematics, and Dr. Charles Katila, Chairperson Department of Computer Science and Information Technology who shared insights into the degree programme .To my father and mother, I say thank you for both financial and emotional support. You will always remain in my heart. To my siblings Esther and Macvivian who give hope in this World. Let this work inspire you.

Thanks to all Master of Science in Data Science students with whom I shared knowledge, staff, and lecturers at The Co-operative University of Kenya for your support.

TABLE OF CONTENTS

DECLARATION i

DEDICATION ii

ACKNOWLEDGEMENT iii

LIST OF TABLES vi

LIST OF FIGURES vii

DEFINITION OF TERMS viii

LIST OF SYMBOLS AND ABBREVIATIONS x

ABSTRACT xi

CHAPTER ONE 1

1. INTRODUCTION 1

1.1 Background study 1

1.2 Problem Statement 3

1.3 Objectives 4

1.4 Significance of the Study 5

1.5 Scope of study 5

1.7 Limitations of study 6

CHAPTER TWO 7

2. LITERATURE REVIEW 7

2.1 Introduction 7

2.2 Research Gap 14

2.3 Assumptions of the Study 16

2.4 Expected Outcomes of the Study 17

2.5 Theoretical frameworks 17

2.6 Conceptual framework 17

CHAPTER THREE 19

3. RESEARCH METHODOLOGY 19

3.1 Introduction 19

3.2 Design Science Research Paradigm 19

CHAPTER FOUR 29

4. DATA ANALYSIS PRESENTATION, AND INTERPRETATION 29

4.1 Introduction 29

4.2 Data Preprocessing and Exploration 29

4.3 Model Development 36
4.4 Model Evaluation 40
4.5 Feature Importance Analysis 43
CHAPTER FIVE 46
5. DISCUSSION OF FINDINGS, CONCLUSIONS, AND RECOMMENDATIONS 46
5.1 Discussion of findings 46
5.2 Conclusion 47
5.3 Limitations and Delimitations 48
5.4 Recommendations 48
5.5 Future work 48
REFERENCES 49
APPENDICES 53

LIST OF TABLES

1. Summary of key research gaps…….15

2. Descriptive Statistics for Key Numeric Features (After Initial Cleaning)…….34

3. Performance of Individual Base Models……...38

4. Model Comparison Table……...40

5. Ensemble Performance Table……42

6. Feature Importance score table……43

LIST OF FIGURES

Figure 1. Conceptual framework……**20**

Figure 2. DSR principles…….19

Figure 3. Distribution of price…….30

Figure 4. Distribution of log transformed price…….……..…..……31

Figure 5. Distribution of year of manufacture…….…..32

Figure 6. Year of manufacture vs price……..…32

Figure 7. Year of manufacture vs log transformed price……..…32

Figure 8. Mileage vs price…….……..…..36

DEFINITION OF TERMS

Machine Learning: A field of artificial intelligence where algorithms learn patterns from data to make predictions without being explicitly programmed. Source: CERN (2024).

Ensemble Learning: A technique that combines multiple models (base learners) to improve prediction accuracy compared to a single model.Source: IBM (2024).

Base Learner: An individual model (e.g., Decision Tree, SVM, KNN) used as a building block in an ensemble method.Source: Wikipedia (2024).

Random Forest: An ensemble method that builds many decision trees and averages their outputs to reduce overfitting and improve prediction performance.Source: IBM (2024).

Support Vector Regression (SVR): A machine learning method that predicts continuous values by fitting a function with the maximum possible margin using support vectors.Source: Coastal Wiki (2024).

K-Nearest Neighbors Regression (KNN): A non-parametric algorithm that predicts a value based on the average of the k nearest data points in feature space.Source: Carnegie Mellon University (CMU) Notes (2023).

Gradient Boosting: An ensemble technique where models are built sequentially, each correcting the errors of the previous model to improve accuracy.Source: Journal of AI & Engineering Applications (2024).

Stacking Regressor: An ensemble method that combines predictions from several base models using a second-level (meta) model to improve final prediction accuracy.Source: Journal of AI & Engineering Applications (2024)..

Mean Absolute Error (MAE): A regression metric that measures the average absolute difference between predicted and actual values. Lower MAE means better accuracy. Source: Standard statistical texts.

Root Mean Squared Error (RMSE): The square root of the average squared prediction errors, expressed in the same units as the target variable.Source: Standard statistical texts.

R-Squared (R^2): A measure of how much of the variation in the target variable is explained by the model.Source: Standard statistical texts.

LIST OF SYMBOLS AND ABBREVIATIONS

Mathematical and Statistical Symbols and abbreviations

ANN - Artificial Neural Network

CC - Engine capacity in cubic centimeters

IQR - Interquartile Range

K-Fold - A cross-validation method where K represents the number of splits in the dataset

KNN - K-Nearest Neighbors

Ksh - Kenyan (currency)

LSTM - Long Short-Term Memory

MAE - Mean Absolute Error

ML - Machine Learning

OLS - Ordinary Least Squares

R^2 - R-squared (coefficient of determination)

RMSE - Root Mean Square Error

SHAP (SHapley Additive exPlanations)

SLA - Service Level Agreement

SVM - Support Vector Machine

VIF - Variance Inflation Factor

ABSTRACT

Most Kenyan car owners prefer used vehicles due to their affordability, leading to a booming used car market. However, the absence of an objective pricing mechanism has led to inconsistent and subjective pricing, with prices varying significantly from seller to seller. This research aimed to provide a data-driven solution by incorporating key vehicle attributes. Using Design Science Research (DSR) methodology, the research implemented machine learning techniques: Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Gradient Boosting, Linear regression as base models, and Permutation for feature explanation to enhance accuracy and interpretability. The individual models were trained and evaluated using 5 cross-validation. Random Forest emerged as the best with a Mean Absolute Error of 0.1174, and Linear regression was the last with a Mean Absolute Error of 0.2635. For performance optimization, the four best baseline models (RF, SVM, KNN, and GB) were combined using a Stacking Regressor, which achieved an **R-squared score** of 0.9725, a mean absolute error (MAE) of 0.1137, and a mean squared error (MSE) of 0.2171, showing an improved predictive performance compared to individual models. Feature importance analysis identified mileage, car age, annual insurance, engine size, and usage type (Kenyan/Foreign) as the most influential features. These findings are significant because they demonstrate that machine learning can provide an objective, reliable, and data-driven pricing mechanism for the Kenyan used car market. The results offer practical value to car buyers, sellers, dealerships, and insurance companies by reducing pricing disparities, improving transparency, and supporting more informed decision-making. The developed ensemble model can therefore be applied as a practical tool for accurate price estimation, contributing to fairness, efficiency, and better market regulation in Kenya's used car industry.

CHAPTER ONE

1. INTRODUCTION

1.1 Background study

The number of vehicles in Kenya has been growing at a rapid rate of 12% annually, with the national registered fleet standing at 4 million as of 2018 (Kenya National Bureau of Statistics, 2023). This significant growth highlights vehicles' critical role in the Kenyan economy. Regular pricing of vehicles is essential for a variety of purposes, including insurance, resale, leasing, and accounting, among others. However, the existing systems for automobile pricing exhibit considerable variability, as values often differ significantly even for identical vehicles. This inconsistency points to the inefficiencies in traditional pricing methods (Lessmann & Voss, 2017).

Currently, obtaining a car price requires contacting licensed experts, such as evaluation firms, insurance agents, or an individual who knows the field. These methods rely heavily on expert opinions and standardized depreciation formulae, which consider factors like mileage, age, transmission, fuel type, engine capacity, horsepower, and condition(accident or accident-free) among others. While these traditional approaches are widely used, they often lack uniformity and consistency. The advent of machine learning has shown promise in automating vehicle pricing, offering more reliable and accessible solutions.

Predicting the prices of used cars is particularly relevant in today’s automotive market, as used vehicles have become a priority for many buyers due to their affordability. For instance, in the United States alone, approximately 40 million used-cars are sold annually. On average, used cars are 50% cheaper than new ones, allowing buyers to save on financing costs and avoid the sharp depreciation new vehicles face, which can be as high as 11% within the first drive from the lot (National Automobile Dealers Association, 2023).

The Kenyan used-car market primarily relies on imports, with Japanese brands of vehicles like Toyota dominating due to their affordability (Doctor, 2024). However, the market faces challenges, such as fluctuating import duties and currency depreciation, which significantly affect car prices (Mobility Foresights, 2024). Recent reports indicate a sharp rise in the cost of used cars in Kenya. For example, between September 2023 and early 2024, the price of a used Mercedes C-Class 2017 model increased from KSh 3.8 million to at least KSh 4.4 million, while the Toyota Harrier's price rose from KSh 3.8 million to KSh 4 million (Munda & Mutua, 2024). Smaller models, such as the Mazda Demio and Honda Fit, have also experienced substantial price hikes. The primary drivers of these price increases include higher import duties—which rose from 25% to 35% in July 2023—excise taxes, VAT, rising credit costs, and the depreciation of the Kenyan shilling against the dollar. These factors have resulted in reduced vehicle imports, with spending dropping by $162 million in 2023 (National Treasury Tax Expenditure Report, 2024). In response, households and dealers have increasingly turned to older, locally available cars due to reduced costs. The Kenya Auto Bazaar Association attributes this trend to declining household purchasing power and limited access to loans for small traders (Africa, 2024)

At the same time, challenges such as varying pricing methodologies complicate the market further. Pricing a car for resale often involves engaging an expert or referencing market trends, but these methods are inconsistent and highly subjective. Sellers may price their vehicles too high and face long delays in selling, or they may set prices too low and incur losses.

Coming up with a standardized model that can be used to predict the prices of the used cars in Kenya is very crucial. The model can provide more accurate and uniform price by integrating key features .This will be very important to both sellers and buyers since it reduces market efficiencies

which brings about transparency and of course fostering trust in the automotive market. From the recent rise in the application of current technologies in the automotive industry ,the use of machine learning gives a very solid chance to modernize car pricing in the Kenyan market. In the market, we have consistent subjective pricing. This does not give a uniform price across board, therefore, adopting the use of data driven approach which is a predictive model can help to fill the gap between real market prices and intuited ones.This study does not only address the use of machine learning to predict the prices of used cars but also contributes immensely to application of data driven approaches in the global trends.The study intends to assist buyers to predict the values of used cars before they decide to purchase them.

1.2 Problem Statement

The reliance on inconsistent and subjective pricing methods has made the pricing of used cars a challenge. Sellers often overestimate vehicle prices to maximize profits, leading to long delays in sales, while buyers struggle with unreliable pricing information, which breeds mistrust. The existing price prediction models have not been tested on the Kenyan market, and due to unique market variabilities, they may not be directly applicable. In addition to that, most of these studies don't explain the contribution of individual factors to the target variable (price). There is currently no data-driven, standardized tool to address these challenges in Kenya's used-car market. To solve this problem, there is a need for a tailored ensemble machine learning model that incorporates key vehicle attributes, for instance, make, year of manufacture, drive, annual insurance, engine size, mileage, fuel type, horsepower, torque, usage type(Kenyan/Foreign), and acceleration. Such a model will provide accurate price predictions, reduce inefficiencies, and improve market transparency.

