2025 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE) | 979-8-3315-0103-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/AMATHE65477.2025.11081288

Enhanced Used Car Price Prediction Using Machine Learning: A Comparative Study of Regression Models

Goutham P Raj, Gregan George, Hadii Hasan, John Ashwin Delmon, Lis Jose

Department of Computer Science, Amal Jyothi College of Engineering (Autonomous), Kottayam, India
gouthampraj2025@cs.ajce.in, gregangeorge2025@cs.ajce.in, hadiihasan2025@cs.ajce.in,
johnashwindelmon2025@cs.ajce.in, lisjose@amaljyothi.ac.in

***Abstract*—The demand for accurate and efficient car price prediction systems is growing due to the rapid expansion of the used car market. This paper presents a comparative analysis of multiple machine learning models for predicting used car prices. The study utilizes a dataset containing various features such as the car's year of manufacture, mileage, engine capacity, fuel type, transmission type, km driven, owner type, seller type, seats, brand and model. Models including Lasso Regression, Multi-Layer Perceptron Regressor (MLPRegressor), Decision Tree Regressor, Gradient Boosting Regressor, Random Forest Regressor, and Extreme Gradient Boosting (XGBoost) were trained and evaluated. Performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) were computed to assess model effectiveness. Among these, XGBoost achieved the highest R^2 value of 0.9118, indicating its superior predictive accuracy and robustness in capturing complex patterns in the dataset. This work highlights the potential of advanced machine learning techniques in streamlining price estimation processes, thereby benefiting stakeholders in the used car industry. The findings also emphasize the importance of feature selection and model tuning in achieving optimal results.**

***Index Terms*—Used Car Price Prediction, Machine Learning Models, XGBoost, Lasso Regression, MLP Regressor, Decision Tree Regressor, Gradient Boosting, Random Forest Regressor**

I. Introduction

The used car market plays a pivotal role in the global automotive ecosystem, driven by the rising demand for cost-effective transportation solutions and growing consumer preference for sustainable practices. Determining the fair market value of a used car is a complex challenge, influenced by numerous factors such as the vehicle's age, mileage, make and model, fuel type, and overall condition. An accurate and efficient pricing mechanism is essential to promote transparency and trust in this highly dynamic and competitive market. Machine learning (ML) offers a transformative approach to tackle this challenge by leveraging data to uncover patterns and relationships that are otherwise difficult to discern. This paper examines the application of various regression-based ML models for predicting used car prices. The models under consideration include Lasso Regression, MLP Regressor, Decision Tree Regressor, Gradient Boosting Regressor, Random Forest Regressor, and XGBoost. Each model is assessed for its ability to handle diverse datasets and capture the intricate dependencies between a car's features and its price. The study is based on a dataset encompassing a wide range of attributes such as year of manufacture, mileage, engine capacity, and ownership history. By evaluating the strengths and limitations of different ML models, this research aims to identify approaches that best suit real-world scenarios. The insights gained from this study aim to benefit stakeholders, including dealers, buyers, and online platforms, by providing a reliable framework for price prediction. This framework has the potential to streamline the buying and selling process, offering greater confidence to all parties involved while fostering transparency and fairness in the marketplace.

II. Literature Survey

[1] presents an enhanced prediction model utilizing the XGBoost algorithm. The authors demonstrate that by optimizing hyperparameters and incorporating feature selection, the model achieves superior predictive performance compared to traditional methods. In [2], researchers investigate various optimization strategies to improve the accuracy of car price prediction models. The paper emphasizes the significance of data preprocessing and the selection of appropriate machine learning algorithms in developing robust predictive models. [3] explores the integration of explainable AI techniques to enhance the interpretability of logistic regression models. While not directly focused on car price prediction, the insights into model transparency and feature importance are applicable to developing interpretable predictive models in various domains. [4] explores various machine learning algorithms, including Linear Regression, Decision Trees, and Support Vector Machines, to predict used car prices. The authors emphasize the importance of feature selection and data preprocessing in enhancing model accuracy. Their findings suggest that ensemble methods, particularly Random Forests, outperform individual models in terms of predictive performance. [5] propose an automated system that leverages machine learning techniques to estimate used car prices. They compare models such as Multiple Linear Regression, Decision Trees, and Gradient

Boosting, concluding that Gradient Boosting provides superior accuracy. The study also highlights the significance of hyperparameter tuning and cross-validation in model optimization. [6] investigates the application of supervised learning algorithms, including K-Nearest Neighbors (KNN) and Support Vector Regression (SVR), for predicting used car prices. The research underscores the effectiveness of SVR in capturing non-linear relationships between features, leading to improved prediction accuracy.

