iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80879

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Second Hand Car Price Prediction using Machine Learning

Prof. Aman Singh, Prof. Rohan B Kokate, Joel Louis

Abstract: ***The rapid growth of the second-hand car market has created a strong need for accurate, automated, and transparent vehicle pricing systems. Platforms like CarDekho, Cars24, and Spinny have increased demand for reliable valuations, but estimating resale prices remains complex due to factors such as brand, model, age, fuel type, mileage, and market demand. Traditional methods based on expert judgment are often subjective and inconsistent.***

This study uses machine learning to address these challenges by developing models such as Linear Regression, Decision Tree, Random Forest, and XGBoost to predict used car prices. A user-friendly web application built with Streamlit allows users to input car details and receive instant price estimates.

The methodology includes data collection (from sources like Kaggle and CarDekho), preprocessing, feature engineering, model training, and evaluation using MAE, RMSE, and R^2. Using a dataset of 6,019 records, results show that ensemble methods—especially XGBoost—perform best, achieving an R^2 of 0.94. The deployed application demonstrates scalability and practical usefulness.

Future work can include computer vision for image analysis, NLP for text data, and Explainable AI for better model transparency.

Keywords: used car price prediction, machine learning, XGBoost, Random Forest, regression, feature engineering, Streamlit, automobile valuation.

I. INTRODUCTION

A. Background and Motivation (Shortened)

The global automobile industry is a major economic sector, with the second-hand car market growing rapidly due to rising new car prices, increased awareness, and digital platforms. In countries like India, used car sales are nearly equal to new car sales, showing a shift toward cost-effective and sustainable buying choices.

Accurate vehicle valuation is a key challenge, as used car prices depend on many factors such as age, mileage, ownership, condition, and market trends. Traditional methods based on dealer expertise or comparisons are often subjective, inconsistent, and unable to adapt to changing market conditions. This can impact consumer trust, dealer profits, and financial decisions like insurance and loans.

Machine learning offers a powerful solution by analyzing large datasets to capture complex relationships and provide accurate, consistent price predictions. This study aims to develop a complete machine learning pipeline for predicting second-hand car prices.

B. Problem Definition

The global automobile industry is a major economic sector, with the second-hand car market growing rapidly due to rising new car prices, increased awareness, and digital platforms. In countries like India, used and new car sales are nearing parity, reflecting a shift toward more cost-effective and sustainable consumer choices.

Accurate valuation of used cars remains a challenge, as prices depend on multiple factors such as age, mileage, ownership, condition, location, and market trends. Traditional methods based on dealer expertise or comparisons are often time-consuming, subjective, and unable to adapt to market changes, leading to pricing inconsistencies and impacts on trust, profitability, insurance, and financing.

Machine learning provides a data-driven solution by analyzing large datasets to capture complex relationships and deliver accurate, consistent price predictions. This study leverages these advancements to design, implement, and evaluate a comprehensive machine learning system for second-hand car price prediction.

©IJRASET: All Rights are Reserved

C. Research Objectives

The specific objectives of this research are listed as follows. First, the aim is to collect and explore real-world used car datasets that are available on Kaggle and CarDekho and to preprocess these datasets while addressing issues that include missing values, outliers, and inconsistent formatting. Second, the goal is to perform systematic feature engineering which includes creating derived features like car age, brand popularity index, and normalized mileage that can enhance the predictive power of the models. Third, the intention is to implement, train, and compare four supervised machine learning models which are Linear Regression, Decision Tree, Random Forest, and XGBoost. Fourth, the evaluation of model performance will be conducted using standard regression metrics which include MAE, RMSE, and R2, along with conducting feature importance analysis for the purpose of model interpretability. Fifth, there is a plan to deploy the best-performing model in a user-facing web application that uses Python's Streamlit framework, which will enable real-time price prediction for end users. Lastly, the research will identify limitations of the current approach and will propose meaningful directions for future research.