1.3 Objectives

1.3.1 Main Objective

To develop an ensemble machine learning model that predicts the prices of used cars in the Kenyan market.

1.3.2 Specific Objectives

1. To identify the most important factors that should be considered when predicting the prices of used cars in Kenya.
2. To develop an ensemble machine learning model that can be used to predict used car prices in Kenya.
3. To validate the developed ensemble model.
4. To evaluate the performance of the validated model using appropriate metrics.

1.3.3 Research Questions

1. What are the most important factors that influence the price of used cars in Kenya?
2. How can an ensemble machine learning model be developed?
3. How well does the developed ensemble model perform when applied to a real-world dataset of used car prices in Kenya?
4. Which performance metrics provide the most reliable assessment of the validated ensemble model's accuracy in predicting used car prices in Kenya?

1.4 Significance of the Study

The purpose of this research was to develop an ensemble machine learning model to predict used car prices in Kenya by combining Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Gradient Boosting, and Permutation for feature explanation. The study sought to address the existing pricing challenges by creating a data-driven solution that integrates vehicle-specific attributes. By providing reliable and standardized price estimates, the model enhanced trust, improved market efficiency, and supported better decision-making for buyers, sellers other stakeholders, and contributed to the growing body of knowledge on applying machine learning in emerging markets.

1.5 Scope of study

The scope of this study was to develop an ensemble machine learning model to predict the prices of used cars in Kenya. This encompasses both vehicles that were initially used within Kenya and those imported from other countries, but have a history of prior use. The model utilized real-world datasets with key attributes: name of the car, model, annual insurance make, year of manufacture, mileage, engine size, fuel type, horsepower, transmission, condition, car name, body type, source (Kenyan/foreign used). Ensemble techniques: Random Forest, Support Vector Machines (SVM), Gradient Boosting, and K-Nearest Neighbors (KNN) were applied to enhance accuracy. Permutation was also used to explain feature contributions.

1.6 Justification of the study

The Kenyan used car market plays a vital role in transportation, offering affordable mobility options to a significant portion of the population. While characterized by a preference for durable

Japanese imports and locally used cars, the market faces challenges such as price uncertainty, limited transparency, and the proliferation of counterfeit parts. (Chepkwony, 2022). These issues hinder consumer confidence, lead to pricing inefficiencies, and stifle market growth. Developing a machine learning-based price prediction model specifically tailored to the Kenyan context was crucial. This model enhanced market transparency by providing objective price estimates, empowering buyers with informed decision-making, and reducing the risk of overpricing. By incorporating key factors like brand, engine size, age, drive, transmission, fuel type, mileage, acceleration, and horsepower, the model offered more accurate predictions than existing global models.

1.7 Limitations of study

The study was limited to the constraint of financial resources. To ensure the feasibility of data collection within budget constraints, the research primarily relied on an online data collection method which is web scraping.

CHAPTER TWO

2. LITERATURE REVIEW

2.1 Introduction

The literature review explores studies conducted by various researchers on car price prediction using machine learning techniques. It examines the key variables employed in these studies, such as mileage, year of production, brand, model, fuel type, and other features influencing car prices. The review also discusses the machine learning models applied, including linear regression, Random Forest, Gradient Boosting, and Neural Networks, highlighting their respective outcomes and performance metrics. Furthermore, the chapter identifies the future recommendations proposed by these studies, such as incorporating advanced algorithms, expanding datasets, or improving model explainability. By synthesizing these findings, the review highlights how these studies address existing challenges in car price prediction while aligning with the research gaps this study aims to address. This serves as a basis to come up with a data driven model suited to this research.

Researchers have done models on the car pricing sector. Despite this significant progress, majority of them have relied on single modeling technique, use of narrow scope or partial variables. For instance Gegic et al. (2019) built a model for both Bosnia and Herzegivina market. They developed it using an ensemble approach by combining Artificial Neural Networks, Support vector machines and Random Forest to improve prediction accuracy. The data was obtained from a website known as autopijaca.ba and it was tested in a java application. The output of the model was 87.38%.On top of that, the researchers recommended the use of diversified datasets and scope that can help to achieve generalization

Chandak et al. (2019) investigated the relationship between car attributes like year of registration, kilometers driven and fuel type. The study made use of **K-Nearest Neighbors (KNN)** and

Classification and Regression Trees (CART) on a dataset containing **300,000 entries.** Out of the two CART performed better that KNN by achieving a lower root mean square error of 4961.64 than KNN, which had 5581. The researchers recommended the addition of more features for future studies, like horsepower and torque, to mention but a few. This will improve accuracy.

Pillai (2020), on the other hand, made use of Artificial Neural Networks to predict the price of used cars in the United States market. The model was trained with a dataset of 140,000 cars and 30 different but popular brands, and it was tested by the use of 35,000 cars. This resulted in result achieved a mean absolute error of 11% and R^2 value of 0.96. This performed better than linear regression and random forests, thus making it more accurate in predicting the prices of used cars. Although the study achieved a significant output, it made use of a single model. It is important to come up with an advanced approach to compare it with other models in order to identify the most important model that can be used to predict the prices of used cars.

Research conducted by Kumar and Samruddhi(2020) utilized the K-Nearest Neighbor model to determine the price of used cars in India. Data was collected from Kaggle, which was used to train the model. Different ratios were used to train and test the data, which resulted in the model achieving an accuracy of 85%. This is evident that the KNN algorithm can actually predict the prices of used cars effectively. However, the researchers recommended the use of other advanced models and expanded scope in order to come up with a more accurate model to predict the prices of used cars.

Bukvić et al. (2020) present an overview of data-driven models for predicting the price of used vehicles in the Croatian market. The authors focused on key factors like the production year and

kilometers driven. They gathered data from the online marketplace "Njuškalo" and cleaned it by removing redundant and missing values. The study used linear regression to predict car prices. They also compared the accuracy of this method with classification algorithms. The aim was to analyze the vehicle market and predict price trends based on available data. They concluded that the predicted model has the highest accuracy with linear regression, where main features (price and model) are available. This study is limited to India and paid attention to only key factors (Production year and Kilometers driven) with linear regression as the specific machine learning technique. This calls for a need to do research that brings many features together and combines various machine learning techniques to identify which one predicts the price better.

Bharambe et al. (2022) proposed three regression algorithms to predict the price of used cars. They considered various factors to ensure reliable and accurate predictions. The study used three supervised machine learning techniques: linear regression, lasso regression, and ridge regression. Python libraries like Numpy, Pandas, and Scikit-learn were used to build the model and design the project's graphical user interface (GUI). The accuracy of the models was compared, with linear regression achieving 83.65%, lasso regression 87.09%, and ridge regression 84.00%. Lasso regression was used to make the final price prediction since it achieved the highest accuracy. The researchers recommended the use of other machine learning techniques and the collection of a larger dataset so that they can be used to improve the accuracy of predicting used car prices. The final price prediction was made using lasso regression, as it provided the highest accuracy.

K-Nearest Neighbors regression model was used by Budiono et al.(2024) to predict the prices of used cars. KNN works well with data which is in a multi-dimensional and its highly resistant to data which is noisy. So this makes it an ideal in predicting the prices of used cars. In total, 504

used car data points were collected using a web scraping method. The model finally achieved an R^2 of 98.8% and an error rate of 8.8%.This result can significantly assist both buyers and sellers to determine the prices of used cars. The study utilized only one single model to predict the prices. This calls for an advanced approach of combining different models in order to determine which one performs better and gives more accuracy.

Due to the lack of existing online services for predicting the prices of used cars in Bangladesh, Amik et al.(2021) proposed a model that can be used to help customers. Their main objective was to come up with a data-driven model that can help used car potential buyers predict the prices. The collected data and performed an analysis. They exposed the data to different machine learning algorithms to mention Lasso regression, Linear regression, random forest, and gradient boosting. Out of them all, XGBoost was selected as the best model after achieving a score of 91%. They identified the key factors to consider to mention them: fuel type, engine capacity, kilometers driven, transmission type, year of manufacture, model, body type, and brand of the car. They obtained a dataset from the Bikroy.com website. For future work, the researchers proposed the collection of a larger dataset with many features in order to improve accuracy.

In Mauritius, Peerun et al.(2022) carried out research using artificial neural networks to assess whether it can be used to predict the price of second-hand cars accurately. They collected a record of 200 cars in total from different sources such as newspaper advertisement and car websites. They collected factors like car brand, engine capacity, manufacturing year and mileage. They developed four machine learning algorithms and out of which, support vector machine performed the best than compared to linear regression and artificial neural networks. Although this produced a good result, there was a challenge with higher priced cars where some predictions were significantly off

from actual prices. They made use of Mean absolute error as a metric to measure performance. This brings about a need to diversify the scope and size of the dataset and to use other machine learning techniques to determine which produces more accurate results.

In order to predict used car prices in India, Varshita et al.(2022) carried out a study that aimed at developing supervised learning models integrating both Random Forest and Artificial Neural Networks. They aimed at developing an accurate system that can predict used car prices using various attributes. Machine learning models like Random Forest and linear regression were tested on datasets, while an Artificial neural network was implemented using the Keras Regressor Algorithm. Out of them all, Random Forest emerged as the best by achieving the lowest error with a mean absolute error of 1.097 and an R^2 value of 0.772. It's quite evident that random forest is better as compared to other simpler models. The researchers, however, suggested advancement in the future by testing different machine learning techniques for refinement of the level of accuracy.

In Morocco, in order to assist both buyers and sellers in predicting accurate prices of used cars, Mustapha et al.(2022) came up with a regression model to determine the resale values of used cars. They collected a dataset from Avito, which is a local online e-commerce platform, where they utilized the BeautifulSoup library. This is a library in Python. They considered key features such as model, fiscal power, fuel type, mileage, and year of production. The dataset used comprised a total of 8000 car records. Out of the models tested, the Gradient boosting regressor outperformed all other models by achieving an R^2 of about 0.80. For future work, they recommended the addition of more features and a larger dataset for improved accuracy of predicted used car prices.

Liang Han et al. (2020) developed a multi-modal system that extracts textual and visual features alongside statistical features to determine whether the item uploaded is suitable for better price suggestion. They developed a price suggestion system for online second hand items using both text descripts and images. The system is intended to help sellers on deciding effective and reasonable prices for their second hand items uploaded to online platforms. The assessment in this case is carried out using a binary classification model, while price suggestions are generated using a regression model. This is done for the items that are deemed as qualified. Besides that, the researchers propose some sets of evaluation metrics to evaluate the performance of their system. And this can be done with experiments using real-world datasets. On top of that, they recommended the addition of more features to improve the accuracy of the predicted prices of the used cars.