III. Methodology

The methodology adopted for this study ensures a systematic approach to preparing the dataset and training machine learning models for accurate used car price prediction. The process is outlined as follows:

The initial dataset comprised various columns describing the features of used cars, such as name, year, selling_price, km_driven, fuel, seller_type, transmission, owner, mileage, engine, and seats. To prepare the data for analysis, the dataset was thoroughly inspected to identify and address issues such as missing values and duplicate entries. Rows containing null values were removed to maintain data integrity, and duplicate records were dropped to avoid redundancy and overfitting during model training. The name column, which contained the combined information about the car's brand and model, was split into two separate columns: brand and model. This step allowed for more granular analysis and enhanced the effectiveness of feature encoding. Categorical variables such as brand, fuel, seller_type, transmission, owner, and model were encoded into numerical representations to facilitate compatibility with machine learning algorithms. Continuous features like mileage and engine, initially stored as strings, were converted to numeric types (float) after extracting their numerical components. This transformation enabled these features to be properly utilized during the training process. The entire dataset was normalized and scaled where necessary to ensure that features with varying ranges did not disproportionately affect the model's performance.

After preprocessing, the dataset was divided into training and testing subsets to evaluate the performance of the models effectively. A standard split of 80% training and 20% testing was used. Various machine learning algorithms were employed for training, including Lasso Regression, Multi-Layer Perceptron (MLP) Regressor, Decision Tree Regressor, Gradient Boosting Regressor (GBR), Random Forest Regressor, and Extreme Gradient Boosting (XGBoost).

Lasso Regression employed regularization to focus on significant features while reducing overfitting. The MLP Regressor, a neural network model with three hidden layers, captured non-linear relationships after the dataset was scaled for optimal performance. The Decision Tree Regressor utilized hierarchical splits to model feature interactions effectively. Gradient Boosting techniques were implemented with GBR, leveraging learning rate adjustments and depth optimization to balance model complexity and overfitting. Random Forest Regressor improved prediction accuracy by averaging multiple decision trees, while XGBoost, a powerful gradient-boosting algorithm, employed features such as early stopping and regularization for enhanced accuracy and generalization.

Each model was trained on the prepared dataset, and its performance was evaluated using testing data. Evaluation metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and R-squared (R^2), were calculated to assess the models' accuracy and error margins. Scatter plots were generated to visualize the alignment between actual and predicted prices, enabling a comparative analysis of each model's effectiveness.

This methodological approach ensured high-quality data preprocessing, the implementation of diverse models, and the evaluation of performance through standardized practices, resulting in robust and interpretable predictions for used car prices.

IV. Results and Discussions

The results of this study provide insights into the performance of various machine learning models for predicting used car prices, evaluated using key metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and R-squared (R^2). Scatter plots depicting the relationship between actual and predicted prices were also generated for each model, offering a visual comparison of their accuracy.

Lasso Regression achieved an MAE of 188,221.90, an RMSE of 280,870.89, and an MSE of 78,888,457,280.09, with an R^2 value of 0.5523. While the model successfully reduced overfitting by penalizing less significant features, its relatively high error metrics and moderate R^2 indicate limitations in capturing the complex relationships inherent in the dataset.

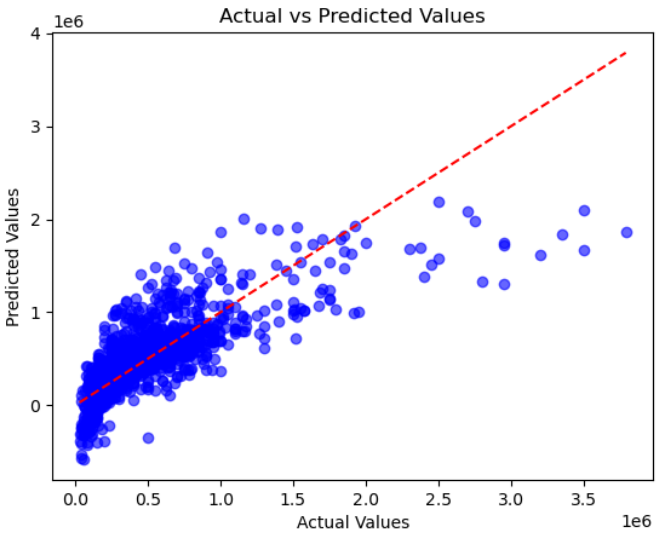

Figure 1: *Actual vs. predicted prices using Lasso Regression*

The MLP Regressor demonstrated significant improvement over Lasso Regression, achieving an MAE of 116,362.63, an RMSE of 225,901.28, and an MSE of 51,031,389,712.13,