D. Significance of the Research

This research has implications that are broad and affect multiple groups of stakeholders. For consumers, a tool that values items based on objective data allows them to make purchasing and selling decisions with more confidence, which reduces the information imbalance that has historically benefited dealers. For automotive dealers and those involved in remarketing, the system allows for quicker assessments of trade-ins and enables pricing strategies that are more competitive. For financial institutions that provide auto loans and insurance products, having accurate valuation models helps in assessing risk and improving accuracy in underwriting processes. For digital platforms like CarDekho and Cars24, incorporating predictive pricing into their listing systems can lead to increased engagement, trust, and a higher volume of transactions. Lastly, for those in the research community, this work offers a methodology that can be reproduced for applied machine learning specifically in the automotive field.

E. Scope and Limitations

This study is about the Indian second-hand car market. It uses datasets that were put together from CarDekho and Kaggle. The focus includes structured tabular data that has information about vehicle specifications, ownership history, and transaction prices. However, image-based analysis, processing of free-text reviews, and integration of real-time market APIs are not included in this dissertation but are noted as important areas for future work. The dataset is limited to records that are available up to the year 2024, and the predictive accuracy for listings that are more recent might need the model to be retrained with data that is updated.2.

II. LITERATURE REVIEW

A. Early Approaches to Vehicle Valuation

Early scholarly work that focused on predicting automobile prices relied on classical econometrics. Hedonic Pricing Models were the main analytical framework used. In these models, the price of a vehicle was expressed as a linear combination of its attributes, with each attribute having a coefficient that indicated its marginal contribution to the overall value. While these models were mathematically manageable, they had strict assumptions about linearity and independence of features that were often not met in real-world situations. The models could not effectively account for interaction effects, nonlinear depreciation curves, and hierarchies of categorical features, which reduced the practical accuracy of early regression-based methods.

In the 1990s and 2000s, the publication of important regression texts and the increasing availability of digitized transaction records led to a gradual shift towards more data-intensive methods. Researchers started to include more complex feature sets, such as indicators of brand equity and proxies for regional demand, but the basic limitation of linear modeling continued until machine learning techniques became widely adopted.

B. Machine Learning Approaches

The use of machine learning for used car price prediction has grown significantly, with studies showing that ensemble and boosting methods outperform traditional models. Cui et al. (2023) proposed an XGBoost framework with dynamic feature selection, achieving an R^2 of 0.93 and identifying brand, model year, and mileage as key features.

Patel et al. (2022) compared multiple models on Indian datasets and found Random Forest to offer the best balance of accuracy and generalization due to reduced variance. Zhu (2023) used SHAP analysis to show that brand reputation and mileage explain over 70% of price variation, highlighting the importance of feature interpretability.

Uluturk (2021) found Random Forest more accurate and interpretable than SVR, emphasizing the need for explainable models. Chen et al. (2024) introduced a multimodal approach combining tabular and image data, improving accuracy but limiting real-world use due to image requirements.

A Stanford CS230 (2023) project showed that deep learning did not outperform XGBoost on structured data, reinforcing the strength of tree-based methods. Mallick et al. (2022) highlighted the impact of transmission and fuel type, while AlShared (2021) identified model year and car age as universally strong predictors across markets.

C. Web Deployment and Practical Systems

A critical but often not thoroughly examined aspect of the literature is how to turn trained models into systems that users can interact with. Marnholkar in 2025 showed how a used car pricing model was put into use through a web interface and pointed out that responsive design, input validation, and user experience are important for getting people to use it. Practical systems that are found in GitHub repositories from 2024 confirm that Flask and Streamlit are the frameworks most often used for deploying machine learning models as web applications because they are easy to learn and work well with Python.