Kanwal et al.(2017) developed a supervised machine linear regression learning model that could be used to predict the price of used cars. It achieved an accuracy of 98%. The model relied heavily on features: version, make, city, color, mileage, alloy rims, power steering, and vehicle model to predict the price. The system simplified the complexity of the model by choosing features that are more relevant from the dataset and getting rid least important variables during the processing phase. They achieved accurate results. However, they recommended future work to be done by integrating more advanced techniques like Fuzzy logic, K-nearest neighbors, and genetic algorithms to achieve more accurate results.

Random forest, which is a supervised machine learning model, was developed by Pal et al.(2018) to predict the prices of used cars. The model was chosen after careful exploration of the data analysis to determine the contribution of each characteristic to the price. The researchers

developed a random forest of 500 decision trees, which was used to train the dataset. The training achieved an accuracy of 95.82%, and the testing accuracy of 83.63%.The construct predicted the car prices effectively by choosing the key correlated features for accurate results. From the research, it's evident that only one machine learning technique was used. To achieve a higher accuracy index,its important to combine multiple of them.

Zulfiqar and Ahtesham (2022) developed a machine learning models to predict the price of used cars. The constructed linear regression, decision tree and gradient boosting algorithms to predict the prices of cars based on their characteristics. They implemented using apache spark which a very powerful data processing tool. They obtained a dataset from PakWheels which had 56,186 car records in total and it comprised of 16 attributes in total. The researchers complimented a significant progress in building models to predict used car prices. However, they also acknowledged that there is a limited work done which has utilized PyPark. The models they build achieved significant results although they varied in nature. Linear regression achieved 50% in terms of accuracy, decision tree achieved 86% and gradient boosting achieved 89%.They however, recommended that future work to integrate other machine learning techniques like K-Nearest neighbors to improve the predicted results.

Damandeep Kaur and Sachin Kumar (2022) developed a data driven tool to predict the car price in the Asian market. They obtained the dataset from Kaggle website and transformed it through a series of preprocessing steps like cleaning to eliminate noise and standardization. This processes help to elevate the performance of machine learning algorithms. They exposed different algorithms in order to identify which best gives the price. Random forest demonstrated the best result in terms of prediction capabilities. Finally they implemented the model into python-driven application by

the use of Flask to enable user interaction privilege. The research gave out impressing results in predicting car prices. In order to achieve better results, the authors recommended the use of diversified datasets and other machine learning techniques.

2.2 Research Gap

Based on the above analyzed research papers, it's very evident that there has been a wonderful progress that has been made in developing machine learning algorithms that can be used to predict both prices of new cars and used cars in the automotive sector. Techniques like K-Nearest Neighbors (KNN), Artificial Neural Networks(ANN) ,Random forest , XGBoost and Support Vector Machines have been used. However, the studies have the following gaps identified:

1. **Limited Use of Ensemble Methods:** While studies like *Gegic et al. (2019)* and *Bharambe et al. (2022)* touched on ensemble techniques, many of the models rely on single algorithms, such as linear regression, KNN, or ANN. Combining multiple models (ensemble methods) has shown potential to improve prediction accuracy, but this has not been widely applied.
2. **Local Context and Limited Scope**: Most studies have been conducted in markets such as the **U.S., Europe, India, and Mauritius**, with **limited research in the African context**, particularly Kenya. The Kenyan used car market has unique challenges, including **economic factors** for instance high import duties, **aging car inventory** and **market inefficiencies**. This requires models tailored to the Kenyan context.
3. **Feature Engineering Gaps:** Some studies (*Chandak et al., 2019*) emphasized the importance of incorporating additional features like **horsepower, torque or maintenance costs** to improve accuracy, but this has not been widely adopted.

4. **Small and Outdated Datasets:** Many studies used small or geographically limited datasets.

5. Absence of feature importance explanation towards the prediction of the price

Table 1 Summary of Key Research Gaps

Study	Techniques	Strengths	Weaknesses	Focus
Gegic et al.(2019)	Ensemble (ANN, SVM,Random Forest)	High accuracy(87.38%)	Limited data source (Bosnia and Herzegovina), needs diverse datasets for improved generalization	Used car market in Bosnia and Herzegovina
Chandak et al. (2019)	KNN, CART	Identified key features (mileage, year, fuel type)	Lower accuracy compared to other studies (RMSE: 4961.64), recommends including more features (horsepower, torque)	Not specified
Samruddhi & Kumar (2020)	KNN	Achieved good accuracy (85%)	Limited data source (Kaggle website), recommends using advanced techniques for improved accuracy	Used car market in India
Bukvić et al. (2020)	Linear Regression	Focuses on key factors (year, mileage), good accuracy with linear regression	Limited to linear regression, needs exploration of advanced techniques and using more features	Used car market in Croatia

Bharambe et al. (2022)	Linear Regression, Lasso Regression, Ridge Regression	Compare multiple algorithms (highest accuracy with Lasso Regression: 87.09%)	Limited data, proposes using advanced techniques (Random Forest, ANN, CNN) for improvement	Not specified
Mustapha et al. (2022)	Gradient Boosting Regressor	Achieved good accuracy (R^2: 0.80)	Lower accuracy in minimizing RMSE, recommends adding more features for improvement	Used car market in Morocco
Pal et al. (2018)	Random Forest	Achieved good accuracy (training: 95.82%, testing: 83.63%)	Used single technique (Random Forest), proposes combining with other models (Gradient Boosting) for improvement	Not specified

2.3 Assumptions of the Study

1. The available data is assumed to be of reasonable quality, with minimal missing values and inconsistencies. Data cleaning and preprocessing steps will be necessary to ensure data accuracy and reliability.
2. It is assumed that the developed ensemble machine learning model will apply to various pricing scenarios and different categories of used cars within the Kenyan market.
3. The model is assumed to be robust to variations in market conditions and changes in vehicle attributes over time.
4. It is assumed that the data to be collected on car prices accurately reflects all local market trends, including fluctuations in demand, supply, depreciation, and applicable costs such as value-added tax (VAT) and other duties.

2.4 Expected Outcomes of the Study

This research was intended to produce a reliable machine learning model that will be used to predict the prices of used cars in Kenya . It was expected to perform better than compared to other traditional methods in terms of reliability, accuracy and consistency. On top of that, the research was intended to identify and select the most significant factors influencing prices of cars. The developed model was to be a practical construct for both buyers and sellers which can enable them to make informed decisions which brings about transparency in the used car market. Above all, the study was intended to contribute significantly to the application of machine learning in the global emerging markets.

2.5 Theoretical frameworks

2.5.1 Hedonic Pricing Theory

This research paper utilized Hedonic pricing theory (HPT) which explains the value of a product as a determination of its individual characteristics (Hargrave, 2020).It lays its foundation in the characteristics theory of value. This elaborates further that consumers of products tend to get utility not from the products themselves rather from the attributes they perceive (Lancaster, 1996).On the other side Rosen(1974) explained further saying that goods are a collection of features and every consumer has his or her own preferences in the features they go for. This is in line with this study since, buyers of cars tend to perceive value from the characteristics of a car like mileage, year of manufacture, brand to mention but a few.

2.6 Conceptual framework

This section analyzes both independent and dependent variables that are going to be used in this research work. This means that, the variables used will immensely contribute to the end result of the ensemble algorithms.

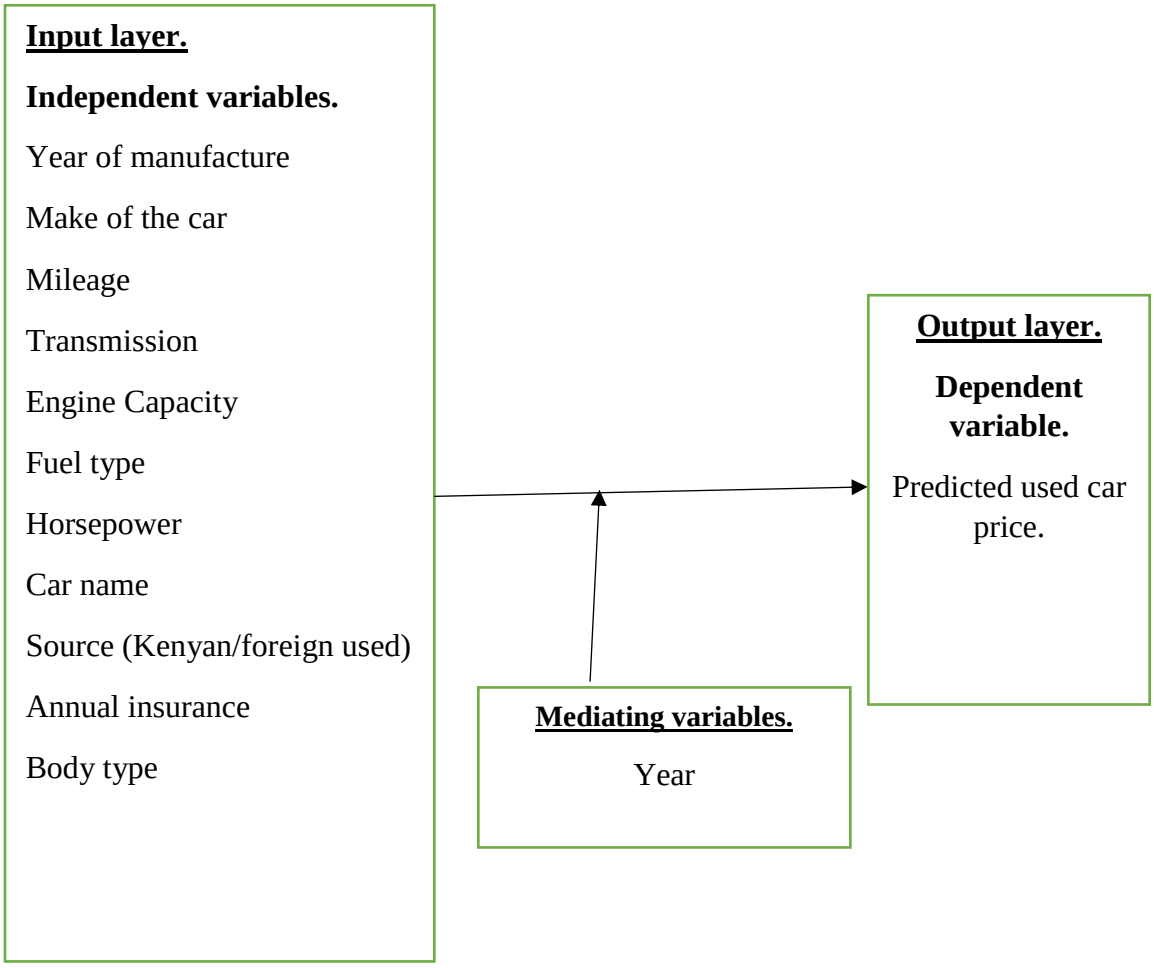


Figure 1 Conceptual Framework

Explanation of the labels

Dependent variable: This is the predicted used car price (output).

