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 R². Linear Regression performed poorly, while ensemble models produced stronger results. Random Forest achieved a testing R² of 0.816 with an MAE of Ksh 683,303, XGBoost reached a testing R² of 0.837 with an MAE of Ksh 672,930, and a Voting Ensemble combining both models performed best, with a testing R² 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and often unreliable, as traditional valuation methods rely heavily on manual appraisals, dealer experience, or outdated reference books.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 R² 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 R² on test data.

Table 2. Model evaluation results

Model	Train R²	Test R²	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