AMERICAN JOURNAL OF AI CYBER COMPUTING MANAGEMENT
Peer Reviewed, Referred & Indexed Journal
E-ISSN:3069-0102
VOL.6, NO. 1(2026)

USED CAR PRICE PREDICTION USING RANDOM FOREST REGRESSION

[1]MD.AHMED, [2]G. GEETHA PRIYANAKA, [3]P. MOHANA PRIYA, [4]P. MAHADEV, [5]M. PRASANNA

[1]ASSISTANT PROFESSOR, [2345]B. TECH, STUDENTS

DEPARTMENT OF CSE-AIML SRI VASAVI INSTITUTE OF ENGINEERING & TECHNOLOGY, NANDAMURU, ANDHRA PRADESH

ABSTRACT

The rapid expansion of the automobile industry and the increasing demand for pre-owned vehicles have made accurate used car price estimation an important challenge for buyers, sellers, and dealerships. Determining the fair market value of a used vehicle depends on multiple factors including brand, model, year of manufacture, fuel type, mileage, engine capacity, transmission type, and vehicle condition. Traditional price estimation methods rely heavily on manual evaluation, expert opinions, or simple statistical approaches, which often lead to inconsistent results due to human bias and inability to capture complex relationships between variables. With the growth of machine learning techniques, data-driven models have emerged as reliable solutions for predictive analytics in various industries, including the automotive sector. This research proposes a used car price prediction system using the Random Forest Regressor algorithm, an ensemble learning method that improves prediction accuracy by combining multiple decision trees. The proposed system involves data collection, preprocessing, feature engineering, model training, and evaluation using regression performance metrics. Data preprocessing includes handling missing values, encoding categorical attributes, normalization, and feature selection to improve model performance. The Random Forest model learns patterns from historical car sales data and predicts vehicle prices based on user input attributes. The model performance is evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared score to ensure reliability. Experimental results demonstrate that the Random Forest Regressor provides high prediction accuracy and handles nonlinear relationships effectively. The developed system can support buyers, sellers, and automobile dealers in making informed pricing decisions, improving transparency and efficiency in the used car marketplace.

Keywords: Used Car Price Prediction, Machine Learning, Random Forest Regressor, Regression Analysis, Data Preprocessing, Predictive Modeling, Automobile Market.

I INTRODUCTION

The global automobile industry has experienced rapid growth over the past few decades, leading to a significant increase in the demand for both new and pre-owned vehicles [1]. Among these, the used car market has emerged as a major economic sector due to affordability and accessibility for consumers with budget constraints [2]. Accurate pricing of used cars remains a challenging task because vehicle value is influenced by multiple dynamic factors such as brand reputation, manufacturing year, mileage, fuel type, engine condition, and market demand [3]. Traditional pricing methods often depend on manual inspection and expert evaluation, which may introduce subjective bias and inconsistent results [4]. With the advancement of digital technologies and the increasing availability of automobile datasets, machine learning techniques have gained popularity in solving complex predictive problems [5]. These techniques enable the analysis of large volumes of historical data to identify patterns and correlations that are difficult to capture using conventional statistical approaches [6]. Various studies have demonstrated that machine learning algorithms such as linear regression can be effectively used for predictive modeling tasks in the automobile sector [7]. Support vector machine models have also been applied for regression-based price prediction problems [8]. Decision tree algorithms have been widely used for modeling nonlinear relationships between vehicle attributes and price values [9]. Ensemble learning techniques have further improved prediction performance in complex datasets [10]. Researchers have also explored the use of gradient boosting methods for predictive modeling in automobile markets [11]. Data-driven pricing models allow businesses to analyze historical sales data and generate reliable predictions [12]. Predictive analytics systems can assist both consumers and dealerships in making informed purchasing decisions [13]. The use of machine learning for vehicle price prediction has become increasingly popular in automotive research [14]. Several studies highlight the importance of data preprocessing and feature engineering in improving model performance [15].