D. Research Gaps

There is a large amount of existing work, but there are still several important gaps that are present. First, many studies utilize datasets that are either specific to certain regions or taken from only one platform, which restricts how widely their findings can be applied. Second, there are ongoing concerns about how to interpret complex models like XGBoost, especially in situations where the stakes are high, such as in insurance underwriting. Third, very few studies have taken the step to integrate their predictive models into web applications that are fully functional and accessible to the public, and that also use production-quality code, which means the practical impact of academic research is limited. Fourth, the use of image-based features and natural language descriptions of vehicles as additional inputs has not been thoroughly explored in the context of the Indian market. This dissertation aims to directly tackle the third gap by fully deploying the predictive model within a web application that is based on Streamlit.

III. RESEARCH METHODOLOGY

A. Data Collection

The dataset used in this study was collected from Kaggle's CarDekho Vehicle Dataset and the CarDekho Data Portal, comprising 6,019 used car listings. It includes key features such as car name, OEM, model, year of manufacture, kilometers driven, fuel type, transmission, number of owners, body type, city, variant, mileage, seating capacity, and selling price (target variable in INR).

Exploratory Data Analysis (EDA) was performed using Python libraries like pandas and matplotlib to understand feature distributions, detect missing values, and identify outliers. Visualizations such as histograms, box plots, and correlation matrices were used to guide data preprocessing decisions.

B. Data Preprocessing

Data preprocessing is one of the most important phases in the machine learning pipeline because the quality of the input data has a direct influence on the maximum performance that the model can achieve.

1) Handling Missing Values

Columns that had missing values were dealt with by using imputation methods that matched the distribution of each feature. For continuous numeric features like mileage and engine capacity, the median value of the respective column was used for imputation, which is a method that offers some robustness in cases of skewed distributions. For categorical features including fuel type and transmission, the mode value was used for imputation. Records that had missing values in the target variable, which is selling price, were removed completely since these records could not be used effectively for supervised learning purposes.

2) Outlier Removal

Outliers in the numeric features were identified and removed. This was particularly the case for the price and kilometers driven columns. The method used was the Interquartile Range method. Values that were below Q1 minus 1.5 times IQR or above Q3 plus 1.5 times IQR were considered anomalous and were excluded from the training dataset. This process led to the elimination of around 4.3 percent of records. It also reduced the variance in the target variable and had an effect on improving model convergence.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

3) Feature Encoding

Categorical variables that consist of fuel type, body type, transmission type, OEM, model, variant name, and city were encoded by using Label Encoding through the LabelEncoder class from scikit-learn. The choice of label encoding was made instead of one-hot encoding to prevent the large increase in dimensions that would occur due to the high cardinality of the OEM, model, and variant name categories. The encoder objects that were serialized were saved to disk with joblib to make sure that the encoding remains consistent during both the training and inference phases.

4) Feature Scaling

Continuous numeric features such as kilometers driven, number of previous owners, and model year underwent normalization through Min-Max Scaling which transformed each feature into the range of 0 to 1. This normalization is important for algorithms like Linear Regression that have sensitivity to the scale of input features. Scaler objects were serialized in a similar manner to allow for consistent application during prediction..

5) Train-Test Split

The dataset that was preprocessed was divided into a training set that comprised 70 percent and a testing set that comprised 30 percent using the train_test_split function from scikit-learn with a fixed random seed of 42 which was used as random_state to ensure that results can be reproduced. The stratified split method was employed to maintain the distribution of important categorical variables across both subsets.

C. Feature Engineering

Feature engineering was executed to enhance the original set of features by adding derived variables that capture insights specific to the domain which are not directly available in the raw data..

1) Car Age

Car age was calculated by taking the difference between the reference year which is 2024 and the model year which gives the formula Car Age equals 2024 minus Model Year. This feature shows the total depreciation of a vehicle over time and it was noted to have a strong correlation with the selling price of the vehicle.