with an R^2 value of 0.7104. The neural network's ability to model non-linear relationships contributed to this performance. However, its results suggest potential overfitting or sensitivity to hyperparameter tuning.

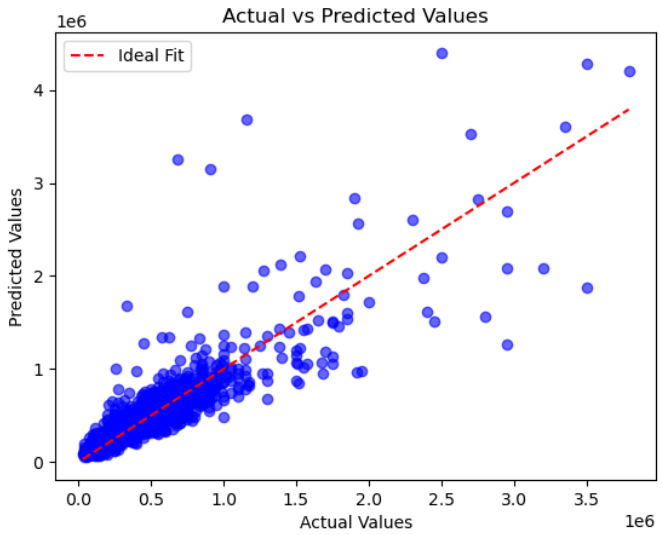

Figure 2: *Actual vs. predicted prices using MLP Regressor*

The Decision Tree Regressor outperformed both Lasso Regression and the MLP Regressor, with an MAE of 101,815.50, an RMSE of 200,641.56, and an MSE of 40,257,037,072.43, coupled with an R^2 value of 0.7715. Its hierarchical structure captured interactions between features effectively, resulting in improved accuracy and better alignment with the data.

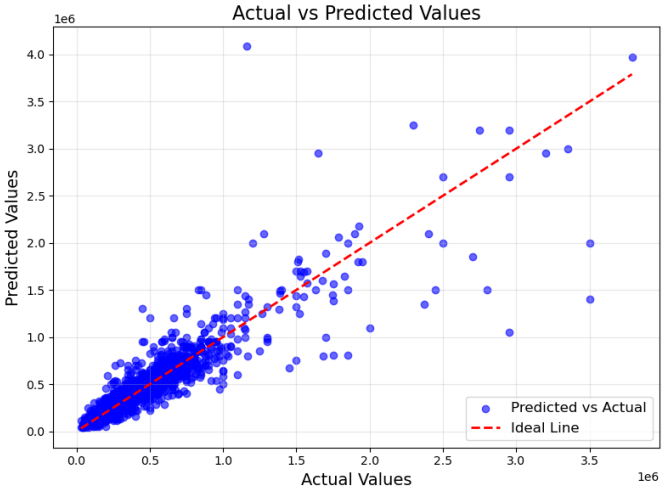

Figure 3: *Actual vs. predicted prices using Decision Tree Regressor*

The Gradient Boosting Regressor provided even stronger performance, achieving an MAE of 89,905.97, an RMSE of 154,948.65, and an MSE of 24,009,084,984.39, with an R^2 value of 0.8637. This model effectively combined weak learners, showcasing its robustness and ability to generalize across the dataset.