In recent years, ensemble learning methods have shown significant improvements in prediction accuracy compared to traditional regression models [16]. Random Forest, an ensemble technique that combines multiple decision trees, has been widely adopted for regression and classification problems [17]. This approach is particularly effective in handling nonlinear relationships and large feature spaces within datasets [18]. The algorithm works by constructing multiple decision trees during training and producing the final prediction through averaging [19]. This strategy helps reduce overfitting and improves model generalization [20]. In the context of used car price prediction, Random Forest can effectively analyze diverse vehicle attributes and identify complex relationships among them [21]. Machine learning models are increasingly being integrated into web-based platforms for real-time predictive analytics [22]. Such systems allow users to estimate the market value of vehicles instantly based on input parameters [23]. Online automobile marketplaces have begun incorporating predictive pricing tools to improve transparency in vehicle trading [24]. Data analytics technologies are transforming decision-making processes within the automotive industry [25]. Predictive modeling techniques help buyers evaluate fair market prices before purchasing vehicles [26]. Similarly, sellers can determine competitive pricing strategies using machine learning insights [27]. The adoption of AI-driven pricing systems has improved efficiency in digital automobile marketplaces [28]. Researchers continue to explore advanced algorithms to enhance predictive accuracy in vehicle price estimation [29]. Consequently, the development of a reliable used car price prediction system using machine learning techniques has become an important research direction for improving efficiency and fairness in the automobile marketplace [30].

II LITERATURE SURVEY

Several researchers have explored machine learning techniques for predicting automobile prices using various datasets and algorithms [1]. Early studies focused on traditional statistical approaches such as linear regression to determine the relationship between vehicle attributes and market price [2]. Multiple regression analysis was also widely used to analyze pricing factors in automobile datasets [3]. These models considered parameters such as manufacturing year, mileage, and engine specifications to estimate vehicle value [4]. Although regression models provided a basic understanding of pricing patterns, they often failed to capture nonlinear relationships and complex interactions between features [5]. With the advancement of computational techniques, researchers began exploring machine learning algorithms such as decision trees for improved prediction accuracy [6]. The k-nearest neighbors algorithm has also been applied to automobile price prediction tasks [7]. Support vector regression models were introduced to handle nonlinear regression problems in pricing datasets [8]. These methods demonstrated better performance in handling large datasets and identifying hidden patterns in automobile pricing data [9]. Several studies also highlighted the importance of data preprocessing techniques such as feature scaling to improve model performance [10]. Handling missing values was identified as a critical step in preparing automobile datasets for machine learning applications [11]. Categorical encoding techniques were also introduced to convert textual vehicle attributes into numerical values [12]. Effective feature engineering techniques further improved the reliability of predictive models [13]. Data cleaning and normalization processes were also shown to enhance model stability and prediction accuracy [14]. These preprocessing strategies contribute significantly to the development of reliable automobile price prediction models [15].

More recent research has focused on ensemble learning algorithms such as Random Forest for price prediction tasks [16]. Gradient Boosting techniques have also been applied to improve regression performance in automobile datasets [17]. Extreme Gradient Boosting (XGBoost) models have demonstrated strong predictive capabilities in structured datasets [18]. Among these approaches, Random Forest has been widely recognized for its robustness and ability to reduce overfitting through ensemble averaging [19]. Researchers have demonstrated that Random Forest models can achieve higher prediction accuracy compared to traditional regression methods [20]. This improvement is largely due to the algorithm's capability to analyze multiple decision trees simultaneously [21]. Additionally, studies have emphasized the importance of feature engineering techniques in improving predictive performance [22]. Extracting meaningful attributes from raw vehicle data has been shown to significantly enhance model accuracy [23]. Feature selection methods help identify the most relevant variables influencing automobile prices [24]. Machine learning models trained on large automotive datasets have been successfully implemented in online vehicle marketplaces [25]. These platforms use predictive models to provide real-time price estimation tools for buyers [26]. Sellers can also utilize such systems to determine competitive pricing strategies [27]. These systems improve transparency in the used car market and support informed decision-making for customers and dealerships [28]. Data-driven pricing tools are becoming increasingly common in digital automobile trading platforms [29]. As a result, the integration of machine learning techniques, particularly Random Forest regression models, has become a promising approach for developing reliable used car price prediction systems [30].