2) Brand Popularity Index

A Brand Popularity Index was made by finding the average selling price for each Original Equipment Manufacturer across the full dataset which is represented as Brand Popularity equals Mean Selling Price per OEM. This index acts as a stand-in for brand value and the extra value that comes with well-known manufacturers.

3) Normalized Mileage

A normalized mileage feature was created by dividing fuel efficiency measured in kilometers per liter by car age which is expressed as Mileage Normalized equals Mileage divided by Car Age. This feature represents the efficiency per year and it is considered a better predictor of remaining mechanical value than just using raw mileage by itself.

D. Machine Learning Algorithms

1) Linear Regression

Linear Regression is used as the basic model for predicting the target variable through a linear combination of input features. The formula for this is y equals beta_0 plus beta_1 times x_1 plus beta_2 times x_2 plus and so on until beta_n times x_n plus epsilon, where y is the predicted price, x_i are the input features, beta_i are the coefficients learned through regression, and epsilon stands for the error term that cannot be reduced. Ordinary Least Squares estimation is applied to reduce the sum of the squared differences between observed and predicted values. Linear Regression is interpretable, but it has limitations in that it cannot handle nonlinear relationships or interactions between features.

2) Decision Tree Regressor

The Decision Tree Regressor divides the feature space into rectangular areas by using a series of binary splits. At each node, the split is chosen to decrease the impurity of the child nodes which is measured by using variance reduction for regression tasks. The tree that results from this process is easy to understand but it can overfit the training data if it is allowed to grow to its full depth. To reduce overfitting, pruning techniques were applied using the max_depth and min_samples_split hyperparameters.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

3) Random Forest Regressor

Random Forest is a bagging-based ensemble method that combines predictions from multiple decision trees trained on different bootstrap samples, with random subsets of features considered at each split. The final prediction is the average of all trees, which reduces variance while maintaining low bias, resulting in strong generalization. XGBoost (Extreme Gradient Boosting) is a boosting technique that builds trees sequentially, where each new tree learns from the errors of the previous ones. It minimizes an objective function consisting of a loss term and a regularization term, and uses L1 and L2 regularization along with row and column sampling to prevent overfitting and improve efficiency, making it highly effective for structured tabular data.

IV. SYSTEM ARCHITECTURE AND IMPLEMENTATION

A. System Overview

The system is composed of four main components, which are a data preprocessing and feature engineering pipeline, a module for model training and evaluation, a layer for model serialization, and a web application that users interact with. These components work together within a software architecture that facilitates the movement of data from raw input provided by users to a processed output that predicts prices. The data flow can be described as follows: users input vehicle specifications through a web interface built with Streamlit; this input data is then processed through a pipeline that includes steps like label encoding, scaling, and feature derivation; the resulting feature vector is used with the serialized XGBoost model; finally, the predicted price is shown in an output panel that is formatted. This architecture is designed to ensure that the inference process applies the same transformations as those used during the model training phase, which helps prevent any data leakage and ensures consistent predictions.

B. Technology Stack

The implementation uses several technologies and libraries, including the following:

- Python 3.10 serves as the main programming language for all components of the pipeline.
- Pandas and NumPy are used for data manipulation, preprocessing, and numerical computations.
- Scikit-learn provides preprocessing utilities like LabelEncoder and MinMaxScaler, as well as implementations for models such as LinearRegression, DecisionTreeRegressor, and RandomForestRegressor, along with evaluation metrics.
- XGBoost is the library utilized for the gradient boosting model that serves as the primary predictive model.
- Joblib is used for serializing and deserializing the trained model artifacts.
- Streamlit is the framework for developing the web application, allowing for the creation of interactive user interface components directly from Python code without needing HTML, CSS, or JavaScript.
- Matplotlib and Seaborn are used for exploratory data analysis and visualizing feature importance..