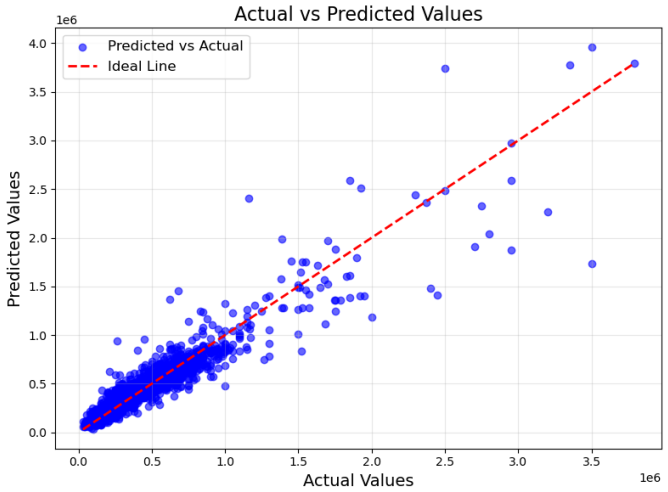

Figure 4: *Actual vs. predicted prices using Gradient Boosting Regressor*

The Random Forest Regressor further enhanced the results, with an MAE of 75,707.60, an RMSE of 147,305.25, and an MSE of 21,698,837,785.98, achieving an R^2 value of 0.8769. By averaging multiple decision trees, the model reduced variance and improved prediction accuracy, making it one of the top-performing models.

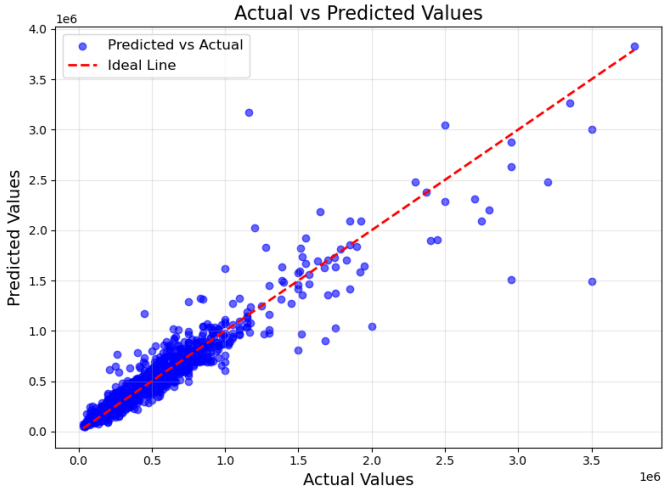

Figure 5: *Actual vs. predicted prices using Random Forest Regressor*

XGBoost delivered the best performance among all the models, with an MAE of 80,875.08, an RMSE of 143,085.30, and an MSE of 20,473,404,696.05, coupled with an exceptional R^2 value of 0.9118. Its advanced features, such as regularization, early stopping, and efficient handling of sparse data, enabled precise predictions with minimal error.

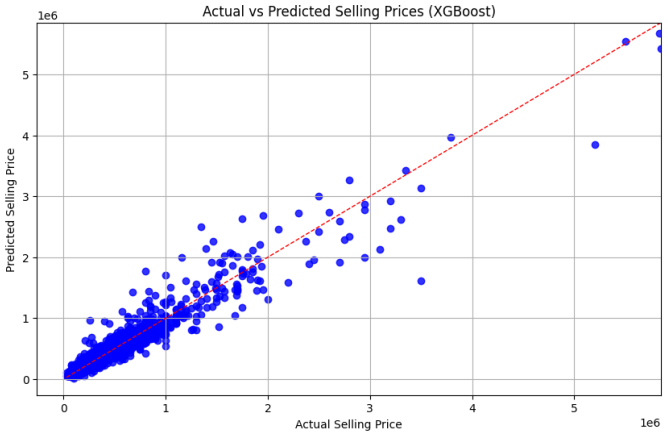

Figure 6: *Actual vs. predicted prices using XGBoost Regressor*

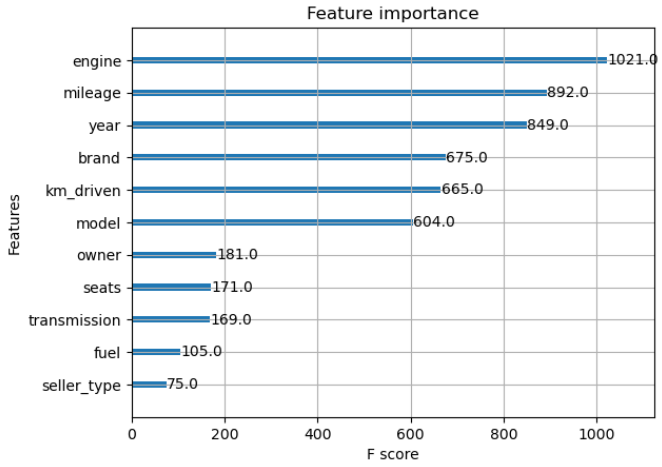

Figure 7: *Feature importance ranking from the XGBoost model*

The feature importance chart (figure 7) highlights the contribution of each input feature to the predictive performance of the XGBoost model. Key features like engine, mileage, and year demonstrate the highest importance, reflecting their strong influence on predicting car prices. In contrast, features such as seller_type and fuel have lower importance, indicating a relatively minor impact on the model predictions. The F-score quantifies each feature's contribution, emphasizing the hierarchical significance of these attributes.