Independent Variables: These are the input features for instance model, age, and mileage that directly influence the price of the car.

Mediating variables: Represent the model year of the car or the year the car was sold. Year mediates the impact of age and mileage on the price because newer models generally depreciate differently compared to older models, even if they have the same mileage.

CHAPTER THREE

3. RESEARCH METHODOLOGY

3.1 Introduction

This section explains the research methodology process that was followed in order to develop an ensemble machine learning model used to predict the price of used cars in Kenya. Due to its complexity in nature, which required a practical solution, the study adopted the Design science research methodology. This lays its foundation in the preparation and evaluation of artifacts (Ahmed,D., &Sundaram, 2011). About this study, an ensemble machine learning model was developed to solve the identified real-world problem that which Kenyan automotive market is facing.

3.2 Design Science Research Paradigm

Design science research gave a systematic approach for carrying out this research as it focuses in the development and evaluation of practical solutions to address real world problems. It integrated the principles of DSR which are:

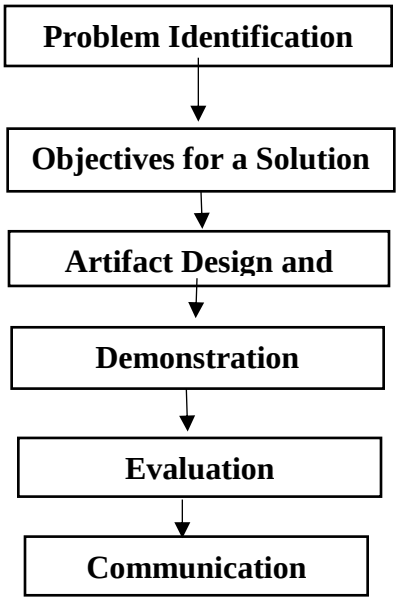


figure 2. DSR principles

3.2.1 Pragmatism as a paradigm in DSR

Pragmatism is an approach that emphasizes the development of practical solutions. Its core principles align perfectly with Design Science Research (Hevner et al., 2004). This philosophically structured approach enables researchers who are interested in Design science research methodology to always prioritize the development of artifacts that solve real-world challenges (Smith and March, 1995).In this study context, this approach paid keen interest in knowing the requirements of both buyers and sellers. This means that gathering or desired features for the data-driven tool is very significant. It also emphasized repeatedly building and evaluating constructs with different approaches of machine learning to finally come up with the most suitable performance. This means that the result was expected to be more accurate.

3.2.2 Problem Identification and Motivation, and Objectives of the Solution

This section on problem identification and motivation with objectives of the solution was fully discussed in chapter one under section 1.2, which is the problem statement and section 1.3, which is the objectives of the study.

3.2.3 Artifact Design and Development

3.2.3.1 Data Collection

i)Target population/Data source

Data was collected from Kenya's online motor bazaars to mention them: Kai and Karo, Gigi, Motor hub, and Maridady. The main reason for selecting these sources is because they consist data which is structured hence consisting minimal inconsistencies. They also consist vehicles of different categories and brands. Their richness in diverse features also makes them stand out from the rest. The data was stored in the laptop's local drive.

ii) Data Features

The following features were collected: Car brand, Annual insurance, Year of manufacture, Horsepower, Engine size, Price, Fuel type, Transmission, name of the car, Source (Kenyan or foreign used).

iii) Data Collection Tools

Since data was collected from cars listed in online motor bazaars in Kenya, BeautifulSoup, a Python library, was used for web scraping. It has the ability to help researchers extract data from both XML and HTML documents.

iv) Programming Language

For programming, the Python programming language was used. It is a simple to learn language which is user-friendly. Its simplicity and readability made it the most suitable language to use in this study.

v) Libraries

- Data Collection**:** BeautifulSoup .
- Data Preprocessing**:** Pandas, NumPy and Scikit-learn.
- Evaluation and Visualization**:** Matplotlib and sklearn.

3.2.3.2 Data Preprocessing

i)Data Cleaning:

All records that consisted of identical features were removed, and any missing value which is numerical was imputed by the use of the mean and median. Categorical data was handled appropriately based on their type: Binary variables were encoded using binary encoding (0/1). Nominal variables were encoded using one-hot encoding to avoid introducing an artificial order.

Outlier Detection: Boxplots and scatterplots were used to identify and handle extreme values in numerical features.

ii) Feature Engineering

a) Derived Features:

"Car Age" was calculated as the difference between the current year and the year of production.

b) **Categorical Encoding:**

For variables like **fuel type** in a used car price prediction model, one-hot encoding was used to ensure that the categorical information is represented numerically without introducing unintended relationships.

c) Scaling and Normalization:

StandardScaler was applied to normalize numerical features, ensuring uniform feature scaling for distance-based models like KNN.

3.2.3.3 Model Development

i) Base Models:

a) Random Forest

In the Kenyan used car market, where a wide variety of brands, models, and conditions exist, Random Forest plays a critical role. It develops multiple decision trees, and each tree provides its prediction score. Then it averages the prediction scores to come up with a weighted average. It performs better when exposed to multiple variables and can deal with overfitting.

b) Support Vector Machines (SVM)

In the Kenyan used car market, factors tend to behave in a non-linear relationship with the prices of cars. This behavior sometimes doesn't exhibit a simple linear pattern. This model is good at dealing with such relationships. It has the ability to handle complex patterns.

c) K-Nearest Neighbors (KNN)

This algorithm is good when capturing localized trends. It's useful in this study for the Kenyan market due to regional forces of demand and supply. You find that, the price of a car can be determined by looking at its closest neighbor

d) Gradient Boosting

This machine learning technique is good since it captures the intricate interplay of the features that influence the prices. This technique works by building a series of decision trees which are termed

as weak learners. Each learner corrects any errors of another learner in a peer review format. Its output is very accurate and is good at understanding any pricing trend.

The development of the used car price prediction model followed a structured experimentation process:

ii) Data Extraction

Data was extracted from various Kenyan online car marketplaces: Kai and Karo, Motor Hub, Gigi, Maridady, using the BeautifulSoup library. The data included key features: Mileage, Year of production, Horsepower, Engine size, Price, Fuel type, Transmission, Car name, Source(Kenyan used or foreign used), Body Type, Annual insurance.

iii) Data Conversion

The collected raw data was cleaned and preprocessed using Pandas and NumPy. Duplicate records were identified and removed to prevent data redundancy. Missing values, which can significantly impact model performance, were addressed. We identified missing values using Pandas functions like “isnull()” and “sum()”. Imputation strategies were then applied: for numerical features, mean or median imputation was used, while mode imputation was employed for categorical variables. One-hot encoding was used to encode categorical variables. For instance, fuel type, we used 0 and 1 to differentiate either petrol or diesel. Whereby, the presence of one type denoted by 1, and the absence of the other denoted by 0. So this conversion was in a format that could be handled by machine learning.

iv) Feature Selection

The study featured filtration to choose the best and most relevant features used in the model training. Importance ranking via random forest was used to perform this task and correlation analysis. This enhanced performance and reduced computational complexity.

v) Model Selection and Application

Once feature selection was done completely, the models listed in this study were trained and tested. Each gave a prediction score that ranked it.

vi) Ensemble Strategy

In order to have more accurate predictions, the research work adopted a Stacking technique in order to bring together the price predictions of each model. This resulted in a final model that produced better results as compared to individual ones.

vii) Model Interpretability

In order to understand the contribution of each feature in the prediction of the price, Permutation were used. This assisted in identifying most important features that influence the price of cars in the Kenyan market.

3.2.3.4 Demonstration

After the model was developed, it was applied to a real-world dataset, which was obtained from the cars listed in the Kenya online motor bazaars. This helped us to understand its ability to predict prices in regard to the input characteristics.

3.2.3.5 Evaluation

i) Evaluation Metrics

Root Mean Squared Error (MSE) ,R-squared (R^2) and Mean Absolute Error were used to evaluate the model performance:

Root Mean Squared Error (RMSE) is a metric that is used widely to evaluate the predictions accuracy especially in regression machine learning tasks. It works by measuring the difference of values that are predicted from the actual ones by the use of Euclidean distance. When the value of RMSE is lower, it means the accuracy is greater. Below is the formula for finding RMSE, whereby "y" represents the actual values and "$\hat{y}$" represents predicted values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

R-squared is used as a metric in regression machine learning, often normalized as Mean Squared It remains unaffected by the scaling of values in the dataset, focusing instead on the proportion of errors relative to the average value. This means that the absolute magnitude of errors does not influence R-squared, only their relative size. Like RMSE, R-squared can also be affected by outliers or unusually large errors. Its values range from 0 to 1, with 1 indicating perfect prediction accuracy, where the model's predictions completely align with the actual data. Higher R-squared values suggest better model performance.

MAPE (Mean Absolute Percentage Error) is another metric used to evaluate prediction accuracy. It calculates the absolute percentage error between predicted and actual values for each instance, then averages these errors to produce the MAPE value. Unlike RMSE and R-squared, MAPE does not square the errors, and the individual errors depend on the relative magnitude of the actual and

predicted values. A MAPE of zero indicates perfect predictions, while lower values signify better model performance. (GeeksforGeeks, 2024).

ii) Cross-Validation

K-Fold Cross-Validation was used: It ensured robust performance evaluation by splitting the data into training and testing subsets multiple times based on the size of data that was collected. 80% of the data was used for training and 20% of the data was used for testing.

iii) Benchmarking

The ensemble model's performance was benchmarked against Individual base models.

iv) Expected Outcomes

The primary objective was for the ensemble model to demonstrate better performance compared to individual machine learning algorithms: KNN, SVM, Random Forest and Gradient boosting, in terms of both accuracy and robustness. By combining the strengths of these diverse models, the ensemble approach achieved more reliable and consistent price predictions. Furthermore, the integration of Permutation provided valuable insights into the key factors that most significantly influence used car prices.

3.2.3.6 Communication of Results

The research findings was disseminated through Academic Publications which is submission of papers to relevant journals.

3.3 Ethical Consideration

Ethical practices are crucial throughout the research process. Data scraping was conducted responsibly, adhering to the terms of service of the online platforms used. Any sensitive or proprietary information encountered was anonymized to protect user privacy and intellectual property rights. Furthermore, the research findings were presented transparently, acknowledging any potential biases or limitations of the model. This transparency fosters trust and ensures responsible use of the developed pricing model.

CHAPTER FOUR

4. DATA ANALYSIS PRESENTATION, AND INTERPRETATION

4.1 Introduction

This chapter outlines the systematic process that was followed to build and evaluate the machine learning model used to predict the prices of used cars in Kenya. It consists the data preprocessing, feature engineering, baseline model selection, hyperparameter optimization, and the building of an advanced ensemble model.