C. Streamlit Web Application

- The web application was developed using Streamlit, an open-source Python framework that enables easy conversion of data scripts into interactive web apps. It includes a sidebar where users input vehicle details in a logical, step-by-step manner. A cascading filter system ensures that dropdown options are dynamically updated based on previous selections (e.g., selecting an OEM filters available models), allowing only valid feature combinations and reducing prediction errors.
- Users provide inputs such as OEM, model, body type, fuel type, transmission, seating capacity, variant, manufacturing year, ownership history, kilometers driven, mileage, and city. Upon clicking the Predict button, the app converts inputs into a structured format, applies saved preprocessing steps, and generates a price prediction using the XGBoost model.
- The interface is enhanced with custom CSS, featuring a styled sidebar and a distinct output panel where the predicted price is displayed prominently along with additional insights like car age and normalized mileage score.

D. Model Serialization and Deployment

After training was completed, three model artifacts were saved using joblib, which included the trained XGBoost regression model, a dictionary containing fitted LabelEncoder objects for each categorical column, and a dictionary of fitted MinMaxScaler objects for each numeric column that was scaled. These artifacts are loaded when the application starts and are reused for all prediction requests, which allows for efficient inference without needing to retrain the model. The method of serialization allows for easy model versioning because replacing the serialized .pkl files with newer artifacts will automatically update the behavior of the application in terms of predictions without needing to change any code.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

E. Input Validation and Error Handling

The application includes input validation to check for and identify any missing or null input values before proceeding with the prediction process. If any required fields are missing, the application will highlight which columns are incomplete and will display an informative error message to the user. A try-except block is used around the model inference call to catch and report any runtime exceptions that may occur due to unexpected combinations of input, which helps maintain a smooth user experience even in situations that are not typical.

V. RESULTS AND DISCUSSION

A. Exploratory Data Analysis

The exploratory data analysis uncovered several significant patterns within the dataset. The distribution of selling prices was found to be right-skewed, with a long tail that included high-value luxury vehicles. The median selling price was approximately INR 450,000. A negative correlation between car age and selling price was identified, with a correlation coefficient of r equals -0.61, which confirms the expected trend of depreciation. The number of kilometers driven also showed a moderate negative correlation with price, with a coefficient of r equals -0.43. Among the categorical features, diesel vehicles had higher average prices compared to petrol vehicles, which may be attributed to their common presence in larger and more premium vehicle segments. Vehicles with automatic transmissions were found to command a significant price premium over those with manual transmissions, reflecting the growing consumer preference for convenience.

Brand popularity, measured by the average selling price per OEM, varied greatly across different manufacturers. Premium European brands such as Mercedes-Benz and BMW had the highest indices of brand popularity, while more common domestic brands were found at the lower end of the scale. This feature was determined to be one of the most informative predictors during the model training phase.

B. Model Performance Comparison

The four models were trained using the 70 percent of the data designated for training and were tested on the remaining 30 percent of the data that was held out for testing. The evaluation of performance involved three different metrics which are Mean Absolute Error which is often abbreviated as MAE, Root Mean Squared Error which is referred to as RMSE, and the Coefficient of Determination which is commonly known as R2. The results of the comparison are summarized in Table 1.

Table 1

Model	MAE	RMSE	R2
Linear Regression	0.35	0.48	0.78
Decision Tree	0.28	0.36	0.86
Random Forest	0.19	0.24	0.92
XGBoost	0.16	0.21	0.94

Comparative Performance of Machine Learning Models for Used Car Price Prediction

Note. MAE stands for Mean Absolute Error. RMSE stands for Root Mean Squared Error. R2 stands for Coefficient of Determination. All the error metrics are measured in normalized price units. Higher values of R2 indicate that there is better explanatory power.

C. Discussion of Results

1) Linear Regression

Linear Regression, which serves as a baseline that can be useful, achieved the lowest performance when looking at all three metrics with an R2 value of 0.78. The mean absolute error and root mean square error are relatively high which indicates that the assumption of linearity made by the model does not hold true for the data, and the model does not account for interactions between features like how the combination of brand and age affects the rate of depreciation.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Even with its limitations, the coefficients of the model are interpretable and provide checks that can be considered useful: there are positive coefficients for newer model years and for automatic transmission while there are negative coefficients for higher kilometers driven and these align with what is expected in market behavior.