TABLE I: Performance Evaluation of Prediction Models

Model	MAE	RMSE	MSE	R^2
Lasso Regression	188221.90	280870.89	78888457280.09	0.5523
MLP Regressor	116362.63	225901.28	51031389712.13	0.7104
Decision Tree	101815.50	200641.56	40257037072.43	0.7715
Gradient Boosting	89905.97	154948.65	24009084984.39	0.8637
Random Forest	75707.60	147305.25	21698837785.98	0.8769
XGBoost	80875.08	143085.30	20473404696.05	0.9118

Table I compares the performance of machine learning models for predicting used car prices using metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and R^2. Among the models, XGBoost achieved the best results, with the MAE (80,875.08), RMSE (143,085.30), and MSE (20,473,404,696.05), along with the highest R^2 (0.9118), indicating superior prediction accuracy.

Ensemble models like Random Forest and Gradient Boosting also performed well, while Lasso Regression showed the weakest performance, with the highest errors and lowest R^2. These results highlight the effectiveness of advanced models like XGBoost for complex datasets and guide model selection for similar predictive tasks.

The results demonstrate the progressive improvement in model accuracy as more complex algorithms are employed. Lasso Regression, though interpretable, struggled to capture the nuances of the dataset. Decision Tree-based models, including Random Forest and Gradient Boosting, excelled due to their ability to model feature interactions effectively. XGBoost stood out as the best-performing model, leveraging its advanced boosting techniques and robust handling of data variations.

These findings highlight the importance of selecting appropriate models and tuning their parameters for predictive tasks. The visual analysis provided by scatter plots further corroborates the quantitative metrics, validating the performance of the top models for practical applications in used car price prediction.

V. Conclusion

This study demonstrates the effectiveness of various machine learning models for predicting used car prices, with a focus on improving accuracy and minimizing errors. Among the evaluated models, XGBoost emerged as the most effective, achieving the highest R^2 score, showcasing its ability to handle complex relationships within the dataset. Ensemble models like Random Forest and Gradient Boosting also delivered strong results, while simpler models like Lasso Regression struggled to capture the dataset's nuances.

The findings underscore the importance of selecting advanced algorithms for predictive tasks involving high-dimensional data. These results can serve as a foundation for building robust applications in the used car market, enabling informed decision-making for both buyers and sellers. Future work can explore integrating real-time data, optimizing feature selection, and enhancing interpretability to further improve prediction accuracy and usability.

References

[1] C. Sheng and H. Yu, "An optimized prediction algorithm based on XGBoost," *2022 International Conference on Networking and Network Applications (NaNA), Urumqi, China, 2022, pp. 1-6.*

[2] D. J. Vijaya, A. Gopu, V. Reddy and A. Kaul, "Optimization Techniques for Car Price Prediction," *2023 International Conference on Computer, Electronics & Electrical Engineering & their Applications (IC2E3), Srinagar Garhwal, India, 2023, pp. 1-6.*

[3] Y. Yang and M. Wu, "Explainable Machine Learning for Improving Logistic Regression Models," *2021 IEEE 19th International Conference on Industrial Informatics (INDIN), Palma de Mallorca, Spain, 2021, pp. 1-6.*

[4] M. Hankar, M. Birjali and A. Beni-Hssane, "Used Car Price Prediction using Machine Learning: A Case Study," *2022 11th International Symposium on Signal, Image, Video and Communications (ISIVC), El Jadida, Morocco, 2022, pp. 1-4.*

[5] S. K. Satapathy, R. Vala and S. Virpariya, "An Automated Car Price Prediction System Using Effective Machine Learning Techniques," *2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, 2022, pp. 402-408.*

[6] F. Wang, X. Zhang and Q. Wang, "Prediction of Used Car Price Based on Supervised Learning Algorithm," *2021 International Conference on Networking, Communications and Information Technology (NetCIT), Manchester, United Kingdom, 2021, pp. 143-147.*