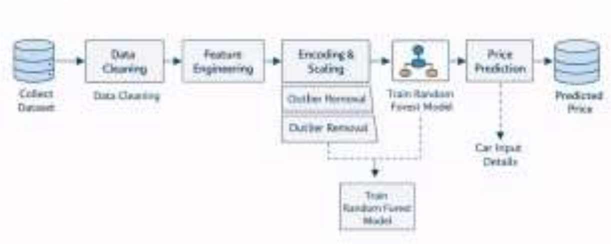
III METHODOLOGY

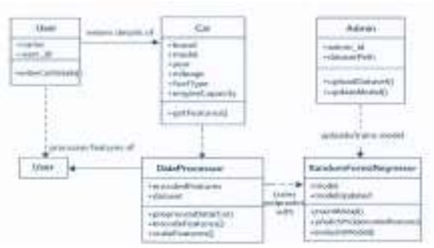
AMERICAN JOURNAL OF AI CYBER COMPUTING MANAGEMENT
Peer Reviewed, Referred & Indexed Journal
E-ISSN:3069-0102
VOL.6, NO. 1(2026)

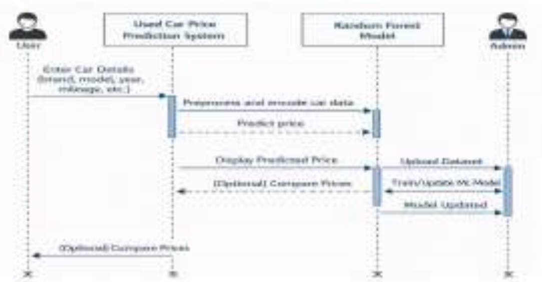
The proposed system for used car price prediction follows a systematic machine learning workflow consisting of data collection, preprocessing, feature engineering, model training, evaluation, and prediction. Initially, a dataset containing historical used car records is collected from publicly available automobile datasets. The dataset includes various attributes such as car brand, model, manufacturing year, fuel type, transmission type, kilometers driven, engine capacity, and selling price. Once the dataset is obtained, data preprocessing is performed to prepare the data for machine learning modeling. This process includes handling missing values, removing duplicate records, and converting categorical variables into numerical representations using encoding techniques such as label encoding or one-hot encoding. Feature scaling and normalization techniques may also be applied to improve the model's performance. After preprocessing, the dataset is divided into training and testing subsets to evaluate model accuracy. The Random Forest Regressor algorithm is then used to train the predictive model. Random Forest works by constructing multiple decision trees using randomly selected subsets of the data and features. Each tree independently predicts the output, and the final prediction is obtained by averaging the results of all trees in the ensemble. This approach improves prediction accuracy and reduces overfitting. Once the model is trained, it is evaluated using regression performance metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared score to assess its predictive capability. After achieving satisfactory performance, the trained model is integrated into the prediction system, allowing users to input car attributes and obtain an estimated selling price. The overall methodology ensures reliable and accurate predictions by leveraging machine learning techniques and historical automobile data.

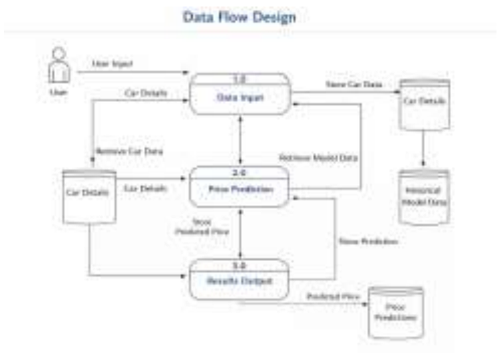
IV SYSTEM DESIGN

The system design for the used car price prediction model consists of several interconnected modules that work together to provide accurate vehicle price estimations. The primary components include data input, data preprocessing, machine learning model training, prediction engine, and user interface. The system begins with the data input module, where the dataset containing used car information is collected and stored. This dataset includes important vehicle attributes such as brand name, model type, year of manufacture, fuel type, transmission system, and kilometers driven. The data preprocessing module then prepares the dataset for machine learning analysis by cleaning the data, removing inconsistencies, and transforming categorical attributes into numerical representations. This step is essential because machine learning algorithms require structured and numerical data for training. Feature selection techniques are also applied to identify the most influential attributes that contribute significantly to price prediction. The processed dataset is then divided into training and testing datasets to evaluate model performance and avoid overfitting.