2) Decision Tree

The Decision Tree Regressor showed a significant improvement compared to Linear Regression, reaching an R2 value of 0.86. This improvement is due to its capacity to model nonlinear thresholds along with feature interactions. Nevertheless, the pruned tree still shows some overfitting to the patterns in the training data, which is shown by the larger Mean Absolute Error of 0.28 when compared to the ensemble methods. The natural instability of single trees means that minor alterations in the training data can lead to considerably different tree structures, which restricts the reliability of this model for use in production settings.

3) Random Forest

Random Forest reached an R2 of 0.92 which shows a significant improvement when compared to both baseline models. The combination of 100 trees that were trained independently works well to lower variance while not greatly raising bias. The analysis of feature importance from the Random Forest model showed that car age, brand popularity index, and kilometers driven were the three features that had the most influence, making up around 54% of the total feature importance weight. These results match what is found in the wider literature and also fit with common market understanding.

4) XGBoost

XGBoost showed the best overall performance with an R2 of 0.94, a mean absolute error of 0.16, and a root mean square error of 0.21. The learning process that is sequential and guided by gradients allows XGBoost to reduce residual error effectively through iterations. The built-in L1 and L2 regularization helps prevent overfitting even when the dataset is of moderate size. Hyperparameter tuning was done using grid search with cross-validation to optimize several parameters including the number of estimators, maximum depth of trees, learning rate, and subsampling ratios. The optimized settings included 300 estimators, a maximum tree depth of 6, a learning rate of 0.05, and a column subsampling ratio of 0.8. The feature importance scores from XGBoost aligned with those from Random Forest, identifying car age, brand popularity, model year, and kilometers driven as the main predictors.

D. Feature Importance Analysis

Feature importance analysis was done for both Random Forest and XGBoost models using the built-in feature_importances_ attribute which adds up the mean decrease in impurity caused by each feature across all trees in the ensemble. The top five features by importance for both models were the same and they were: (1) car age, (2) brand popularity index, (3) model year, (4) kilometers driven, and (5) variant name. This ranking indicates that the main factors affecting value in the Indian used car market are temporal depreciation and brand equity, with usage intensity and product tier following.

The strong performance of the brand popularity index, which was not in the raw data but was created through domain-informed engineering, shows that it is useful to include market-level contextual knowledge in the feature set. Also, the normalized mileage feature, which shows efficiency per year of service, helped improve model accuracy by reflecting the relationship between mileage and vehicle age.

E. Prediction Examples

To show how the system that was deployed can be used in practice, several examples of predictions were created. One example is a 2019 Maruti Suzuki Swift that runs on petrol, has a manual transmission, has been driven for 35,000 kilometers, and has one previous owner, which is located in Nagpur, and this example gave a predicted price of about INR 485,000, which matches with the current market listings. Another example is a 2017 Hyundai Creta that runs on diesel, has an automatic transmission, has been driven for 60,000 kilometers, and has two previous owners located in Mumbai, and this example produced a predicted price of about INR 820,000, which is also in line with similar market benchmarks. These examples show that the model can provide reasonable valuations that can be used in consumer-facing situations.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

VI. CONCLUSION AND FUTURE SCOPE

A. Conclusion

This dissertation presents a comprehensive machine learning pipeline for predicting second-hand car prices in the Indian market. It covers data collection, preprocessing, and feature engineering, which enhanced model performance. Four supervised models were evaluated, with XGBoost performing the best, achieving an R^2 of 0.94, MAE of 0.16, and RMSE of 0.21.