4.2 Data Preprocessing and Exploration

This section gives an overview of the steps that were taken to transform the raw dataset into a suitable format that can be used for machine learning algorithms.The original dataset consisted of 13,667 entries in total. After examining the dataset carefully, it revealed some notable data quality challenges of missing values across various features. The price column, which is the central variable for this analysis, was confirmed to be numeric, and it consisted only of positive values, although it had a missing data rate of 4.62%. The year of manufacture also revealed a significant anomaly. About 2308 instances were recorded as 0, which is an invalid manufacturing year. The 0 values, however, were converted to NaN, which led to an increased count of missing values for the critical feature. The currency column also presented a recognizable challenge for price standardization.

4.2.1 Feature Engineering for Predictive Power

In order for the model to have a better predictive capability, action was taken in feature engineering. The price variable, which is the target for prediction, showed a severe right-skewness with a skewness value of 6.86. To normalize this distribution, it's important to reduce the undue

influence of extreme outliers so that it can be more amenable to various regression algorithms. An np.log1p transformation was applied, which resulted in a new feature called price_log. This transformation managed to compress the range of values, and it reduced the skewness significantly to 0.64, thus yielding a more symmetrical distribution.

Distribution of Price (Original and Log-Transformed)

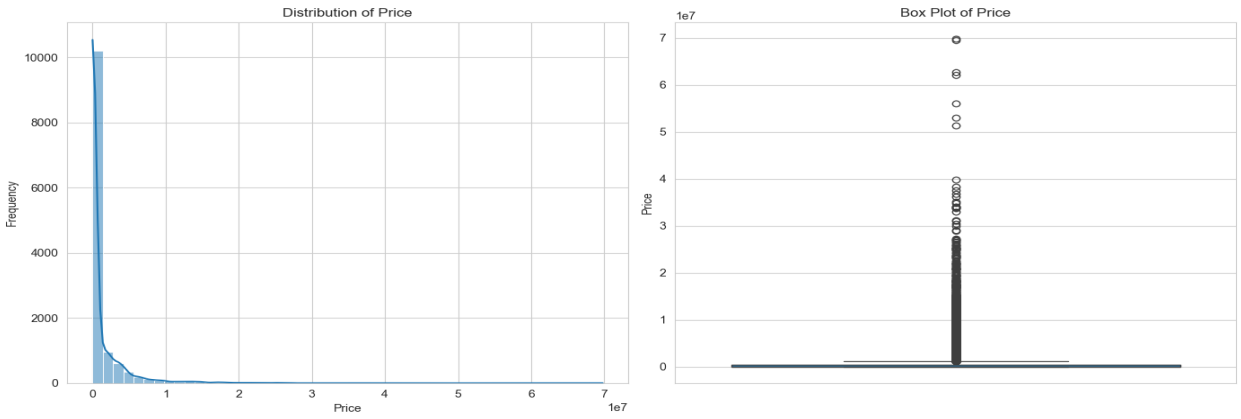


Figure 3 Distribution of price

It's evident from the histogram that the initial price displays a distribution with a lot of right skew. This means that the majority of cars are clustered around lower prices, and few are much higher. The box plot shows clearly that there are many outliers as well.

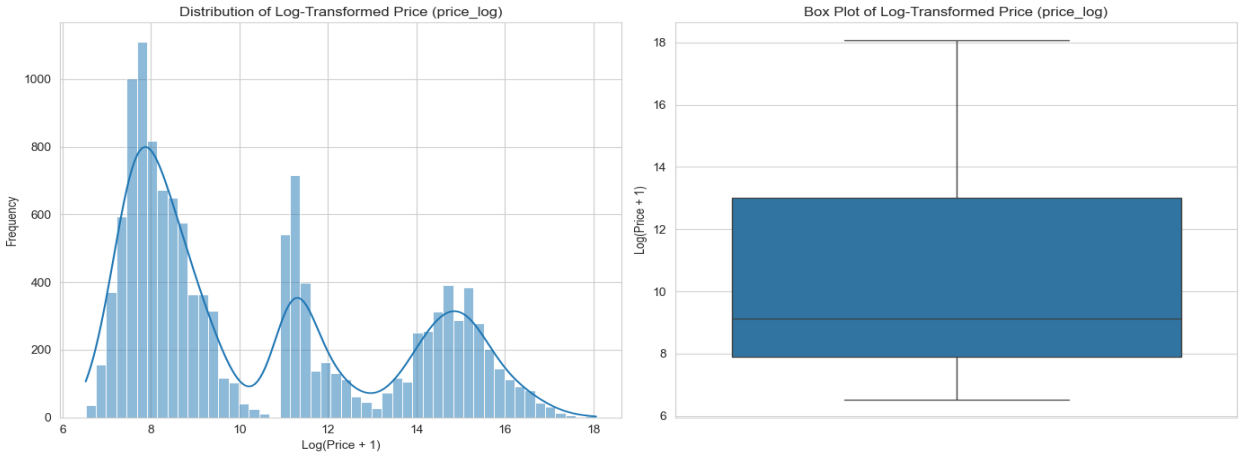


Figure 4 Distribution of Log Transformed price

After applying the np.log1p transformation, we noticed that the price_log histogram is much more symmetrical than before. It also has less skewness as well as fewer extreme outliers in the box plot. Nevertheless, it still retains some slight bimodal or multimodal traits, which indicate possible distinct price groups.

From the year_of_manufacture variable, two new temporal features were derived: car_age and car_age_squared. The car_age feature directly captures the linear depreciation of a vehicle over time, thus informing of the potential price that car_age_squared was brought on board to account for any potential non-linear depreciation pattern. These two car age-related features managed to demonstrate high performance in the predictive models

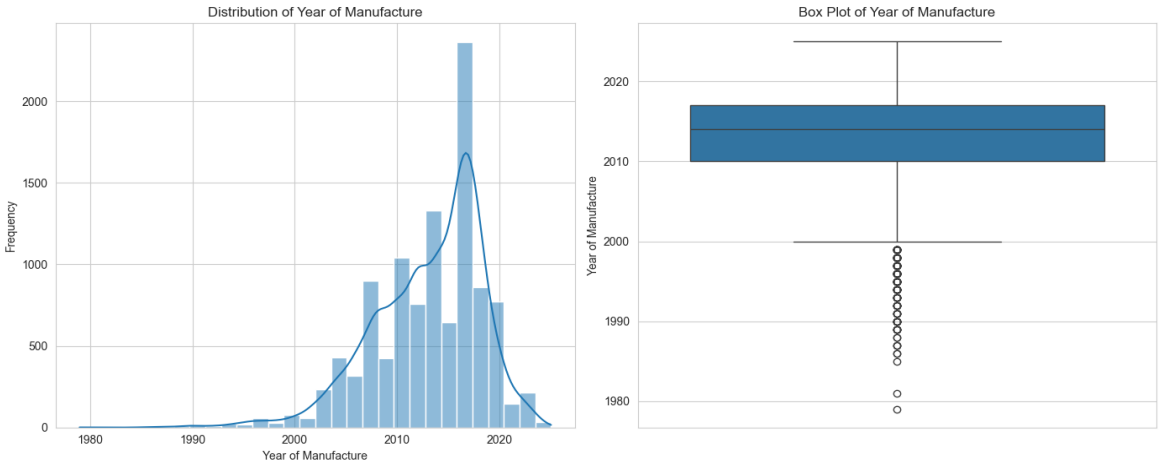


Figure 5 Distribution of Year of Manufacture pg 32

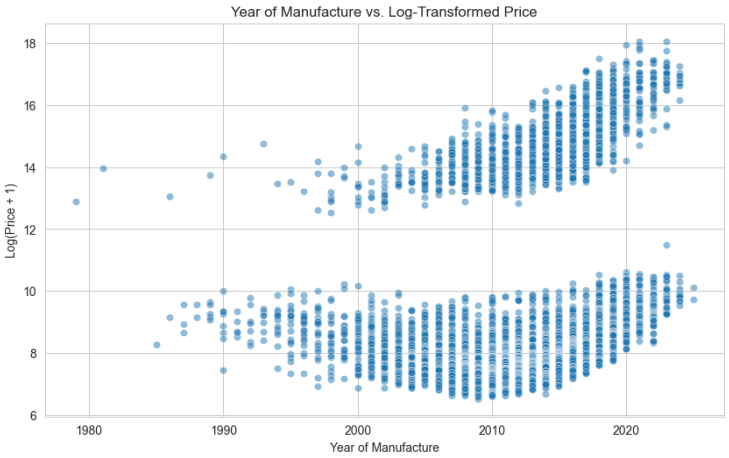


Figure 6 Year of manufacture vs Price

Besides the target variable, which is the price, several other features also showed a significant right skewness, to mention but a few mileage (skewness:3.15), engine size(skewness:5.17. horse_power (skewness:2.17) and torque(skewness:1.28). Just like the price variable, logarithmic transformations were applied to these features. The main reason for this approach was to normalize their distributions, stabilize variance, and potentially linearize their relationships with price_log, thus contributing to improved model performance

Advanced Missing Value Imputation

Straightforward imputation methods were used in the initial data cleaning efforts, like mode imputation for currency_code and mileage_unit. However, due to a substantive missingness in the annual_insurance feature during baseline modeling, it resulted in the application of a more sophisticated imputation strategy. IterativeImputer, which is a model-based imputation technique, was specifically used for annual insurance. This method operates by treating each feature with missing values as a target variable and modeling it as a function of other features in the dataset. It repetitively refines these estimates in a round-robin fashion until it converges.

4.2.4 Feature Scaling and Categorical Encoding Techniques

In order for the machine learning to work perfectly with the dataset, numerical features underwent a scaling to prevent any single feature from disproportionately influencing the learning process due to its magnitude. StandardScaler was selected for this purpose. This is a method that is used to transform data to achieve a mean of 0 and a standard deviation of 1. This method is more often used for algorithms that assume data follows a normal distribution, making it a suitable choice following the log transformations applied to skewed features.

The categorical features that are non-numeric by nature were converted into a numerical format using OneHotEncoder. This approach introduces a new binary column for each unique category within a feature, assigning a value of 1 if the data point belongs to that category and 0 otherwise. This approach is important since it avoids imposing arbitrary ordinal relationships that LabelEncoder might introduce, which can mislead algorithms into inferring ranking where none exists.

One of the most pressing problems came from the large cardinality of particular categorical features, especially model_name with 999 unique values and make_name with 54 unique values. If OneHotEncoder was applied to these features, the processed data could be expanded to 999 total features, which is an exceedingly wide feature space. The resulting data would have high dimensionality, straining computation, memory use, and increasing risks of overfitting. To address this problem without losing predictive power from these features, a specific scheme was designed. After initially applying OneHotEncoder, additional steps were taken to reduce dimensions using SelectFromModel. This method uses the feature importance from a well-performing tree-ensemble model, e.g., Random Forest, to select non-zero one-hot encoded features, then outputs only the most important features. This is expected to enable a reduction of the feature count to 500, which was achieved in this case. Such an approach is effective and practical when dealing with high-cardinality categorical features in machine learning projects.