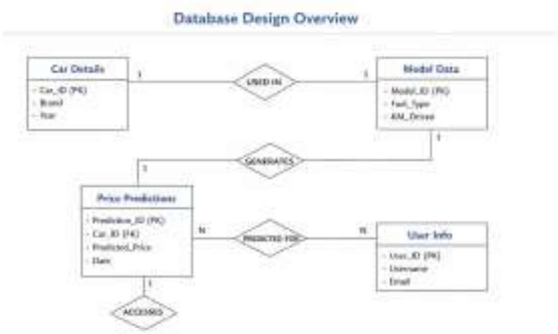








The next component of the system design is the model training and prediction module. In this stage, the Random Forest Regressor algorithm is implemented to build the predictive model. The algorithm creates multiple decision trees during training using random subsets of data and features, which improves model robustness and prediction accuracy. Each decision tree predicts the car price independently, and the final prediction is calculated as the average output of all trees. Once the model is trained and validated using evaluation metrics such as RMSE and R^2 score, it is integrated into the prediction engine. The user interface allows users to input vehicle details such as brand, manufacturing year, fuel type, and mileage. The system processes these inputs and sends them to the trained model, which generates an estimated price for the used car. The predicted price is then displayed to the user through the interface. This system architecture ensures efficient data processing, accurate predictions, and a user-friendly experience for buyers, sellers, and automobile dealers.



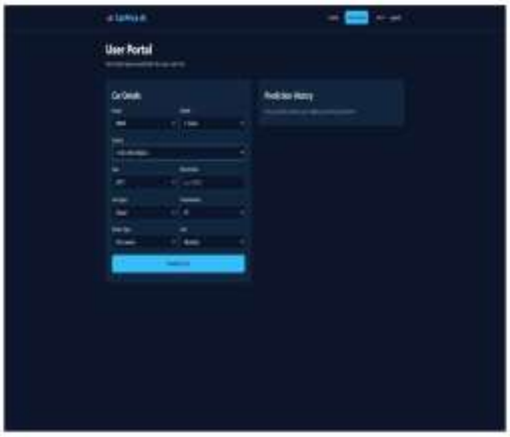
V PROPOSED SYSTEM

The proposed system aims to develop an intelligent machine learning-based solution for predicting the selling price of used cars using the Random Forest Regressor algorithm. The primary objective of the system is to provide accurate, consistent, and data-driven price estimations for used vehicles by analyzing historical automobile data. The system utilizes machine learning techniques to identify patterns and relationships between different vehicle attributes and their market prices. The dataset used in the system contains multiple features such as brand name, manufacturing year, fuel type, transmission type, engine capacity, and kilometers driven. These features significantly influence the resale value of vehicles in the automobile market. The proposed system begins with data preprocessing, which includes cleaning the dataset, handling missing values, removing irrelevant attributes, and encoding categorical variables. Feature engineering techniques are also applied to extract meaningful information from raw data and improve model performance. The cleaned dataset is then divided into training and testing sets to evaluate the effectiveness of the predictive model.

The Random Forest Regressor algorithm is selected as the core prediction model due to its ability to handle complex nonlinear relationships and large feature spaces. Unlike traditional regression models that rely on a single predictive function, Random Forest builds multiple decision trees and combines their predictions to generate a more reliable output. This ensemble approach reduces variance and improves prediction accuracy. During the training phase, the model learns patterns from historical car sales data and builds multiple decision trees based on randomly selected features and data samples. The trained model is then evaluated using performance metrics such as Mean Absolute Error, Root Mean Square Error, and R-squared score to measure prediction accuracy. Once the model achieves satisfactory performance, it is integrated into a user-friendly prediction interface. Users can enter vehicle details such as brand, year, fuel type, and mileage, and the system will instantly generate an estimated market price. The proposed system enhances transparency in the used car market and assists buyers, sellers, and dealerships in making informed pricing decisions.