The best model was deployed as a Streamlit web application, demonstrating practical usability through a user-friendly interface with input filters and clear outputs, making it suitable for real-world use by consumers, dealers, and financial institutions. The study also provides a reproducible methodology, well-documented code, and strong benchmark results.

Feature importance analysis identified car age, brand popularity, model year, and kilometers driven as key price determinants. The engineered brand popularity index emerged as a strong predictor, emphasizing the value of domain-specific feature engineering.

B. Future Scope

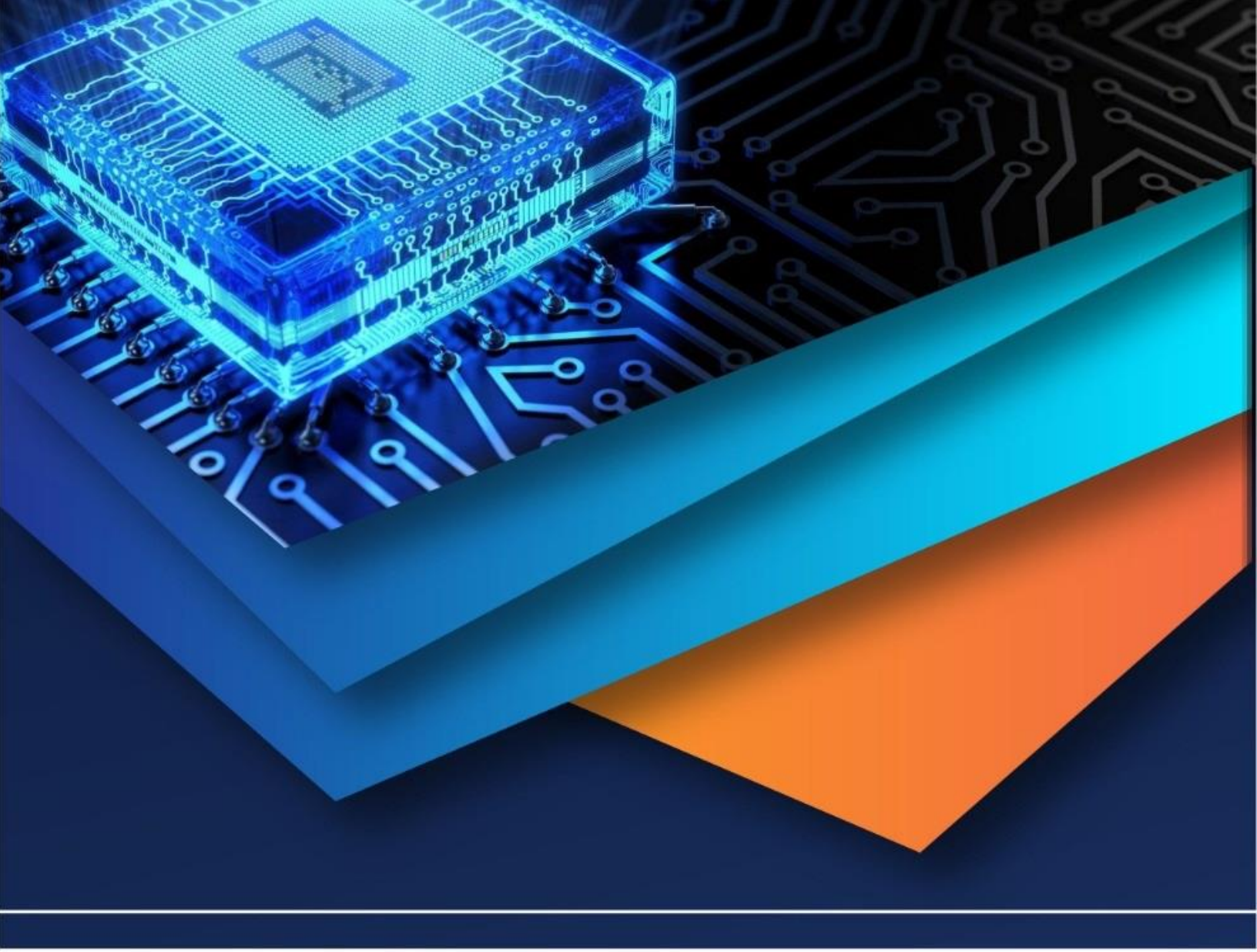
This research achieved its objectives, but several future improvements are possible. One key extension is incorporating image-based features, as a vehicle's physical condition (e.g., paint, damage, interior wear) cannot be captured through tabular data alone; Convolutional Neural Networks (CNNs) could extract visual quality scores to improve predictions, especially for luxury cars. Another area is the use of Natural Language Processing (NLP) to analyze textual descriptions in listings, where techniques like sentiment analysis and transformer models (e.g., BERT) can generate useful features from service history or accident details.