4.2.5 Visualizations of Data Characteristics

Table 2 Descriptive Statistics for Key Numeric Features (After Initial Cleaning)

Feature	count	mean	std	Min	25%	50%(Mean)
price	12843	1.18e+06	3.39e+06	680	2680	9050
mileage	12900	1.01e+05	8.33e+04	0	46000	92000
engine_size_cc	13069	2318.39	1409.99	0	1490	1990
horse_power	13069	194.78	129.57	46	106.7	154
torque	13069	286.16	187.30	50	146.1	216.5
acceleration	13069	9.29	2.69	2.8	7.7	9.3
seats	13069	5.12	1.47	2	5	5

year_of_manufacture	10764	2013.01	5.38	1979	2010	2014

Scatter Plots of Price vs. Key Numeric Features

A prominent pattern was revealed by the scatter plots of price_log against mileage, year of manufacture, engine_size cc, horsepower, torque, and acceleration. Two horizontal bands were formed by the data points. This suggested an underlying systematic factor as a result of most likely the presence of prices which are in different currencies, that is, Ksh and USD, which were not fully standardized. This underscored the crucial need for robust currency standardization in preprocessing

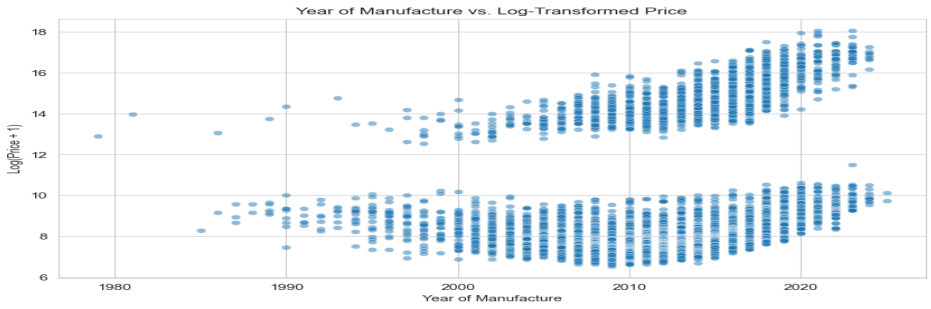


Figure 7 Year of Manufacture vs log transformed Price

The year_of_manufacture displayed a clear positive trend with price_log, with newer cars being more expensive.

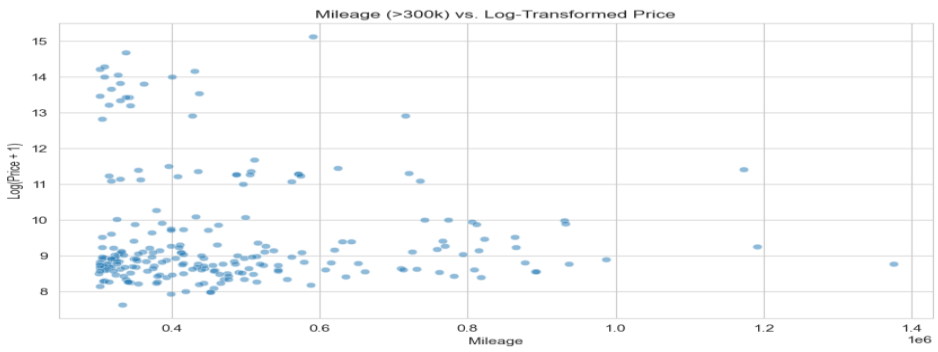


Figure 8: Mileage vs price

Mileage displayed a general negative trend with price_log, which means cars with a higher mileage are cheaper. Factors such as horsepower, engine size (cc), and torque displayed a positive trend with price_log, indicating that larger and more powerful engines are more expensive, as shown above.

Acceleration, on the other hand, displayed a negative trend with price_log, meaning that faster cars, that is, the ones with lower acceleration times, are more expensive

4.3 Model Development

This section deals with the initial training and evaluation of the regression models in order to come up with a performance benchmark with a key revelation on their strengths and weaknesses on the preprocessed dataset.

4.3.1 Overview of Selected Regression Algorithms

The following are the machine learning regression algorithms that were selected for the initial baseline evaluation

Random forest regressor: This is an ensemble method that builds multiple decision trees and is well known for its high accuracy, robustness to fitting, and feature importance scores

Gradient Boosting Regressor: This is a powerful ensemble technique that builds models sequentially, correcting the errors of its predecessors, and is well-known for its high predictive power.

K-Nearest Neighbors(KNN) Regressor: It is a non-parametric /instance–based algorithm that gives a prediction based on the average of "K" nearest neighbors

Support Vector Regressor(SVR): It is a variant of SVMs for regression, and it aims to find a hyperplane that maximizes the margin between predicted and observed values.

4.3.2 Training/Test Split and Cross-Validation

The dataset was split into 80% for training and 20% for testing using a random_state=42 in order to ensure reproducibility. This resulted in 10,773 samples used for training and 2,694 for testing across 999 features after one-hot encoding. K-Fold cross-validation was used for hyperparameter tuning and robust model evaluation with 5 folds, which ensures that the model performance estimates were not overly dependent on a single data partition

4.3.3 Hyperparameter Tuning for Optimal Performance

About baseline model evaluation, the hyperparameter tuning was performed for the algorithms: Gradient Boosting, Random Forest, Support Vector Regressor (SVR), and K-Nearest Neighbors (KNN). RandomizedSearchCV was used for the tuning process, making use of 5-fold cross-validation to ensure a robust evaluation of each parameter set. It's important to note that the tuning was conducted on the data that had undergone IterativeImputer for annual_insurance. The result of the four models demonstrated very high cross-validated R-squared scores, mostly exceeding

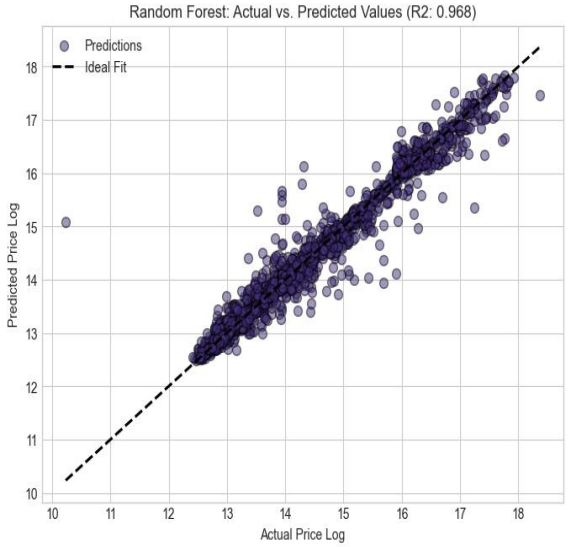
0.95, which underscored the quality of the engineered features and the effectiveness of the IterativeImputer.

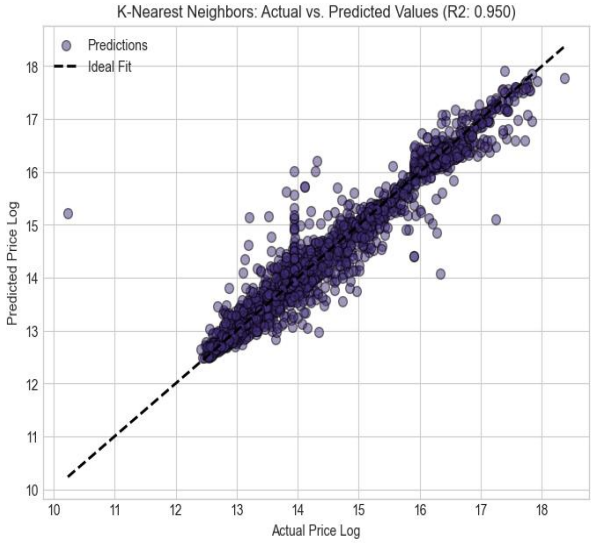
Table 3: Performance of Individual Base Models

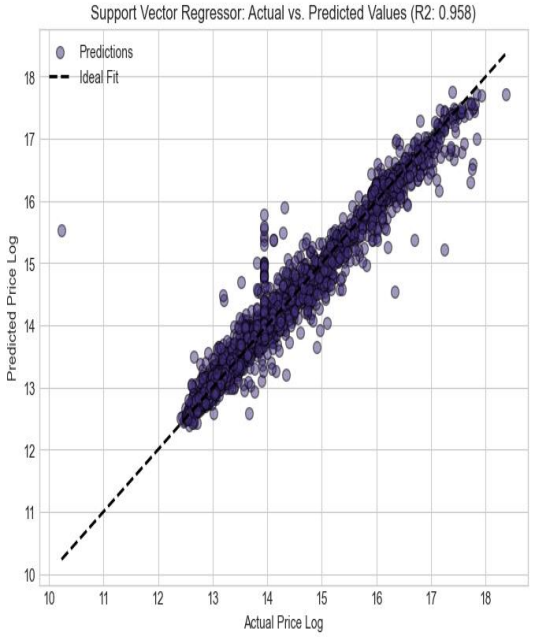
Model	Best R-squared	Key-optimal parameters
Random forest	0.968249	n_estimators=200, max_features=0.7, max_depth=20
Gradient boosting	0.936533	n_estimators=300, learning_rate=0.05, max_depth=7
Support vector regressor (SVR)	0.958412	kernel='rbf', gamma='scale', epsilon=0.01, C=10
K-Nearest neighbors	0.950422	n_neighbors=6, weights='distance', p=1

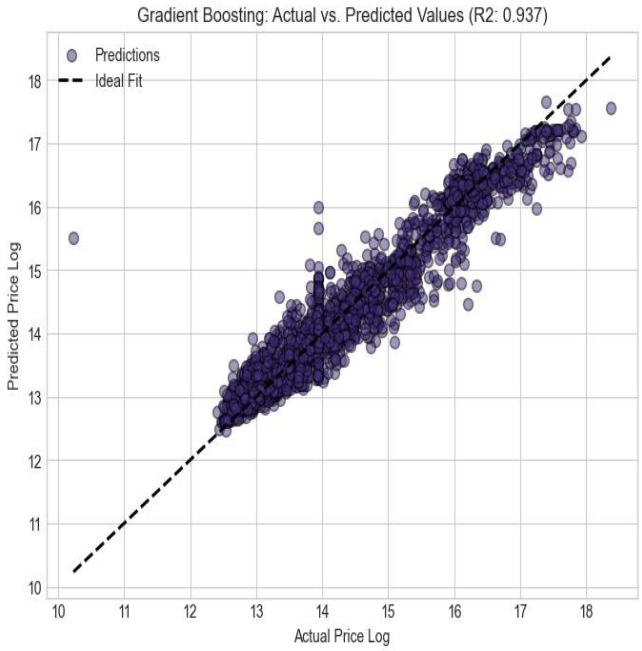
From the above table, it's clear that Random Forest performed the best with R^2 score of 0.968249, benefiting from deep trees. It is followed by Support Vector Regressor with a score of 0.958412, followed by K-Nearest Neighbors with a score of 0.950422, and finally Gradient Boosting with a score of 0.936533 following the models I had chosen for the project.