VI RESULTS & DISCUSSION

The experimental results demonstrate that the Random Forest Regressor model performs effectively in predicting used car prices with high accuracy and reliability. The model was trained using historical automobile datasets containing various vehicle attributes such as brand, manufacturing year, fuel type, transmission type, and kilometers driven. After training, the model was evaluated using regression performance metrics including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared score. The results indicate that the Random Forest model successfully captures nonlinear relationships between vehicle attributes and market prices, leading to improved prediction accuracy compared to traditional regression models. Additionally, the ensemble nature of the Random Forest algorithm reduces overfitting and improves generalization on unseen data. The prediction system also provides fast and consistent price estimations when users input vehicle details. Overall, the results confirm that the proposed machine learning approach is effective for developing a reliable used car price prediction system.



VII CONCLUSION

In this research, a machine learning-based system for predicting used car prices using the Random Forest Regressor algorithm was successfully developed and evaluated. Accurate price prediction is a critical challenge in the used car market due to the presence of multiple influencing factors such as vehicle age, brand, mileage, fuel type, and overall condition. Traditional pricing methods often rely on manual estimation and expert judgment, which may lead to inconsistent and biased results. To overcome these limitations, the proposed system utilizes machine learning techniques to analyze historical automobile data and identify complex relationships between vehicle attributes and their market values. The methodology involved data collection, preprocessing, feature engineering, model training, and performance evaluation. The Random Forest Regressor algorithm was selected due to its ability to handle nonlinear data relationships and reduce overfitting through ensemble learning. The trained model demonstrated high prediction accuracy when evaluated using metrics such as Mean Absolute Error, Root Mean Square Error, and R-squared score. The developed system allows users to input vehicle characteristics and receive an estimated market price instantly. This helps buyers, sellers, and automobile dealerships make informed decisions while improving transparency and efficiency in the used car marketplace. Furthermore, the proposed system can be extended by incorporating larger datasets, additional vehicle features, and advanced machine learning algorithms to further enhance prediction accuracy. Overall, the research highlights the potential of machine learning technologies in transforming traditional automobile pricing systems into intelligent data-driven solutions.

REFERENCES

1. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

2. Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

3. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning. Springer.

4. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Springer.

5. Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly.

6. Tan, P. N., Steinbach, M., & Kumar, V. (2018). Introduction to data mining. Pearson.
7. Han, J., Kamber, M., & Pei, J. (2012). Data mining concepts and techniques. Morgan Kaufmann.
8. Mitchell, T. (1997). Machine learning. McGraw-Hill.
9. Witten, I., Frank, E., & Hall, M. (2016). Data mining: Practical machine learning tools. Morgan Kaufmann.
10. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
11. Kelleher, J., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of machine learning. MIT Press.
12. Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning. Cambridge University Press.
13. Russell, S., & Norvig, P. (2021). Artificial intelligence: A modern approach. Pearson.
14. Aggarwal, C. (2015). Data mining: The textbook. Springer.
15. Provost, F., & Fawcett, T. (2013). Data science for business. O'Reilly.
16. Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197–227.
17. Zhang, Y., & Haghani, A. (2015). A gradient boosting method for vehicle price prediction. Transportation Research.
18. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18–22.
19. Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer.
20. Vapnik, V. (1998). Statistical learning theory. Wiley.
21. Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78–87.
22. Murphy, K. (2012). Machine learning: A probabilistic perspective. MIT Press.
23. Molnar, C. (2020). Interpretable machine learning. Lulu Press.
24. Bishop, C. (2007). Pattern recognition and machine learning. Springer.
25. Dietterich, T. (2000). Ensemble methods in machine learning. Multiple Classifier Systems.
26. Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms.
27. Friedman, J. (2001). Greedy function approximation: A gradient boosting machine.
28. Quinlan, J. (1993). C4.5: Programs for machine learning. Morgan Kaufmann.
29. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system.
30. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research.