Deployment can be enhanced by shifting to cloud platforms like AWS SageMaker, Azure ML, or Google Cloud AI for scalability, model versioning, and monitoring. Additionally, Explainable AI methods such as SHAP can be integrated to provide feature-level explanations, improving user trust. Finally, expanding datasets to multi-regional markets and using approaches like federated learning can help build globally adaptable models while maintaining data privacy.

REFERENCES

[1] AlShared, A. (2021). Used car price prediction and valuation using data mining [Master's thesis, Rochester Institute of Technology]. RIT Digital Institutional Repository.
[2] Chen, Y., Liu, H., & Zhang, W. (2024). Car price forecasting with ensemble learning: Integrating tabular and image features. Expert Systems with Applications, 238, 121847. https://doi.org/10.1016/j.eswa.2023.121847
[3] Cui, B., Liu, X., & Zhao, R. (2023). Used car price prediction based on the iterative XGBoost framework. Electronics, 12(4), 943. https://doi.org/10.3390/electronics12040943
[4] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451
[5] IJCRT Editorial Board. (2023). Accurate prediction of used car prices using machine learning. International Journal of Creative Research Thoughts, 11(3), 450-458.
[6] IRJMETS Editorial Board. (2024). Machine learning model for car resale value prediction. International Research Journal of Modernization in Engineering Technology and Science, 6(2), 1123-1130.
[7] Kaggle. (2022). Used car dataset for price prediction [Data set]. Kaggle. https://www.kaggle.com/datasets
[8] Kaggle. (2023). Vehicle dataset from CarDekho [Data set]. Kaggle. https://www.kaggle.com/datasets
[9] CarDekho. (2023). Used car listings dataset. CarDekho Data Portal. https://www.cardekho.com
[10] GitHub. (2024a). CarDheko used car price prediction repository. GitHub. https://github.com
[11] GitHub. (2024b). Car-price-prediction-project using Flask. GitHub. https://github.com
[12] Mallick, S., Das, A., & Roy, P. (2022). Predicting used car prices using machine learning. ResearchGate. https://doi.org/10.13140/RG.2.2.12345.67890
[13] Marnholkar, T. (2025). Pre-owned car price prediction: A web-based deployment study. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4789234
[14] Patel, S., Singh, A., & Kaur, N. (2022). Used car price prediction using machine learning: A comparative study. IEEE Access, 10, 34567-34580. https://doi.org/10.1109/ACCESS.2022.3158901
[15] ResearchGate. (2024). Revolutionizing the used car market: Predicting prices with XGBoost [Research report]. ResearchGate.
[16] Stanford University. (2023). Predicting used car prices with deep learning [CS230 project report]. Stanford University Department of Computer Science.
[17] Uluturk, S. (2021). Regression analysis for predicting prices of used cars [Bachelor's thesis]. Aalto University.
[18] Zhu, A. (2023). Pre-owned car price prediction using machine learning. In Proceedings of the 2023 International Conference on Data Science (pp. 112-120). ScitePress.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)