Graphs for algorithms









4.3.4 Tools and Libraries Used

The following tools were used in the development process, being products of the Python programming language:

- Pandas and NumPy for data manipulation and numerical operations.
- Scikit-learn is used for machine learning algorithms and preprocessing (OneHotEncoder, StandardScaler, and IterativeImputer), model selection (SelectFromModel, RandomizedSearchCV, and KFold), and finally ensemble methods (StackingRegressor)
- Seaborn and Matplotlib for data visualization.
- Joblib was used for saving and loading trained preprocessors and models

4.4 Model Evaluation

4.4.1 Evaluation Metrics

The following metrics were used to evaluate the performance of the models.

R-squared: This one is used to measure the proportion of variance in the dependent variable that is predictable from independent variables. 1 indicates a perfect fit, while 0 indicates no linear relationship

Mean Squared Error: This one is used to measure the average of the squared differences between actual and predicted values.

Mean absolute Error: This one is the average of the absolute differences between predicted and actual values

Root Mean Squared Error: is the square root of MSE.

4.4.2 Model Comparison Table

Table 4: Model Comparison Table

Model	R-squared	MAE	RMSE
Linear Regression	0.9115	0.2635	0.3894
Random Forest	0.968249	0.117401	0.233277
Support Vector Regressor	0.958412	0.155636	0.266979
Gradient Boosting	0.936533	0.231147	0.329814
K-Nearest Neighbors	0.950422	0.163419	0.291500

4.4.3 Advanced Ensemble Techniques: Stacking Regressor

In order to improve the accuracy and generalization, the base models were further combined in an advanced ensemble technique using a Stacking Regressor. Stacking combines the predictions of multiple diverse base models through a meta-model that learns the optimal way to integrate these predictions. This approach brings together the strengths of different models, leading to a superior performance as compared to individual models. The four models were combined and the output is as shown below.

Table 5: Ensemble Performance Table

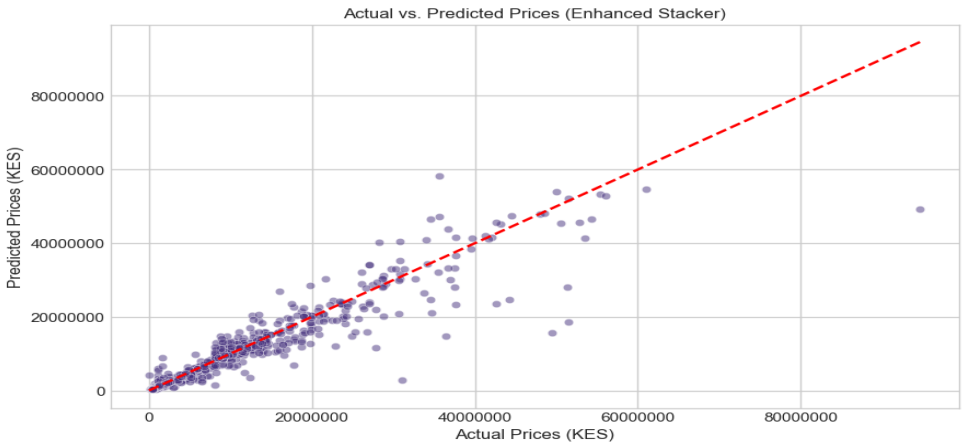
Metric	Score
R-squared (R2)	0.9725
Mean Absolute Error (MAE)	0.1137
Root Mean Squared Error (RMSE)	0.2171

4.4.4 Model Selection and Robust Evaluation

From the analysis above on the performance metrics, it's evident that the ensemble strategy produced better results as compared to individual models. This is after the Stacking Regressor achieved an R-squared (R2) score of 0.9725, a Mean Absolute Error (MAE) score of 0.1137, and a Root Mean Squared Error (RMSE) score of 0.2171, which are better than individual base models. In conclusion, the final stacking model demonstrated a very strong performance. Its high R-squared value proved a robust predictive power

Actual vs. Predicted Plot (Champion Stacking Regressor)

The scatter plot of actual versus predicted values showed points tightly clustered around the diagonal "Ideal Fit" line, visually confirming the high R^2 value. While the overall fit was excellent, there was a slight tendency for the model to under-predict some of the very highest actual prices and a single instance of significant over prediction at the very low end of the price scale.



4.5 Feature Importance Analysis

It is important to understand the features that drive a model's predictions, especially when using a complex ensemble model like a Stacking regressor. This will help in identifying and ranking the features in the order in which they contribute towards the final price prediction. In order to provide this interpretability, Permutation importance analysis was conducted on the champion model. Permutation importance is a model-agnostic technique that is used to quantify feature importance by measuring the decrease in the model's performance when a single feature's values are randomly shuffled in the test. If a feature causes a larger drop in performance when perturbed, it is considered more important. The following table shows the key features identified by permutation importance for the Stacking Regressor.

Table 6: Feature Importance score table

Rank Feature	Importance score
Mileage	0.147
Car_age	0.0897
Annual_insurance	0.0640
Usage_type (Kenyan or Foreign)	0.0404

Engine_size	0.0252
Body_type	0.0094
Car_make	0.0063

From the above analysis, it's evident that mileage, age, insurance, usage type(local/foreign used), engine size, Body_type and Car_make among others, are the best top seven features considered when purchasing a car, and this aligns with the domain knowledge of car valuation.

4.6 Discussion of Findings

The developed model has shown its strength in predicting the prices of the user cars in Kenya. It has resulted in important findings that are aligned with and expand existing literature. The individual models achieved a good predictive score, with Random Forest emerging as the best amongst the five. Their outputs were as follows: Random Forest performed the best with R2 score of 0.968249, benefiting from deep trees. It was followed by Support Vector Regressor with a score of 0.958412, followed by K-Nearest neighbors with a score of 0.950422, Gradient Boosting with a score of 0.936533 and finally Linear regression with an R2 value of 0.9115. These scores were, however, combined into an ensemble using a Stacking approach. The developed model, as a result of the combination, outshone all of them by achieving a score of R-squared of 0.9725, a Mean Absolute Error (MAE) score of 0.1137, and a Root Mean Squared Error (RMSE) score of 0.2171.

The MAE value of 0.1137 means that, on average, the model's predictions deviate from the actual car prices by about 0.1137 units in the scale of the log-transformed target variable. When converted back to the real car price scale, this represents a very small average error, indicating that the model makes highly accurate predictions.

Similarly, the RMSE of 0.2171 indicates that the typical size of the prediction error is 0.2171 units in log-price scale. RMSE penalizes large errors more than MAE, so a low value shows that the model rarely makes large mistakes.

Because both error values are close to zero, this confirms that the ensemble model predicts used car prices with very high precision.

This indicates that it explains over 97% of the used car prices. This high accuracy furthermore has outshone some existing research, for instance, KNN (85% accuracy) by Samruddhi and Kumar (2020), Random forest (95% R2) by Alshared (2021). This is a reference to individual models. However, about ensemble models, this model performed better than some models like Gegic et al (2019), who achieved 87.38% accuracy, and CatBoost, which achieved 86.05% accuracy in Bishkek. This model demonstrates a superior performance, which directly addresses the limited use of ensemble methods as identified in the literature review.

On the other hand, the feature importance analysis revealed key features that play a crucial role in the determination of the price of a used car in Kenya. The mileage, Car_age, Annual_insurance, usage_type (Kenyan or Foreign), engine_size, body_type, and car_make emerged as among the best seven features with the highest score, as demonstrated earlier. This one addresses the objective of identifying the key features that are used to predict the prices of used cars in Kenya. This also addresses the gap in feature engineering as mentioned by Chandak et al (2019), which emphasized the addition of more features like horsepower, torque, insurance, and many more. The performance metrics were well used, which explained the predictive nature of the ensemble model. In conclusion, the Stacked model performed better in the prediction by achieving an R-squared score value of 0.9725, which is 97% accurate than compared to individual models.

CHAPTER FIVE

5. DISCUSSION OF FINDINGS, CONCLUSIONS, AND RECOMMENDATIONS

5.1 Discussion of findings

The major aim of this study was to come up with an advanced data-driven model that is an ensemble to predict the prices of used cars in Kenya. The research incorporated several data science techniques, including data preprocessing, feature engineering, baseline model evaluation, hyperparameter tuning, and ensemble learning, all aimed at enhancing predictive performance. The dataset used in the research was obtained through web scraping under an authorizing license issued by the National Commission of Science, Technology and Innovation (NACOSTI). It consisted of 13,667 records that underwent rigorous preprocessing, including handling missing values using Iterative Imputation, applying logarithmic transformations, encoding categorical features through OneHotEncoding, and performing feature selection to reduce high dimensionality.

5.1.1 Factors influencing used car prices

The study began by identifying the key factors influencing used car prices. Feature importance analysis using the permutation method revealed that mileage, car age, annual insurance, usage type (Kenyan or Foreign), and engine size were the most influential attributes in determining the price of a used car in Kenya. These findings highlighted the real-world variables that significantly contribute to price variation in the Kenyan automotive market.

5.1.2 Developing of the ensemble model

The research further involved building several baseline machine learning models, including Random Forest, Support Vector Regressor, K-Nearest Neighbors, Gradient Boosting, and Linear Regression. These models were trained and evaluated using R^2, Mean Absolute Error (MAE), and

Root Mean Squared Error (RMSE) as performance metrics. Random Forest emerged as the best-performing individual model with an R^2 score of 0.968249, followed by Support Vector Regressor with 0.958412, K-Nearest Neighbors with 0.950422, Gradient Boosting with 0.936533, and finally, Linear Regression with an R^2 of 0.9115.

To improve prediction accuracy and generalization, the four best individual models were combined using a stacking ensemble strategy. The stacked ensemble model outperformed all the baseline models, achieving an R^2 score of 0.9725, an MAE of 0.1137, and an RMSE of 0.2171. These performance values indicate that the ensemble model explains 97.25% of the variation in used car prices, with small average prediction errors. Specifically, the MAE value shows that the model's predictions differ only slightly from actual values on average, while the low RMSE value reflects that large prediction errors are rare.

Overall, the study successfully developed and validated a high-performing ensemble machine learning model for predicting used car prices in Kenya. It identified the key predictive factors, built and tested multiple models, and demonstrated that the ensemble approach provides the highest accuracy and reliability when evaluated using appropriate metrics.

5.2 Conclusion

In conclusion, the study developed an ensemble model for predicting the prices of used cars. Data was obtained from Kenya's online Motor Bazaars, and it was preprocessed to ensure it is fit for the study. The study found out that mileage, car age (year of manufacture), annual insurance, usage type (Kenyan/Foreign), engine size, body type, and car make were the most important features. Four models were combined in an ensemble strategy using a Stacking regressor: Random Forest, Support Vector Machines, K-Nearest Neighbors, and Gradient Boosting, which was validated using real-world data obtained from Kenya online motor bazaars. The developed champion model

demonstrated a superior performance by achieving an R^2 score of 0.9725, an MAE of 0.1137, and an RMSE of 0.2171. This one provides a practical artifact that can be used to predict the prices of used cars in Kenya.

5.3 Limitations and Delimitations

The study was limited to financial budget constraints and notable inconsistencies with the data. However, the online method of data collection, which is web scraping, was used, and the data underwent a series of preprocessing stages to reduce inconveniences.

5.4 Recommendations

The industry stakeholders and policy makers should support and spearhead the use of data-driven solutions, thus utilizing a standardized tool in the automotive industry for consistency and informed decision-making when pricing cars. Car dealers are also encouraged to make use of predictive data-driven tools to ensure fairness and competitive pricing of cars. Finally, future researchers can expand on the work by integrating real-time features like currency exchange rates and economic trends, and the use of multimodal models, and also do comparisons between the stacked ensemble approach with advanced deep learning models like artificial neural networks

5.5 Future work

Future researchers can incorporate natural language processing for interpreting user-submitted descriptions of car conditions, come up with a web application powered by the model to assist consumers in getting instant price predictions, and discover more features that can be on boarded to improve the accuracy of the model.

REFERENCES

Africa, O. (2024). Surging Prices and Shifting Trends: The Challenges Facing Kenya's Second-Hand Car Market. OfficePhase Africa. https://officephase.com/2024/02/12/surging-prices-and-shifting-trends-the-challenges-facing-kenyas-second-hand-car-market/#:~:text=The%20Kenya%20Auto%20Bazaar%20Association,diminished%20purchasing%20power%20of%20households.

Africa, O. (2024). Surging Prices and Shifting Trends: The Challenges Facing Kenya's Second-Hand Car Market. OfficePhase Africa. https://officephase.com/2024/02/12/surging-prices-and-shifting-trends-the-challenges-facing-kenyas-second-hand-car-market/

Ahtesham, M., & Zulfiqar, J. (2022). Used Car Price Prediction with Pyspark. In Lecture notes in networks and systems (pp. 169–179). https://doi.org/10.1007/978-3-031-01942-5_17

Amik, F. R., Lanard, A., Ismat, A., & Momen, S. (2021). Application of machine learning techniques to predict the price of Pre-Owned cars in Bangladesh. Information, 12(12), 514. https://doi.org/10.3390/info12120514

AnalytixLabs. (2023, December 26). Random Forest Regression — How it Helps in Predictive Analytics? Medium. https://medium.com/@byanalytixlabs/random-forest-regression-how-it-helps-in-predictive-analytics-01c31897c1d4

Bhatnagar, P., Lokesh, G. H., Shreyas, J., Flammini, F., & Gautam, S. (2024). An Analysis of Car Price Prediction using Machine Learning. An Analysis of Car Price Prediction Using Machine Learning, 4, 11–15. https://doi.org/10.1145/3674029.3674032

Blog Post | Peach Cars. (n.d.). https://peachcars.co.ke/blog/challenges-of-selling-your-car-in-kenya-and-how-to-overcome-them

Budiono, D. A., Utomo, K. S., Wibowo, K. J., & Wiradinata, M. J. (2024). Used Car Price Prediction Model: A Machine Learning approach. Budiono | International Journal of Computer and Information System (IJCIS). https://doi.org/10.29040/ijcis.v5i1.147

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

Climate Change 2022: Impacts, adaptation and Vulnerability. (n.d.). IPCC. https://www.ipcc.ch/report/ar6/wg2/

GeeksforGeeks. (2024a, February 26). Regression in machine learning. GeeksforGeeks. https://www.geeksforgeeks.org/regression-in-machine-learning/

GeeksforGeeks. (2024b, July 29). Regression metrics. GeeksforGeeks. https://www.geeksforgeeks.org/regression-metrics/

Han, L., Yin, Z., Xia, Z., Tang, M., & Jin, R. (2020). Price Suggestion for Online Second-hand Items with Texts and Images. Proceedings of the 30th ACM International Conference on Multimedia, 4, 2784–2792. https://doi.org/10.1145/3394171.3413759

Hao, L., Umar, M., Khan, Z., & Ali, W. (2020). Green growth and low carbon emission in G7 countries: How critical the network of environmental taxes, renewable energy and human capital is? The Science of the Total Environment, 752, 141853. https://doi.org/10.1016/j.scitotenv.2020.141853

Hargrave, M. (2020, November 18). Hedonic pricing: definition, how the model is used, and example. Investopedia. https://www.investopedia.com/terms/h/hedonicpricing.asp

Kenya used car market size | Mordor Intelligence. (n.d.). https://www.mordorintelligence.com/industry-reports/kenya-used-car-market

Mobility Foresights. (2024a, January 22). Kenya Used Car market 2024-2030. https://mobilityforesights.com/product/kenya-used-car-market/

Mobility Foresights. (2024b, January 22). Kenya Used Car market 2024-2030. https://mobilityforesights.com/product/kenya-used-car-market/

News Desk. (2022, January 5). Used Car imports up in Kenya. Trendsnafrica | 24/7 Africa News. https://trendsnafrica.com/used-car-imports-up-in-kenya/

Pal, N., Arora, P., Kohli, P., Sundararaman, D., & Palakurthy, S. S. (2018). How much is my car worth? A methodology for predicting used cars' prices using Random Forest. In Advances in intelligent systems and computing (pp. 413–422). https://doi.org/10.1007/978-3-030-03402-3_28

Pillai, A. S. (2022, June 16). A deep learning approach for used car price prediction. Journal of Science & Technology. https://www.thesciencebrigade.com/jst/article/view/140

Prediction of Prices Car Price Prediction with Machne Learning. (2022, October 6). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9995772?casa_token=25vLVXNinokAAAAA:n7PAYqIyFcNtMSN7_Gwr8ELH0UZUrozxa3qWrBXZZ23q7b8pUJQ9NoeU9oEfZ0LdB6HzWeQQzD93oSrc

Prediction of used car prices using artificial neural networks and machine learning. (2022a, January 25). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9740817

Prediction of used car prices using artificial neural networks and machine learning. (2022b, January 25). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9740817/references#references

Price prediction of used cars using machine learning. (2021, November 22). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9696839?casa_token=9f4-JFiGMmQAAAAA:t-SChgZjsuVlUeR0MgBqPNQa8i4FSYoO40Mj1wGFxyUwvn1xYsGnWRpuZLCQ008eD1C1Ia6gcBtaIVHr

Tang, J., Gong, R., Wang, H., & Liu, Y. (2023). Scenario analysis of transportation carbon emissions in China based on machine learning and deep neural network models. Environmental Research Letters, 18(6), 064018. https://doi.org/10.1088/1748-9326/acd468

Used Car Price Prediction using Machine Learning: A Case Study. (2022, May 18). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9800719?casa_token=HHOMbnUT60kAAAAA:LFcinEBfGYrN2NtKl_VLFyVPpOhYglEfaNbASBS4534mm2--0YoT1vgo0d4nFUxeCmBsH_Uboj4400Kr

Wikipedia contributors. (2024, October 2). Random forest. Wikipedia. https://en.wikipedia.org/wiki/Random_forest

Yu, W., Xia, L., & Cao, Q. (2024). A machine learning algorithm to explore the drivers of carbon emissions in Chinese cities. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-75753-y

Doctor, A. (2024, October 28). The evolution of car technology in Kenya: Past, present, and future. Garage in Nairobi | Automotive Doctor Motor Garage Limited. https://automotivedoctor.co.ke/the-evolution-of-car-technology-in-kenya-past-present-and-future/

APPENDICES

Appendix A: Article

Science Frontiers
2025, Vol. 6, No. 3, pp. 96-105
https://doi.org/10.11648/j.sf.20250603.15

Research Article

Ensemble Machine Learning Model for Predicting Prices of Used Cars in Kenya

Moses Onserio[1, *], Fidelis Mukudi[2]

[1]Department of Computer Science and Information Technology, Co-operative University of Kenya, Nairobi, Kenya
[2]Department of Mathematical Sciences, Co-operative University of Kenya, Nairobi, Kenya

Abstract

Most Kenyan car owners prefer used vehicles due to their affordability, leading to a booming used car market. However, the absence of an objective pricing mechanism has led to inconsistent and subjective pricing, with prices varying significantly from seller to seller. This research aimed to provide a data-driven solution by incorporating key vehicle attributes. Using Design Science Research (DSR) methodology, the research implemented machine learning techniques: Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Gradient Boosting, Linear regression as base models, and Permutation for feature explanation to enhance accuracy and interpretability. The individual models were trained and evaluated using 5 cross-validation. Random Forest emerged as the best with a Mean Absolute Error of 0.1174, and Linear regression was the last with a Mean Absolute Error of 0.2635. For performance optimization, the four best baseline models (RF, SVM, KNN, and GB) were combined using a Stacking Regressor, which achieved an R-squared score of 0.9725, a mean absolute error (MAE) of 0.1137, and a mean squared error (MSE) of 0.2171, showing an improved predictive performance compared to individual models. Feature importance analysis identified mileage, car age, annual insurance, engine size, and usage type (Kenyan/Foreign) as the most influential features.

Keywords

Used Car Market, Machine Learning, Price Prediction, Used Car Valuation

1. Introduction

The automotive industry is a key sector that contributes significantly to the growth of the economy. In Kenya, the industry has experienced significant growth. Playing a key role in the economy, it requires a reliable and consistent pricing mechanism that can be used for various purposes such as reselling, insurance, accounting, and leasing, among others. The existing pricing systems have consistently exhibited inconsistent and varying prices due to overreliance on subjective pricing. With the increased demand for used cars, there is a need for a standardized pricing tool that is data-driven [1, 2].

Significant progress has been made in the application of machine learning technology globally in the automotive industry. However, in Kenya, a unique challenge exists: the absence of a data-driven and standardized tool that can be used to predict the prices of used cars. Most existing studies focused on the use of single machine learning algorithms and partial variables and markets, particularly in Europe, India,

*Corresponding author: jonserio@cuk.ac.ke (Moses Onserio)

Received: 31 July 2025; **Accepted:** 12 August 2025; **Published:** 28 August 2025

Appendix B: NACOSTI

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **432889**

Date of Issue: **20/May/2025**

RESEARCH LICENSE

This is to Certify that Mr.. Moses Onserio Onserio 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: ENSEMBLE MACHINE LEARNING MODEL FOR PREDICTING USED CAR PRICES IN KENYA. for the period ending : 20/May/2026.

License No: **NACOSTI/P/25/4174315**

432889

Applicant Identification Number

Deputy Director
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